Advancing Semantic Understanding in Multilingual and Multimodal Contexts

Miaoran Zhang

A dissertation submitted towards the degree of

Doctor of Engineering (Dr.-Ing.)

of the Faculty of Mathematics and Computer Science of Saarland University

Saarbrücken, 2025

DAY OF COLLOQUIUM:
23.03.2026

DEAN OF THE FACULTY:
Prof. Dr. Roland Speicher

EXAMINATION COMMITTEE:

Chair:	Prof. Dr. Josef van Genabith
First Reviewer, Advisor:	Prof. Dr. Dietrich Klakow
Second Reviewer:	Prof. Dr. Michael Hahn
Committee Member:	Dr. Ji-Ung Lee

ABSTRACT

Human beings possess a remarkable ability to understand the meaning of messages and conversations by leveraging their knowledge of language and contextual cues. However, for natural language processing (NLP) systems, accurately capturing semantic information in texts remains a significant challenge. This challenge stems from the inherent complexity of human language, such as linguistic ambiguity and long-range dependencies, and becomes even more pronounced in low-resource languages with limited data. Furthermore, language does not exist in isolation – it is often intertwined with other modalities. To develop NLP systems for real-world applications, it is crucial to effectively encode and model semantic information in dynamic multilingual and multimodal scenarios.

In this dissertation, we present a series of studies to enhance the semantic understanding of NLP systems across diverse tasks. First, we investigate the key factors that influence word embedding learning in multiple languages. By systematically evaluating the effects of the learning algorithm, corpus size, and training parameters, we provide actionable insights to generate high-quality word representations. Next, we introduce a novel method to learn sentence embeddings by exploiting visual and textual information via a multimodal contrastive objective. This approach demonstrates significant performance improvements on semantic similarity tasks and offers a versatile technique for integrating multimodal data into text representation learning. Third, we develop a framework to predict semantic relatedness for under-represented languages, addressing data scarcity through data augmentation and facilitating effective cross-lingual transfer using adapters. Finally, we examine the impact of few-shot demonstrations in in-context learning across a wide range of languages and tasks that require nuanced semantic understanding, revealing that their impact may have been overestimated in prior work and is highly context-dependent. Overall, these studies combine technical innovations with in-depth analysis, facilitating the development of more robust, multilingual, and multimodal intelligent systems.

ZUSAMMENFASSUNG

Menschen verfügen über die bemerkenswerte Fähigkeit, die Bedeutung von Botschaften und Gesprächen zu verstehen, indem sie ihr Wissen über Sprache und kontextuelle Hinweise nutzen. Für Systeme zur Verarbeitung natürlicher Sprache (NLP) bleibt die genaue Erfassung semantischer Informationen in Texten jedoch eine große Herausforderung. Diese Herausforderung ergibt sich aus der inhärenten Komplexität der menschlichen Sprache, wie z. B. sprachlicher Mehrdeutigkeit und weitreichenden Abhängigkeiten, und wird bei Sprachen mit geringen Ressourcen und begrenzten Daten noch deutlicher. Darüber hinaus existiert Sprache nicht isoliert, sondern ist oft mit anderen Modalitäten verflochten. Um NLP-Systeme für reale Anwendungen zu entwickeln, ist es entscheidend, semantische Informationen in dynamischen mehrsprachigen und multimodalen Szenarien effektiv zu kodieren und zu modellieren.

In dieser Dissertation stellen wir eine Reihe von Studien vor, die das semantische Verständnis von NLP-Systemen bei verschiedenen Aufgaben verbessern sollen. Zunächst untersuchen wir die Schlüsselfaktoren, die das Lernen von Wort-Embeddings in mehreren Sprachen beeinflussen. Durch die systematische Bewertung der Auswirkungen des Lernalgorithmus, der Korpusgröße und der Trainingsparameter liefern wir umsetzbare Erkenntnisse zur Generierung hochwertiger Wortdarstellungen. Als Nächstes stellen wir eine neuartige Methode zum Lernen von Satz-Embeddings vor, bei der visuelle und textuelle Informationen über ein multimodales kontrastives Ziel genutzt werden. Dieser Ansatz zeigt signifikante Leistungsverbesserungen bei Aufgaben zur semantischen Ähnlichkeit und bietet eine vielseitige Technik zur Integration multimodaler Daten in das Lernen von Textdarstellungen. Drittens entwickeln wir ein Framework zur Vorhersage der semantischen Verwandtschaft für unterrepräsentierte Sprachen, das Datenknappheit durch Datenvergrößerung behebt und einen effektiven sprachübergreifenden Transfer mithilfe von Adaptern ermöglicht. Schließlich untersuchen wir die Auswirkungen von Few-Shot-Demonstrationen beim kontextbezogenen Lernen über eine Vielzahl von Sprachen und Aufgaben hinweg, die ein nuanciertes semantisches Verständnis erfordern, und zeigen, dass ihre Auswirkungen in früheren Arbeiten möglicherweise überschätzt wurden und in hohem Maße kontextabhängig sind. Insgesamt verbinden diese Studien technische Innovationen mit eingehenden Analysen und erleichtern so die Entwicklung robusterer, mehrsprachiger und multimodaler intelligenter Systeme.

ACKNOWLEDGMENTS

Pursuing a PhD has been a long and challenging journey, and I would like to express my sincere gratitude to those people who supported and helped me along the way.

First, I would like to thank my advisor, Dietrich Klakow, who offered me the opportunity to start this journey. He provided an inclusive environment and gave me the freedom to focus on my own research without other pressure. Thank you for your constant support.

I am also thankful to many great collaborators I had the privilege of working with during my PhD: Marius Mosbach, Vagrant Gautam, Xiaoyu Shen, David Ifeoluwa Adelani, Michael A. Hedderich, Jesujoba Oluwadara Alabi, Mingyang Wang, Dawei Zhu, Anupama Chingacham, and Dana Ruiter. Thank you, Marius and Xiaoyu, for your consistently insightful and critical feedback during our discussions, which inspired me to become a better researcher. Vagrant, thank you for your unwavering support, giving me both scientific suggestions and the emotional strength to keep me moving forward. David and Michael, thank you for your advice on getting my first paper published. JJ and Mingyang, thank you for making our best paper award a reality. Thank you, Dawei, for your sincere advice and perspective during our discussions. Dana and Anu, thank you for your encouragement, which helped me overcome my initial self-doubt.

Furthermore, I am truly grateful to meet wonderful colleagues at LSV: Badr M. Abdullah, Paloma Garcia de Herreros, Zena Al-Khalili, Aravind Krishnan, Julius Steuer, Volha Petukhova, Alexander Blatt, Israel Abebe Azime, Pushkar Jajoria, and our associate member Koel Dutta Chowdhury. Thank you, Badr, for your various suggestions over the years. Zena and Paloma, thank you for the wonderful time we spent together and for helping me through the thesis submission process. Aravind, thank you for being our office therapist in the corridor. Koel, thank you for joining us, and your companionship helped light my way. I also want to express my thanks to our secretary, Claudia Verburg, and IT administrator, Nicolas Louis. They have handled a lot of difficult matters and made our office life so much easier.

My PhD life in Saarbrücken was largely enriched by many lovely friends I met here: Bixian Ying, Dingfan Chen, Dingding Li, Hui-po Wang, Yaoyao Liu, Wenjia Xu, Muqing Li, Fangzhou Zhai, Yuan Xin, Rui Ye, Yiting Xia, Hejing Li, and Hanwei Zhang. We have many unforgettable experiences together, such as cooking traditional Chinese cuisine, exploring restaurants, traveling, hiking, kayaking, and swimming. We also shared our honest opinions about research and

supported one another through the highs and lows of the doctoral study. I am grateful for the warmth and joy you brought into my life.

Finally, and most importantly, I wish to thank my family for their unconditional love and endless support. To my husband, Yang Chen – you are the quiet strength behind every brave step I have taken. Thank you for your unwavering love throughout all these years.

CONTENTS

1 INTRODUCTION

To enable machines to process human language, the field of Natural Language Processing (NLP) has achieved remarkable progress in recent years, powered by the rapid development of language models (Vaswani et al., 2017; Brown et al., 2020; OpenAI et al., 2023, *inter alia*). A major aspect of NLP technologies is the semantic understanding of text – the ability to interpret the meaning of linguistic elements, such as words and sentences, and recognize their relationships within a specific context. This task faces significant challenges due to the intricate and dynamic nature of human language. For instance, language use can be highly creative, as seen in figurative expressions; words may carry different meanings depending on the context information, which introduces layers of ambiguity. Achieving human-like comprehension remains a challenge for advanced NLP systems.

Beyond the inherent difficulties, there are two key factors that impact the semantic understanding capability of models in real-world applications. First, more than 7,100 living languages exist worldwide (Eberhard et al., 2025), yet the development of NLP has focused mainly on a very small number of languages, particularly English. The main reason is the uneven distribution of resources between languages (Joshi et al., 2020), such as labeled and unlabeled corpora on the web. Modern NLP systems, which are heavily dependent on these resources, exhibit a strong bias toward high-resource languages. Meanwhile, their performance is neither always language-agnostic nor easily transferable across languages. Substantial efforts have been made to improve model performance in low-resource settings (Hedderich et al., 2021; Haddow et al., 2022; Qin et al., 2024); however, the gap between high- and low-resource languages still exists.

Second, while text is an essential medium for human communication, it is often intertwined with other modalities (e.g., vision) in daily use, influencing the comprehension of meaning. On the one hand, non-linguistic contexts can alter the intention and sentiment conveyed through textual expressions, as in the case of sarcasm without explicit linguistic markers (Castro et al., 2019). On the other hand, incorporating multimodal information can improve model performance on textual tasks by providing additional contextual cues (Yao and Wan, 2020; Xu et al., 2021). Recently, there has been growing interest in developing multimodal models capable of processing and integrating different modalities simultaneously (Alayrac et al., 2022; Liu et al., 2023; Team et al., 2023), and these models have shown great poten-

tial in achieving a more comprehensive understanding in complex scenarios.

In light of these facts, this dissertation explores and proposes methods to enhance the semantic understanding capability of language models in the context of *language diversity* and *multimodal richness*. Our research focuses on two key perspectives: (1) **encoding** text into representations that reflect the underlying meaning, and (2) **modeling** semantic relationships to perform various tasks.

Specifically, to obtain meaningful text representations, we explore distributional learning methods to generate word embeddings for multiple languages, particularly in low-resource settings. We also propose a novel contrastive learning approach that leverages multimodal data to produce high-quality sentence embeddings. To model contextual relationships, we develop a framework that captures semantic relatedness for a wide range of languages, incorporating techniques such as data augmentation and parameter-efficient training. Lastly, we extend our study to a diverse set of tasks that require nuanced semantic understanding, analyzing the impact of demonstrations on the target input in multilingual in-context learning. Overall, this dissertation combines technical innovations with in-depth analysis, offering new insights to enhance both the capability and inclusivity of modern NLP systems.

1.1 THESIS OUTLINE

This dissertation is organized as follows: Chapter 1 and Chapter 2 establish the foundation by introducing the research motivation and providing a structured overview of related work. In Chapter 3, we explore how to learn good word embeddings. Chapter 4 focuses on sentence embedding learning with multimodal signals. In Chapter 5, we present a framework for modeling semantic relatedness. In Chapter 6, we provide a systematical analysis of multilingual in-context learning. Finally, Chapter 7 summarizes our key contributions and discusses future directions. An overview for each chapter is shown as follows:

- In Chapter 2, we provide a comprehensive review of related work, covering the development of text embeddings, advancements in multilingual and multimodal learning, techniques for language model adaptation, and an outline of evaluation benchmarks.

- In Chapter 3, we explore key factors that influence the quality of word embeddings through a controlled study on the learning algorithm, corpus size, and training parameters. Our experiments span six languages in low-resource settings and uncover universal and language-specific trends. To identify the impact of each

parameter, we adopt linear regression to quantify their effects. Our findings provide practitioners with actionable guidelines for further improvements.

- In Chapter 4, we propose a novel sentence embedding learning approach that integrates visual and textual information via a multimodal contrastive objective. Our experiments on various semantic similarity tasks demonstrate its superior performance. We also analyze the geometric properties of the embedding space, uncovering the inner workings that explain its effectiveness.
- In Chapter 5, we develop a framework for measuring semantic relatedness for 14 African and Asian languages. To mitigate data scarcity in under-represented languages, we use machine translation as a data augmentation strategy. We also explore various training paradigms and source language selection methods to facilitate supervised learning and cross-lingual transfer performance.
- In Chapter 6, we conduct a comprehensive analysis of multilingual in-context learning, examining the role of demonstrations across models, tasks, and languages. Our experiments involve 5 models, 9 datasets, and 56 typologically diverse languages. Through a granular analysis, we reveal misunderstandings in this field, highlighting the need for fair evaluation and a deeper understanding of in-context learning.
- In Chapter 7, we conclude by summarizing the key contributions of this dissertation and outlining promising directions for future research.

1.2 PUBLICATIONS

This section presents my research contributions to multiple projects. Main publications (Section 1.2.1) are primary works that form the content of this dissertation. Additional publications (Section 1.2.2) are collaborative works to which I contributed my expertise. All publications are listed in chronological order.

1.2.1 *Main Publications*

[1] MCSE: Multimodal Contrastive Learning of Sentence Embeddings (Zhang et al., 2022)
***Miaoran Zhang**, Marius Mosbach, David Ifeoluwa Adelani, Michael A. Hedderich, Dietrich Klakow*
In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies (NAACL 2022)
URL: https://aclanthology.org/2022.naacl-main.436/

Abstract: Learning semantically meaningful sentence embeddings is an open problem in natural language processing. In this work, we propose a sentence embedding learning approach that exploits both visual and textual information via a multimodal contrastive objective. Through experiments on a variety of semantic textual similarity tasks, we demonstrate that our approach consistently improves the performance across various datasets and pre-trained encoders. In particular, combining a small amount of multimodal data with a large text-only corpus, we improve the state-of-the-art average Spearman's correlation by 1.7%. By analyzing the properties of the textual embedding space, we show that our model excels in aligning semantically similar sentences, providing an explanation for its improved performance.

[2] **AAdaM at SemEval-2024 Task 1: Augmentation and Adaptation for Multilingual Semantic Textual Relatedness (Zhang et al., 2024b)**
Miaoran Zhang, *Mingyang Wang, Jesujoba O. Alabi, Dietrich Klakow*
In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval 2024)
URL: https://aclanthology.org/2024.semeval-1.114/

Abstract: This paper presents our system developed for the SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian Languages. The shared task aims at measuring the semantic textual relatedness between pairs of sentences, with a focus on a range of under-represented languages. In this work, we propose using machine translation for data augmentation to address the low-resource challenge of limited training data. Moreover, we apply task-adaptive pre-training on unlabeled task data to bridge the gap between pre-training and task adaptation. For model training, we investigate both full fine-tuning and adapter-based tuning, and adopt the adapter framework for effective zero-shot cross-lingual transfer. We achieve competitive results in the shared task: our system performs the best among all ranked teams in both subtask A (supervised learning) and subtask C (cross-lingual transfer).

[3] **The Impact of Demonstrations on Multilingual In-Context Learning: A Multidimensional Analysis (Zhang et al., 2024a)**
Miaoran Zhang, *Vagrant Gautam, Mingyang Wang, Jesujoba O. Alabi, Xiaoyu Shen, Dietrich Klakow, Marius Mosbach*
In Findings of the Association for Computational Linguistics: ACL 2024
URL: https://aclanthology.org/2024.findings-acl.438/

Abstract: In-context learning is a popular inference strategy where large language models solve a task using only a few labeled demonstrations without needing any parameter updates. Although there have been extensive studies on English in-context learning, multilingual in-context learning remains under-explored, and we lack an in-depth understanding of the role of demonstrations in this context. To address this gap, we conduct a multidimensional analysis of multilingual in-context learning, experimenting with 5 models from different model families, 9 datasets covering classification and generation tasks, and 56 typologically diverse languages. Our results reveal that the effectiveness of demonstrations varies significantly across models, tasks, and languages. We also find that strong instruction-following models including Llama 2-Chat, GPT-3.5, and GPT-4 are largely insensitive to the quality of demonstrations. Instead, a carefully crafted template often eliminates the benefits of demonstrations for some tasks and languages altogether. These findings show that the importance of demonstrations might be overestimated. Our work highlights the need for granular evaluation across multiple axes towards a better understanding of in-context learning.

1.2.2 *Additional Publications*

[1] Preventing Author Profiling through Zero-Shot Multilingual Back-Translation (Adelani et al., 2021)

*David Adelani, **Miaoran Zhang**, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, Dietrich Klakow*

In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)

URL: https://aclanthology.org/2021.emnlp-main.684/

Abstract: Documents as short as a single sentence may inadvertently reveal sensitive information about their authors, including e.g. their gender or ethnicity. Style transfer is an effective way of transforming texts in order to remove any information that enables author profiling. However, for a number of current state-of-the-art approaches the improved privacy is accompanied by an undesirable drop in the down-stream utility of the transformed data. In this paper, we propose a simple, zero-shot way to effectively lower the risk of author profiling through multilingual back-translation using off-the-shelf translation models. We compare our models with five representative text style transfer models on three datasets across different domains. Results from both an automatic and a human evaluation show that our approach achieves the best overall performance while requiring no training data. We are able to lower the adversarial prediction of gender and race by up to 22% while retaining 95% of the original utility on downstream tasks.

[2] Knowledge Base Index Compression via Dimensionality and Precision Reduction (Zouhar et al., 2022)

*Vilém Zouhar, Marius Mosbach, **Miaoran Zhang**, Dietrich Klakow*

In Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge @ ACL 2022

URL: https://aclanthology.org/2022.spanlp-1.5/

Abstract: Recently neural network based approaches to knowledge-intensive NLP tasks, such as question answering, started to rely heavily on the combination of neural retrievers and readers. Retrieval is typically performed over a large textual knowledge base (KB) which requires significant memory and compute resources, especially when scaled up. On HotpotQA we systematically investigate reducing the size of the KB index by means of dimensionality (sparse random projections, PCA, autoencoders) and numerical precision reduction. Our results show that PCA is an easy solution that requires very little data and is only slightly worse than autoencoders, which are less stable. All methods are sensitive to pre- and post-processing and data should always be centered and normalized both before and after dimension reduction. Finally, we show that it is possible to combine PCA with using 1bit per dimension. Overall we achieve (1) 100× compression with 75%, and (2) 24× compression with 92% original retrieval performance.

[3] **A Lightweight Method to Generate Unanswerable Questions in English (Gautam et al., 2023)**

*Vagrant Gautam, **Miaoran Zhang**, Dietrich Klakow*

In Findings of the Association for Computational Linguistics: EMNLP 2023

URL: https://aclanthology.org/2023.findings-emnlp.491/

Abstract: If a question cannot be answered with the available information, robust systems for question answering (QA) should know *not* to answer. One way to build QA models that do this is with additional training data comprised of unanswerable questions, created either by employing annotators or through automated methods for unanswerable question generation. To show that the model complexity of existing automated approaches is not justified, we examine a simpler data augmentation method for unanswerable question generation in English: performing antonym and entity swaps on answerable questions. Compared to the prior state-of-the-art, data generated with our training-free and lightweight strategy results in better models (+1.6 F1 points on SQuAD 2.0 data with BERT-large), and has higher human-judged relatedness and readability. We quantify the raw benefits of our approach compared to no augmentation across multiple encoder models, using different amounts of generated data, and also on TydiQA-MinSpan data (+9.3 F1 points with BERT-large). Our results establish swaps as a simple but strong baseline for future work.

[4] **Human Speech Perception in Noise: Can Large Language Models Paraphrase to Improve It? (Chingacham et al., 2024)**

*Anupama Chingacham, **Miaoran Zhang**, Vera Demberg, Dietrich Klakow*

In Proceedings of the 1st Human-Centered Large Language Modeling Workshop @ ACL 2024

URL: https://aclanthology.org/2024.hucllm-1.1/

Abstract: Large Language Models (LLMs) can generate text by transferring style attributes like formality resulting in formal or informal text.However, instructing LLMs to generate text that when spoken, is more intelligible in an acoustically difficult environment, is an under-explored topic.We conduct the first study to evaluate LLMs on a novel task of generating acoustically intelligible paraphrases for better human speech perception in noise.Our experiments in English demonstrated that with standard prompting, LLMs struggle to control the non-textual attribute, i.e., acoustic intelligibility, while efficiently capturing the desired textual attributes like semantic equivalence. To remedy this issue, we propose a simple prompting approach, prompt-and-select, which generates paraphrases by decoupling the desired textual and non-textual attributes in the text generation pipeline.Our approach resulted in a 40% relative improvement in human speech perception, by paraphrasing utterances that are highly distorted in a listening condition with babble noise at signal-to-noise ratio (SNR) -5 dB. This study reveals the limitation of LLMs in capturing non-textual attributes, and our proposed method showcases the potential of using LLMs for better human speech perception in noise.

[5] **Exploring the Effectiveness and Consistency of Task Selection in Intermediate-Task Transfer Learning (Lin et al., 2024)**
*Pin-Jie Lin, **Miaoran Zhang**, Marius Mosbach, Dietrich Klakow*
In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) @ ACL 2024
URL: https://aclanthology.org/2024.acl-srw.24/

Abstract: Identifying beneficial tasks to transfer from is a critical step toward successful intermediate-task transfer learning. In this work, we experiment with 130 source-target task combinations and demonstrate that the transfer performance exhibits severe variance across different source tasks and training seeds, highlighting the crucial role of intermediate-task selection in a broader context. We compare four representative task selection methods in a unified setup, focusing on their effectiveness and consistency. Compared to embedding-free methods and text embeddings, task embeddings constructed from fine-tuned weights can better estimate task transferability by improving task prediction scores from 2.59% to 3.96%. Despite their strong performance, we observe that the task embeddings do not consistently demonstrate superiority for tasks requiring reasoning abilities. Furthermore, we introduce a novel method that measures pairwise token similarity using maximum inner product search, leading to the highest performance in task prediction. Our findings suggest that token-wise similarity is better predictive for predicting transferability compared to averaging weights.

[6] **Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice? (Zhu et al., 2024a)**
*Dawei Zhu, Pinzhen Chen, **Miaoran Zhang**, Barry Haddow, Xiaoyu Shen, Dietrich Klakow*
In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)
URL: https://aclanthology.org/2024.emnlp-main.24/

Abstract: Traditionally, success in multilingual machine translation can be attributed to three key factors in training data: large volume, diverse translation directions, and high quality. In the current practice of fine-tuning large language models (LLMs) for translation, we revisit the importance of these factors. We find that LLMs display strong translation capability after being fine-tuned on as few as 32 parallel sentences and that fine-tuning on a single translation direction enables translation in multiple directions. However, the choice of direction is critical: fine-tuning LLMs with only English on the target side can lead to task misinterpretation, which hinders translation into non-English languages. Problems also arise when noisy synthetic data is placed on the target side, especially when the target language is well-represented in LLM pre-training. Yet interestingly, synthesized data in an under-represented language has a less pronounced effect. Our findings suggest that when adapting LLMs to translation, the requirement on data quantity can be eased but careful considerations are still crucial to prevent an LLM from exploiting unintended data biases.

[7] **Unveiling the Key Factors for Distilling Chain-of-Thought Reasoning (Chen et al., 2025)**

*Xinghao Chen, Zhijing Sun, Wenjin Guo, **Miaoran Zhang**, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, Xiaoyu Shen*

In Findings of the Association for Computational Linguistics: ACL 2025

URL: https://aclanthology.org/2025.findings-acl.782/

Abstract: Large Language Models (LLMs) excel in reasoning tasks through Chain-of-Thought (CoT) prompting. However, CoT prompting greatly increases computational demands, which has prompted growing interest in distilling CoT capabilities into Small Language Models (SLMs). This study systematically examines the factors influencing CoT distillation, including the choice of granularity, format and teacher model. Through experiments involving four teacher models and seven student models across seven mathematical and commonsense reasoning datasets, we uncover three key findings: (1) Unlike LLMs, SLMs exhibit a *non-monotonic* relationship with granularity, with stronger models benefiting from finer-grained reasoning and weaker models performing better with simpler CoT supervision; (2) CoT format significantly impacts LLMs but has *minimal* effect on SLMs, likely due to their reliance on supervised fine-tuning rather than pretraining preferences; (3) Stronger teacher models do *NOT* always produce better student models, as diversity and complexity in CoT supervision can outweigh accuracy alone. These findings emphasize the need to tailor CoT strategies to specific student model, offering actionable insights for optimizing CoT distillation in SLMs.

[8] **AFRIDOC-MT: Document-level MT Corpus for African Languages (Alabi et al., 2025)**

*Jesujoba O. Alabi, Israel Abebe Azime, **Miaoran Zhang**, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O Ademuyiwa, Andrew Caines, Dietrich Klakow*

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)
URL: https://aclanthology.org/2025.emnlp-main.1413/

Abstract: This paper introduces AFRIDOC-MT, a document-level multi-parallel translation dataset covering English and five African languages: Amharic, Hausa, Swahili, Yorùbá, and Zulu. The dataset comprises 334 health and 271 information technology news documents, all human-translated from English to these languages. We conduct document-level translation benchmark experiments by evaluating the ability of neural machine translation (NMT) models and large language models (LLMs) to translate between English and these languages, at both the sentence and pseudo-document levels, the outputs being realigned to form complete documents for evaluation. Our results indicate that NLLB-200 achieves the best average performance among the standard NMT models, while GPT-4o outperforms general-purpose LLMs. Fine-tuning selected models leads to substantial performance gains, but models trained on sentences struggle to generalize effectively to longer documents. Furthermore, our analysis reveals that some LLMs exhibit issues such as under-generation, over-generation, repetition of words and phrases, and off-target translations, specifically for translation into African languages.

2

BACKGROUND

This chapter provides a structured overview of key concepts and methodologies central to this dissertation. We begin with text embeddings in Section 2.1, covering both word embeddings and sentence embeddings, which enable machines to understand the meaning of discrete linguistic units for text processing. Next, we discuss multilingual and multimodal learning in Section 2.2, focusing on generalized models across languages, as well as cross-lingual transfer and multimodal alignment techniques. We then present representative language model adaptation strategies in Section 2.3, including fine-tuning and in-context learning, which tailor pre-trained models to various downstream tasks. Finally, we introduce evaluation benchmarks for model performance in Section 2.4, including both embedding evaluation and downstream evaluation.

2.1 TEXT EMBEDDINGS

Text embeddings are continuous vector representations of linguistic units, ranging from words and phrases to entire sentences or documents, that capture their semantic meaning. By mapping discrete textual elements to a vector space, they enable machine learning systems to process and understand textual data effectively. Early works focus on word embeddings, which have evolved from static to contextual embeddings. In static embeddings, such as Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b), a word is always associated with the same vector regardless of the context. In contrast, more recent contextual embeddings, produced by neural language models such as BERT (Devlin et al., 2019), assign different vectors to the same word depending on the context in which it appears. Building on this, sentence embeddings (Hill et al., 2016; Reimers and Gurevych, 2019) represent entire sentences as fixed-dimensional vectors and capture higher-level semantics for tasks such as retrieval and clustering.

Formally, let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ represent a sequence of tokens, where x_i is the i-th token in the sequence. An embedding model $\mathcal{M}$ parameterized by θ processes this sequence to produce text representations:

$$\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\} = \mathcal{M}_\theta(\boldsymbol{x}), \boldsymbol{v} = \mathrm{Agg}(\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\}). \tag{2.1}$$

Here, $\boldsymbol{h}_i$ represents the output hidden state for token x_i, which are then aggregated into a single sequence-level embedding $\boldsymbol{v}$ by a function $\mathrm{Agg}(\cdot)$, using methods such as special-token selection (e.g., `[CLS]`)

or mean/max pooling. With this formalization, in this section, we first review representative work on word embeddings, and then discuss major research directions and techniques in sentence embedding learning.

2.1.1 *Word Embeddings*

Word embedding learning is based on the distribution hypothesis (Harris, 1954), which posits that words in similar contexts have similar meanings. Early work in this area can be broadly categorized into two types: prediction-based methods and count-based methods. A representative example of prediction-based methods is Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b), which introduces two efficient frameworks for learning word embeddings: continuous bag-of-words (CBOW) and skip-gram (SG). Both variants use shallow neural networks for word prediction, while CBOW predicts a target word based on its surrounding context words and SG predicts the context words given a target word. FastText (Bojanowski et al., 2017) extends the skip-gram model by incorporating character n-grams, allowing the model to capture morphological information and better handle rare words. As a well-known count-based approach, GloVe (Pennington et al., 2014) constructs a word co-occurrence matrix from the entire corpus and then factorizes it to produce embeddings that explicitly preserve global statistical patterns.

Despite their wide application, early word embedding methods are inherently *static* – they assign a single, context-independent vector for each word, making it difficult to distinguish between different senses of polysemous words or to capture contextual nuances. In contrast, recent approaches have shifted toward *contextual* embeddings driven by the innovations in language models, where word representations are dynamically generated based on their surrounding context. For example, ELMo (Peters et al., 2018) produces contextualized representations by combining the hidden states of a bidirectional LSTM trained via language modeling. With the advent of the transformer architecture (Vaswani et al., 2017), a significant breakthrough in this area came with BERT (Devlin et al., 2019), which uses masked language modeling (MLM) and bidirectional attention to produce contextualized embeddings. Subsequent models, such as RoBERTa (Liu et al., 2019), further improve the model through training and architectural optimizations. However, for all their ability to capture transferable linguistic properties, these pre-trained language models (PLMs) yield suboptimal off-the-shelf embeddings for similarity or retrieval tasks (Reimers and Gurevych, 2019; Li et al., 2020). This gap highlights the need for approaches to produce semantically meaningful embeddings from these general-purpose models.

2.1.2 *Sentence Embeddings*

Learning semantic representations for longer text units, such as sentences, is often more challenging than for individual words due to increased compositional complexity and variability. To derive semantically meaningful sentence embeddings, existing methods can be broadly categorized into unsupervised methods, supervised methods, and multi-stage methods. Beyond this taxonomy, in this section, we also highlight *contrastive learning* (Chen et al., 2020a) – a dominant representation learning regime in recent studies, and discuss the emerging frontier of adapting decoder-only language models as embedding models.

UNSUPERVISED METHODS Unsupervised methods can be grouped into traditional distributional approaches, geometry-aware techniques, and self-supervised learning objectives. Traditional methods draw inspiration from word embedding learning. A simple baseline is to average word vectors such as GloVe to form sentence representations. SkipThought (Kiros et al., 2015) and Sent2Vec (Pagliardini et al., 2018) extend the Word2Vec objectives to learn embeddings at the sentence level rather than the word level. Focusing on the geometry of the embedding space, BERT-flow (Li et al., 2020) observes that BERT produces highly anisotropic representations and proposes transforming them into an isotropic Gaussian distribution using normalizing flows. Su et al. (2021) apply a simple whitening transformation to achieve similar isotropy, yielding competitive performance while reducing dimensionality and improving efficiency. In addition, self-supervised objectives leverage the intrinsic structure of the text. For example, IS-BERT (Zhang et al., 2020b) maximizes the mutual information between a global sentence embedding and its local contextual representations. SimCSE (Gao et al., 2021b) and Mirror-BERT (Liu et al., 2021) adopt unsupervised contrastive learning using identical or slightly perturbed text pairs as positives without requiring additional data. Trans-Encoder (Liu et al., 2022a) alternates between unsupervised bi-encoder and cross-encoder formulations and use each to generate pseudo-labels for the other.

SUPERVISED METHODS Supervised methods rely on labeled data to learn high-quality sentence embeddings. Conneau et al. (2017) demonstrate that natural language inference (NLI) data provides an effective supervision signal for training universal sentence embeddings. This is further supported by USE (Cer et al., 2018a), which augments unsupervised learning with supervised training on NLI. With the rise of transformer-based models, Sentence-BERT (Reimers and Gurevych, 2019) introduces the concept of *sentence transformers*, proposing a BERT-based siamese framework and training the model on NLI and semantic

textual similarity (STS) datasets. SimCSE (Gao et al., 2021b) is a well-known contrastive learning framework available in both unsupervised and supervised settings; in the supervised setting, entailment pairs in NLI examples are treated as positives and contradiction pairs as hard negatives. In addition to NLI, Wieting et al. (2020) show that bilingual and back-translation corpora provide useful supervision for learning semantic sentence embeddings. Supervised methods typically achieve higher performance than their unsupervised counterparts, but their reliance on annotated data can limit scalability and applicability in real-world scenarios.

MULTI-STAGE METHODS Multi-stage methods build on the strengths of both unsupervised and supervised learning by sequential training phases. GTE (Li et al., 2023b) follows this paradigm by combining large-scale unsupervised pre-training with a subsequent supervised contrastive fine-tuning stage over a diverse mixture of datasets. E5 (Wang et al., 2024a) adopts a similar two-stage design, but relies on weak supervision at scale by conducting contrastive pre-training on a massive curated text-pair corpus.

CONTRASTIVE LEARNING Contrastive learning is a representation learning paradigm that trains models to map semantically similar inputs (i.e., "*positive pairs*") to nearby points in the embedding space, while pushing dissimilar inputs (i.e., "*negative pairs*") farther apart. A prevalent direction focuses on creating positive pairs without manual data annotation, often through data augmentation. For instance, Wu et al. (2020) and Yan et al. (2021) apply various text augmentation techniques to construct positive pairs, such as token shuffling, deletion and substitution. There are other methods exploring minimal or targeted augmentation: SimCSE (Gao et al., 2021b) uses standard dropout as implicit noise, DiffCSE (Chuang et al., 2022) uses stochastic masking and sampling from a masked language model to capture semantic nuances, and PromptBERT (Jiang et al., 2022) incorporates prompt-based templates with denoising.

Beyond unsupervised positive-pair construction, a key concern is the quality of the negative samples. Standard approaches often rely on in-batch or random negatives, which may introduce false negatives and bias. Zhou et al. (2022) thus mitigate this by weighting instances and generating noise-based negatives, while Wang et al. (2021) constructs semantically informed negative examples to improve robustness against small adversarial perturbations.

Another direction focuses on generating synthetic data for contrastive learning. Approaches such as DINO (Schick and Schütze, 2021a) and SynCSE (Zhang et al., 2023a) use large LLMs to generate entire datasets of labeled sentence pairs from scratch. This enables

smaller models to acquire high-quality embeddings without any additional supervision.

LLMS AS EMBEDDERS The advent of generative large language models (LLMs) has marked a significant milestone in NLP, achieving state-of-the-art performance on many tasks. However, because of their autoregressive nature, their potential as text embedding models has been far less explored than that of bidirectional encoder-only models. This gap has motivated a growing body of work on adapting powerful decoder-only LLMs into effective text embedders. LLM2Vec (BehnamGhader et al., 2024) modifies the causal attention mask to enable bidirectional attention and fine-tunes the model with masked next token prediction and unsupervised contrastive learning. Wang et al. (2024b) rely on diverse synthetic data to fine-tune decoder-only LLMs using standard contrastive loss. NV-Embed (Lee et al., 2025) combines architectural designs, training procedures, and curated datasets to significantly enhance the performance of LLM as an embedding model. Muennighoff et al. (2025) introduce generative representational instruction tuning, where a LLM is trained to handle both generative and embedding tasks by distinguishing between them via instructions. In parallel, prompt-based methods adapt LLMs without any parameter updates. For example, Lei et al. (2024) guide LLMs to produce meaningful embeddings through a series of carefully designed prompts that address multiple representational aspects. Springer et al. (2025) show that simply repeating the input twice and extracting embeddings from the second occurrence yields strong embeddings.

2.2 MULTILINGUAL AND MULTIMODAL LEARNING

2.2.1 *Multilingual Models*

Existing NLP models are predominantly trained on English-centric corpora, leading to imbalanced multilingual capabilities (Joshi et al., 2020). Bridging the gap between English and non-English languages is essential for supporting a wider range of language users and communities. In recent years, multilingual language models have received increasing attention, offering the advantage of handling many languages within a unified framework (Qin et al., 2024). We introduce multilingual models relevant to this dissertation as follows.

LaBSE (Feng et al., 2022) is a multilingual encoder model that produces sentence embeddings for over 100 languages. It employs a BERT-style architecture with a 12-layer, 110M-parameter Transformer. All languages are processed using a customized tokenizer created from the training data with WordPiece algorithm (Johnson et al., 2017). Training includes two stages: first, encoder pre-training with MLM and

translation language modeling; second, bi-encoder fine-tuning with a translation ranking objective using additive margin softmax.

AfroXLMR (Alabi et al., 2022; Adelani et al., 2024) provides a series of multilingual encoder models adapted from XLM-R (Conneau et al., 2020) for African languages. The model is fine-tuned using MLM on a large collection of monolingual corpora from a mix of languages, including widely spoken African languages as well as a few high-resource languages such as English, Arabic, and French. It also examines the effect of vocabulary reduction: removing non-African sub-tokens from the original tokenizer leads to a 50% reduction in model size while still remaining competitive in performance.

XGLM (Lin et al., 2022) is a multilingual decoder model trained on a large-scale corpus of 500 billion tokens for 30 languages, with upsampling of low-resource languages to improve data balance. A joint vocabulary for all languages is created using SentencePiece (Kudo and Richardson, 2018). The model architecture follows GPT-3 (Brown et al., 2020), with additional embedding parameters to accommodate the expanded vocabulary. XGLM is trained using causal language modeling (CLM) and is available in sizes up to 7.5B parameters.

BLOOMZ (Muennighoff et al., 2023b) is a multilingual decoder model adapted from the pre-trained BLOOM (Workshop et al., 2023). The training data is xP3, a multilingual and multitask dataset based on the P3 (Sanh et al., 2022) taxonomy and extended with non-English datasets. It includes 46 languages and a variety of task types. The model is fine-tuned via multitask instruction tuning, where natural language prompts are added to examples to specify the task. BLOOMZ is released in sizes ranging from 560M to 176B parameters.

mT0 (Muennighoff et al., 2023b) is a multilingual encoder-decoder model adapted from mT5 (Xue et al., 2021). It is fine-tuned in the same manner as BLOOMZ but uses a different base model and architecture. The model is available in sizes ranging from 300M to 13B parameters. Results show that mT0 (13B) outperforms BLOOMZ (176B) on held-out tasks, despite having an order of magnitude fewer parameters. This is likely due to the encoder-decoder architecture and the longer pre-training of mT5.

2.2.2 *Cross-Lingual Transfer*

Although multilingual models facilitate multi- and cross-lingual applications for languages seen during training, their performance on low-resource or unseen languages remains limited because such languages are under-represented or absent from the training data (Lai et al., 2023; Ojo et al., 2025). This gap motivates the development

of cross-lingual transfer techniques to effectively transfer knowledge from high-resource to low-resource languages.

An interesting direction is the use of modular approaches, which enhance zero-shot cross-lingual transfer while mitigating interference and overfitting. For example, MAD-X (Pfeiffer et al., 2020) consists of language- and task-specific adapters. At inference time, the source language adapter is replaced with a target language adapter, while the task adapter trained on labeled data in the source language is retained. Similarly, Ansell et al. (2022) propose composable sparse fine-tuning, which learns sparse task- and language-specific masks that can be applied without modifying the model architecture.

When a small amount of low-resource data is available, few-shot learning offers a practical alternative. Lauscher et al. (2020) demonstrate that inexpensive few-shot fine-tuning can substantially mitigate transfer loss and improve performance across various NLP tasks. Zhu et al. (2024a) also show that LLMs exhibit strong translation capabilities after fine-tuning on as few as 32 parallel sentences, and that fine-tuning in a single translation direction can transfer translation capability to multiple directions.

Yet, the mechanisms underlying cross-lingual knowledge transfer are poorly understood. Philippy et al. (2023) identify several contributing factors, including shared subword representations, linguistic similarity, pre-training data distribution, and fine-tuning dynamics. Choenni et al. (2023) analyze cross-lingual sharing from the influential data perspective, showing that sharing increases during fine-tuning, with languages supporting one another through both reinforcing and complementary roles. These studies indicate that effective cross-lingual transfer arises from a complex interplay of linguistic, representational, and training factors.

2.2.3 *Multimodal Alignment*

The rapid advancement of NLP models, particularly LLMs, has driven the development of multimodal models (Yin et al., 2024). By integrating the multiple modalities through which humans interact with the world, these models serve as an important step toward general-purpose intelligent systems.

A key component of multimodal learning is multimodal alignment — the process of mapping data from different modalities into a shared representation space, thereby establishing semantic correspondences across modalities (Li and Tang, 2025). This alignment is often achieved through various pre-training strategies. For example, contrastive learning, as used in CLIP (Radford et al., 2021), pulls matched image–label pairs together while separating mismatched pairs. LXMERT (Tan and Bansal, 2019) and UNITER (Chen et al., 2020b) combine multiple cross-modal objectives, such as image–text matching and visual question

answering, to model rich interactions across modalities and enhance the model's ability to understand and reason over multimodal data.

Aligned representations provide a foundation for grounding language in perceptual knowledge, which can further improve language understanding (Bisk et al., 2020). By incorporating information from other modalities, NLP systems can capture aspects of meaning and context that are difficult to infer from text alone. Approaches such as Vokenization (Tan and Bansal, 2020) and VidLanKD (Tang et al., 2021) demonstrate that grounding language in visual or video information can enhance textual representations and improve performance on diverse NLP tasks.

2.3 LANGUAGE MODEL ADAPTATION

Before the emergence of generative LLMs, modern NLP models are developed under a "pre-train then fine-tune" paradigm. In this pipeline, a model is first pre-trained on large-scale unlabeled corpora to acquire broad knowledge, resulting in a general-purpose pre-trained language model (PLM). Different model architectures typically adopt different unsupervised pre-training objectives: encoder-only models such as BERT (Devlin et al., 2019) use masked language modeling (MLM); decoder-only models such as GPT-2 (Radford et al., 2019) rely on causal language modeling (CLM), i.e., next token prediction; and encoder–decoder models such as T5 (Raffel et al., 2020) employ unified text-to-text objectives including span corruption. Although these pre-training tasks endow models with broad knowledge, they do not directly align with the target tasks to solve. Therefore, a second stage of task-specific fine-tuning on labeled datasets is often required to adapt the pre-trained models to specific downstream tasks.

Driven by the scaling law (Kaplan et al., 2020), the growing size of models and the availability of massive text corpora facilitated the rise of LLMs (Touvron et al., 2023; OpenAI et al., 2023, *inter alia*). While LLMs still begin with the large-scale pre-training stage, their subsequent adaptation phase has evolved. The focus has shifted away from task-specific fine-tuning towards post-training techniques, including instruction tuning and reinforcement learning from human feedback (Ouyang et al., 2022), to improve the model's general instruction-following and alignment. Importantly, LLMs have emerged with strong in-context learning capabilities (Brown et al., 2020): they can adapt to new tasks through task descriptions and a few demonstrations provided at inference time without any parameter updates.

In this section, we introduce these two major adaptation approaches in detail: (1) fine-tuning, including vanilla fine-tuning (i.e., full parameter fine-tuning) and parameter-efficient fine-tuning, and (2) in-context learning, an inference-time strategy that enables task adaptation without modifying model parameters.

2.3.1 *Fine-Tuning*

Vanilla fine-tuning is the standard method to adapt a PLM to various downstream tasks (Devlin et al., 2019; Dodge et al., 2020). In this case, all model parameters are optimized by a task-specific loss function computed over a labeled dataset. Formally, given a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with inputs x_i and corresponding labels y_i (which may be scalar or vector targets), vanilla fine-tuning seeks to minimize the empirical risk:

$$\mathcal{L}(\theta; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i), \tag{2.2}$$

where $f_\theta(\cdot)$ denotes the output of the model parameterized by θ, $\ell(\cdot)$ is the task-specific loss function (e.g., cross-entropy for classification, mean squared error for regression, or contrastive loss for embedding learning), and $\mathcal{L}(\theta; \mathcal{D})$ is the average loss over the training dataset. The optimal parameters θ^* are then obtained by minimizing this loss across the entire parameter space:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta; \mathcal{D}). \tag{2.3}$$

Since vanilla fine-tuning updates all model parameters, it becomes computationally and memory-intensive as model size and the number of tasks increase. To address this limitation, recent work has proposed a variety of parameter-efficient fine-tuning (PEFT) methods. The key idea is to freeze the pre-trained model parameters θ and updates only a small, task-specific set of (extra) parameters ϕ. The optimization objective is to minimize the empirical risk with respect to only ϕ:

$$\mathcal{L}(\phi; \theta, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_{\theta,\phi}(x_i), y_i). \tag{2.4}$$

The small set of trainable parameters is often introduced through plug-and-play modules, such as adapters (Houlsby et al., 2019), prefix embeddings (Li and Liang, 2021), or low-rank matrices (Hu et al., 2022). For example, in Houlsby et al. (2019), two adapter modules are inserted into every Transformer layer, and each adapter consists of a bottleneck which contains few parameters relative to the attention and feedforward layers in the original transformer model (typically 0.5%-5% of the full model). PEFT methods have shown comparable performance to full parameter fine-tuning, and can even surpass it in low-data or out-of-domain scenarios (Chen et al., 2022a; Whitehouse et al., 2024).

2.3.2 *In-Context Learning*

In-context learning (ICL) is an inference strategy in which models perform tasks without any parameter updates (Brown et al., 2020). In

contrast to fine-tuning methods, a model generates predictions for a test example by simply prefixing it with a handful of labeled examples (i.e., in-context demonstrations).

A prevalent research focus is finding high-quality demonstrations. Liu et al. (2022b) propose retrieving semantically similar demonstrations from an existing labeled data pool for each test input. Expanding on this, SU et al. (2023) introduce a two-step framework: they first select a diverse, representative subset of examples from unlabeled data for annotation, and then retrieve examples from this small annotated pool at test time. Complementing these instance-dependent methods, task-level selection approaches identify demonstrations that are broadly effective across all test cases. For example, Chang and Jia (2023) evaluate an example's efficiency by measuring its average dev-set ICL performance when combined with random examples. Another line of research uses LLMs to generate in-context demonstrations by themselves (Kim et al., 2022), further reducing reliance on external datasets.

However, several studies suggest that the exact content of the demonstrations may be less critical than expected, emphasizing the need for a deeper understanding of what actually drives ICL performance. Min et al. (2022c) show that randomly replacing labels in demonstrations barely hurts performance across several tasks. Similarly, Yoo et al. (2022) find that the importance of correct input-label mappings varies substantially with experimental configuration. Wang et al. (2023a) further reveal that chain-of-thought (CoT) reasoning performance can be largely retained even when demonstrations contain invalid reasoning steps.

Beyond the quality of demonstrations, analyses focusing on the quantity of demonstrations also point to unexpected robustness: Zhang et al. (2023b) observe that demonstrations may offer limited benefit or even degrade performance on code-switching tasks, and Chen et al. (2023) show that a single positive demonstration in ICL can significantly outperforms multiple demonstrations. These findings highlight that it is crucial to understand what truly drives ICL performance.

2.4 EVALUATION BENCHMARKS

Evaluating NLP models is crucial for assessing their effectiveness and reliability. In this dissertation, we consider two types of evaluation datasets. Section 2.4.1 covers embedding evaluation, which directly measures the quality of word and sentence representations and reveals how well models capture semantic meaning. Section 2.4.2 describes downstream evaluation, which evaluates model performance on practical, real-world tasks, with a particular focus on multilingual benchmarks.

2.4.1 *Embedding Evaluation*

RG-65 (Rubenstein and Goodenough, 1965) is a pioneering English word similarity dataset comprising 65 noun pairs annotated by 15 human annotators. Each pair is assigned a similarity score on a [0, 4] scale, with higher values indicating greater semantic similarity. The dataset has been widely used as a benchmark for evaluating distributional semantic models.

WordSim-353 (Finkelstein et al., 2002) is also an English word similarity dataset containing 353 word pairs, annotated by 13–16 human judges. Annotators rate the similarity of each pair on a [0, 10] scale, where 0 indicates completely unrelated words, and 10 indicates very closely related or identical words.

RW (Luong et al., 2013) is the Stanford rare word similarity dataset. English word pairs are selected based on their low frequency in Wikipedia, ensuring coverage of infrequent lexical items. Similarity judgments are collected via Amazon Mechanical Turk, with 10 annotators scoring each pair on a [0, 10] scale. After quality filtering, the final dataset comprises 2,034 word pairs.

MEN (Bruni et al., 2012) is an English word similarity dataset containing 3,000 word pairs. Unlike previous datasets, which rely on absolute scores from multiple annotators, MEN uses a relative judgment approach: each pair is ranked by a single annotator against 50 alternative pairs, and a raw score out of 50 is computed based on how often the pair is judged more related than alternatives. This score is then normalized to the [0, 1] range.

SimLex-999 (Hill et al., 2015) is an English word similarity dataset containing 999 noun, verb, adjective, and adverbs pairs. Each pair is rated on a [0, 6] scale via Amazon Mechanical Turk. Notably, SimLex-999 focuses explicitly on similarity rather than association or relatedness, ensuring that pairs of related but dissimilar concepts receive low scores. The dataset also includes a mix of concrete and abstract words, along with independent ratings of concreteness for each pair, enabling more fine-grained analyses.

FS300 (Venekoski and Vankka, 2017) is a Finnish word similarity dataset translated from SimLex-999. The translations are validated by two fluent bilingual researchers, and only words with a single unambiguous sense in both languages are included in the dataset to ensure comparability, resulting in 300 word pairs. Finnish participants rate each pair on a [0, 10] scale, where 0 indicates no similarity, and 10 indicates that the words are synonymous.

jSIM (Karpinska et al., 2018) is a revised Japanese word similarity dataset based on the original dataset (Sakaizawa and Komachi, 2018),

containing 4,851 word pairs in total. The words are sourced from a lexical simplification dataset and annotated by 10 native speakers on a [0, 10] similarity scale. To reduce ambiguity introduced by Japanese tokenization, the dataset is re-categorized into three versions: a full version with all pairs, a tokenized version containing only words recognizable after tokenization, and an unambiguous version with only recognized, unambiguous words. We use the tokenized version which contains 2,790 word pairs.

SemR-11 (Barzegar et al., 2018) is a multilingual word similarity dataset covering 11 languages. It consists of 15,917 word pairs, which are translated from English datasets MC-30 (Miller and Charles, 1991), RG-65, WordSim-353, and SimLex-999. All translations are carefully produced and reviewed by professional translators. The dataset assumes that these translations preserve the similarity scores of the original English annotations.

STS12–STS16 (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015; Agirre et al., 2016) are a series of datasets from SemEval shared tasks designed to evaluate semantic textual similarity (STS) between a pair of sentences. Each dataset contains sentence pairs annotated with rated similarity scores by human judges on a [0, 5] scale. These datasets cover multiple domains, including news, tweets, question–answer pairs, machine-translated text, and video or image descriptions. The STS-16 additionally introduce cross-lingual sentence pairs. These STS datasets have been widely used for research on sentence-level similarity and semantic representations.

STS-B (Cer et al., 2017), the STS Benchmark, is a curated dataset selected from the English STS shared task data between 2012 and 2017. It includes sentence pairs from three genres: news, captions, and forums. The dataset contains a total of 8,628 sentence pairs, split into training, development, and test sets.

SICK (Marelli et al., 2014) is a dataset designed to evaluate distributional semantic models. It contains approximately 10,000 English sentence pairs sourced from caption and video-description corpora. Each pair is annotated for two tasks: semantic relatedness, rated on a [0, 5] scale, and textual entailment, labeled as entailment, contradiction, or neutral. Human judgments are collected via crowd-sourcing. We use the semantic relatedness annotations, commonly referred to as the SICK-R dataset.

For all datasets, we use Spearman's rank correlation coefficient (ρ) between human ratings and predicted similarity scores (e.g., cosine similarity of two embeddings) as the evaluation metric. It is defined as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_i is the difference between the ranks of the i-th word or sentence pair, and n is the total number of pairs. This metric is widely used for evaluating semantic similarity tasks, as it measures the extent to which predicted scores preserve the relative ordering of human judgments.

2.4.2 *Downstream Evaluation*

XNLI (Conneau et al., 2018) is a cross-lingual natural language inference (NLI) dataset. It is created by extending MNLI (Williams et al., 2018) to 15 languages using professional translators, including low-resource languages such as Swahili and Urdu. It provides examples for three-way NLI classification: given a premise and a hypothesis, the task is to predict whether the relationship is entailment, contradiction, or neutral.

IndicNLI (Aggarwal et al., 2022) is an NLI dataset that follows the same data format as XNLI but focuses on the Indic language family. It is created by translating the English XNLI examples (premises and hypotheses) into 11 Indo-Aryan languages using IndicTrans (Ramesh et al., 2022). The high translation quality of IndicTrans has been demonstrated through manual human validation and the automatic metric BERTScore (Zhang et al., 2020a).

PAWS-X (Yang et al., 2019) is a paraphrase identification dataset. It is constructed by translating the Wikipedia portion of the original English PAWS (Zhang et al., 2019a) corpus into 6 other languages. The development and test sets are translated by professional translators, while the training set is translated using a neural machine translation service. Given a pair of sentences, the goal is to classify whether they are paraphrases or not.

XCOPA (Ponti et al., 2020) is a dataset for causal commonsense reasoning. It is created by carefully translating and re-annotating the validation and test sets of the English COPA (Gordon et al., 2012) (Choice of Plausible Alternatives) dataset into 11 languages from distinct families. The task is framed as a two-way classification problem: given a premise sentence, the goal is to select one from two alternatives that is more likely to have a causal relationship with the premise.

XStoryCloze (Lin et al., 2022) is also a commonsense reasoning dataset. It is created by professionally translating the validation split of the English StoryCloze dataset (Mostafazadeh et al., 2016) into 10 typologically diverse languages. Given a four-sentence story (the context) and

two alternative endings, the task is to choose the correct ending that completes the story.

AfriSenti (Muhammad et al., 2023) is a sentiment analysis dataset for African languages. It covers 14 languages and contains examples sourced from tweets. These tweets are collected via the Twitter Academic API using location-based and vocabulary-based strategies. Language identification is then performed using pre-existing tools, native speakers, and language models. Given a sentence, the task is to classify its sentiment as positive, negative, or neutral.

XQuAD (Artetxe et al., 2020) is a cross-lingual question answering (QA) dataset. It is created by translating the English SQuAD v1.1 (Rajpurkar et al., 2016) dataset, based on Wikipedia, into 10 languages. Both the context paragraphs and the questions are translated by professional human translators. As an extractive question answering task, the goal is to identify the answer span in a given context paragraph for a provided question.

TyDiQA (Clark et al., 2020) is also a QA dataset covering 11 typologically diverse languages. Human annotators create questions from short prompts, and each question is matched to a Wikipedia article using a Google search restricted to that language's Wikipedia. Annotators then identify the passage within the article that best answers the question. The dataset also includes a simplified version TyDiQA-GoldP, where only the answer passage is provided instead of the full Wikipedia article.

MAFAND (Adelani et al., 2022) is a sentence-level machine translation (MT) dataset covering 16 African languages. Sentences are collected from local newspapers published in English and French, then translated into other languages by professional translators. Quality control is performed by native speakers, who review and correct problematic translations. The task is to translate between 16 African languages and English or French.

Regarding evaluation metrics, tasks including NLI, paraphrase identification, and commonsense reasoning are evaluated using standard classification accuracy, which measures the proportion of correctly predicted instances. For QA datasets, evaluation is based on the token-level F1 score, which measures the overlap between predicted and ground-truth answer tokens. Let TP denote the number of true positives (tokens correctly predicted), FP the number of false positives (tokens predicted but not in ground truth), and FN the number of false negatives (tokens in ground truth but missing in prediction). The F1 score is then defined as the harmonic mean of precision and recall:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \tag{2.5}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.6}$$

For the MT task, we use chrF++ (Popović, 2017), an extension of the character n-gram F-score metric chrF (Popović, 2015). chrF++ combines character-level n-grams with short word-level n-grams (unigrams and bigrams). The chrF F-score is defined as:

$$\text{chrF}_\beta = (1 + \beta^2) \frac{P_{\text{ng}} \cdot R_{\text{ng}}}{\beta^2 \cdot P_{\text{ng}} + R_{\text{ng}}}, \tag{2.7}$$

where P_{ng} and R_{ng} are the averaged n-gram precision and recall over orders $n = 1$ to N. The parameter β controls the importance weight between recall and precision. In chrF++, the recommended configuration uses character n-grams up to order 6 along with word unigrams and bigrams, and sets β as 2.

Compared with traditional word-level metrics such as BLEU (Papineni et al., 2002), chrF and chrF++ exhibit more robust performance, particularly for morphologically rich languages, including many African languages.

3

WHAT MAKES GOOD WORD EMBEDDINGS

This chapter explores key factors that contribute to obtaining good static word embeddings. While transformer-based models can produce advanced contextualized embeddings, traditional word embedding techniques remain relevant in certain domains such as healthcare and embedded systems due to their simplicity and efficiency. Many word embedding algorithms have been proposed and gained popularity so far, yet questions remain about how certain factors influence embedding quality in a language-specific context. To close this gap, we take a systematic approach to examine the impact of the learning algorithm, corpus size, and training parameters on the quality of embeddings for diverse languages. Our study also quantifies the contribution of each parameter through a linear regression analysis. Finally, we discuss the implications of our findings for practitioners and suggest strategies for further improvements.[1]

3.1 INTRODUCTION

Word embeddings, which are fixed-length distributional representations of words, play a crucial role in the field of NLP. The traditional paradigm for generating such representations is based on the distributional hypothesis (Harris, 1954) that words appearing in similar contexts have similar meanings. This has given rise to the development of well-known word embedding models such as Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and GloVe (Pennington et al., 2014). Embedding models are prevalent not only in English but also in many other languages. For instance, FastText (Bojanowski et al., 2017) provides pre-trained embeddings for 157 languages, showcasing the widespread use of word embeddings.

Recent efforts have primarily focused on developing new mathematical models and techniques. While claims are often made that proposed methods can improve performance on evaluation benchmarks, the impact of each design choice on embedding quality remains unclear. To understand this, Lai et al. (2016) provide a unified comparison of several aspects of word embedding training. Lison and Kutuzov (2017) and Hellrich et al. (2019) analyze the role of specific training parameters (e.g., context window) in detail which is often overlooked in previous literature. Given the broad use of word embeddings across language communities, it is also important to fairly compare these

1 This chapter is based on an exploration project supervised by Dana Ruiter and Dietrich Klakow.

system designs on diverse languages not limited to English, particularly when the training data is limited – a common challenge for low-resource languages. This leads us to the question: *How do different factors impact word embedding learning in low-data multilingual settings?*

To answer it, we conduct a comprehensive evaluation of the impact of multiple design choices across six languages with distinct linguistic features: Arabic, Chinese, English, Finnish, French, and Japanese. The word embedding models and key factors considered are introduced in Section 3.3. All experimental details are provided in Section 3.4. In Section 3.5, we report the main findings from our extensive experiments (18,000 in total) and analyze the results using linear regression analysis. Further discussion and insights for potential improvements are presented in Section 3.6.

3.2 RELATED WORK

Representing words as dense vectors has a long history and there are many studies on learning word embeddings (Almeida and Xexéo, 2023). Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) is among the best known models and has two variants: continuous bag-of-words (CBOW) and skip-gram (SG). CBOW predicts the current word from its surrounding context, while SG does the reverse by predicting the context given a word. FastText (Bojanowski et al., 2017) incorporates character *n*-grams into the skip-gram model, allowing the model to capture morphological information and better handle rare words. In contrast to these prediction-based models, GloVe (Pennington et al., 2014) employs a count-based approach based on the global co-occurrence statistics of the entire corpus. Moreover, Jiang et al. (2018) and Jungmaier et al. (2020) investigate learning better word embeddings for low-resource languages.

The quality of learned word embeddings can be influenced by various factors. For example, Levy et al. (2015) reveal that certain system design choices and hyperparameter tuning have a substantial impact on performance. Lai et al. (2016) examine several critical aspects of word embedding training, including the model, the corpus, and the training parameters. The importance of hyperparameters is further emphasized in Caselles-Dupré et al. (2018) when applying word embeddings to recommendation systems. Zooming in on specific parameters, Lison and Kutuzov (2017) and Ri and Tsuruoka (2020) conduct a systematic analysis on the role of context window, while Hellrich et al. (2019) investigate the impact of downsampling strategies. Our study is inspired by these works but emphasizes a holistic evaluation across diverse languages, including the exploration of simulated low-resource scenarios.

3.3 PRELIMINARIES

3.3.1 *Word Embedding Algorithms*

In this study, we consider three representative word embedding algorithms, including two methods from Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) – **CBOW** and **SG**, as well as **SVD-PPMI**$_\lambda$ (Jungmaier et al., 2020), which we introduce below.

CBOW The training objective is to obtain word representations that effectively predict the current word given its context. More formally, given a target word w_t and its surrounding context $C = \{w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}\}$ with a window size c, the objective is to minimize the negative log-likelihood:

$$\mathcal{L} = -\log P(w_t|C) = -\log \frac{\exp(\boldsymbol{h}_{w_t}^\top \boldsymbol{h}_C)}{\sum_{w' \in \mathcal{V}} \exp(\boldsymbol{h}_{w'}^\top \boldsymbol{h}_C)} \tag{3.1}$$

Here, $\mathcal{V}$ represents the vocabulary, $\boldsymbol{h}_{w_t}$ denotes the vector representation of w_t, and $\boldsymbol{h}_C$ denotes the context vector, which is calculated as the average of the vector representations of the context words.

SG Instead of predicting the a given word based on its context, the training objective of SG is to maximize the prediction of context words based on the center word. The objective is defined as:

$$\mathcal{L} = -\sum_{\substack{j=c \\ j\neq 0}}^{c} \log P(w_{t+j}|w_t) = -\sum_{\substack{j=c \\ j\neq 0}}^{c} \log \frac{\exp(\boldsymbol{h}_{w_{t+j}}^\top \boldsymbol{h}_{w_t})}{\sum_{w' \in \mathcal{V}} \exp(\boldsymbol{h}_{w'}^\top \boldsymbol{h}_{w_t})} \tag{3.2}$$

The original CBOW and SG models (Mikolov et al., 2013a) use *hierarchical softmax* to approximate full softmax for computational efficiency. Mikolov et al. (2013b) introduce *negative sampling* as an effective alternative, which we adopt for both CBOW and SG in this study.

SVD-PPMI$_\lambda$ This approach is built upon Positive Pointwise Mutual Information (PPMI) embeddings (Bullinaria and Levy, 2007). The standard PPMI of two words w and c is defined as:

$$\text{PPMI}(w, c) = \max(\text{PMI}(w, c), 0) = \max(\log \frac{P(w, c)}{P(w)P(c)}, 0) \tag{3.3}$$

where probabilities are estimated from the co-occurrence matrix by maximum likelihood estimation. To reduce the bias towards rare words, SVD-PPMI$_\lambda$ applies Dirichlet smoothing to obtain non-zero probabilities for unseen events by adding a small pseudo count λ in likelihood estimation. This is followed by truncated Singular Value Decomposition (SVD) on the smoothed PPMI matrix for dimensionality reduction.

Factor	Values
algorithm	{CBOW, SG, SVD-PPMI$_\lambda$}
corpus size	{50k, 100k, 200k, 400k, 800k}
context window size	{1, 3, 5, 7, 9, 11, 13, 15, 17, 19}
subsampling threshold	{1e-1, 1e-2, 1e-3, 1e-4}
minimum word count	{1, 2, 3, 4, 5}

Table 3.1: Summary of influential factors and their selected values.

3.3.2 *Influential Factors*

The quality of word representations depends not only on the word embedding algorithm and training data, but previous work (Lai et al., 2016; Caselles-Dupré et al., 2018; Ri and Tsuruoka, 2020) also shows that certain training parameters can significantly influence the performance of the generated embeddings. Table 3.1 presents an overview of all factors that we explored in this systematic study.

For the training parameters, context window size determines how many adjacent words for a center word are taken into account when estimating the word representations. Subsampling (Mikolov et al., 2013b) is a technique of diluting highly frequent words. Each word is discarded with a probability $p = 1 - \sqrt{t/f}$, where f is the word frequency and t is a pre-defined threshold. The lower the threshold, the more frequent words are downsampled. Another common design is to ignore words that are rare in the training corpus, which is controlled by the minimum word count parameter.

We study monolingual embeddings for six languages in total, including Arabic, Chinese, English, Finnish, French, and Japanese. These languages are selected for their diverse scripts and morphological typologies. Our goal is to examine how various factors affect word embeddings for different languages, and see if we can quantify the effects across linguistic differences.

3.4 EXPERIMENTAL SETUP

3.4.1 *Training Data*

We use Wikipedia dumps from 2019[2] as the training corpus. At the preprocessing stage, language identification is performed using Polyglot[3], followed by punctuation normalization for Latin languages using Moses[4]. We apply language-specific toolkits for tokenization: Moses

2 https://dumps.wikimedia.org
3 https://github.com/aboSamoor/polyglot
4 https://github.com/moses-smt/mosesdecoder

for Latin languages, Farasa[5] for Arabic, Jieba[6] for Chinese, and Mecab[7] for Japanese. Sentences are shuffled and split into datasets in varying sizes. Our final corpus sizes are relatively small, as we hope to simulate scenarios with limited resources.

3.4.2 *Evaluation*

The quality of word embeddings is evaluated by word similarity and relatedness tests. These tests are commonly used intrinsic evaluations, aiming at measuring how well the embeddings capture the semantic relationship between a pair of words. For English, our evaluation datasets are RG-65 (Rubenstein and Goodenough, 1965), WordSim-353 (Finkelstein et al., 2002), SimLex-999 (Hill et al., 2015), MEN (Bruni et al., 2012), and RW (Luong et al., 2013). For Chinese, French, and Arabic, we use the multilingual translations of MC-30 (Miller and Charles, 1991), RG-65, and WordSim-353 provided in SemR-11 Barzegar et al. (2018). We evaluate Finnish embeddings on FS300 (Venekoski and Vankka, 2017) and Japanese embeddings on jSIM (Karpinska et al., 2018). The evaluation metric is Spearman's rank correlation coefficient ($\times 100$) between the cosine similarities of word embeddings and human judgment scores. For each language, we report the macro average score over their evaluation datasets.

3.4.3 *Implementation Details*

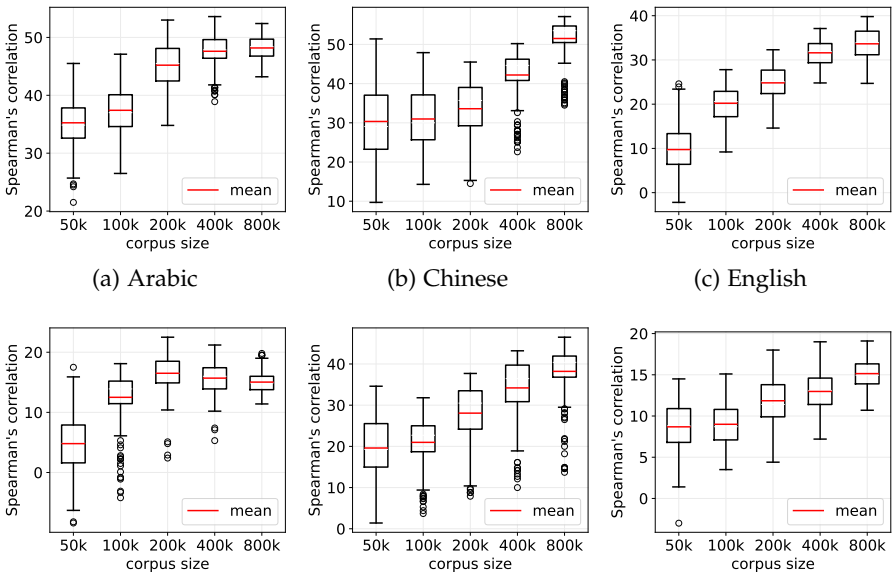
We use the word2vec toolkit[8] from Google and the official codebase of SVD-PPMI$_\lambda$[9] for our experiments. Besides the three training parameters that we investigate, we use default values for all other parameters. Particularly, the word embedding dimensionality is set to 100 and the smoothing factor λ in SVD-PPMI$_\lambda$ is set to 1e-4. During evaluation, we skip the out-of-vocabulary (OOV) words for simplicity.

3.5 RESULTS AND ANALYSIS

3.5.1 *Empirical Results*

We conduct a total of 18,000 experiments to analyze the impact of various factors. First, we show the performance variances across different training parameters in Figure 3.1 for CBOW, Figure 3.2 for SG, and Figure 3.3 for SVD-PPMI$_\lambda$. There are 200 experimental results for each language and corpus size, and we can observe that **the**

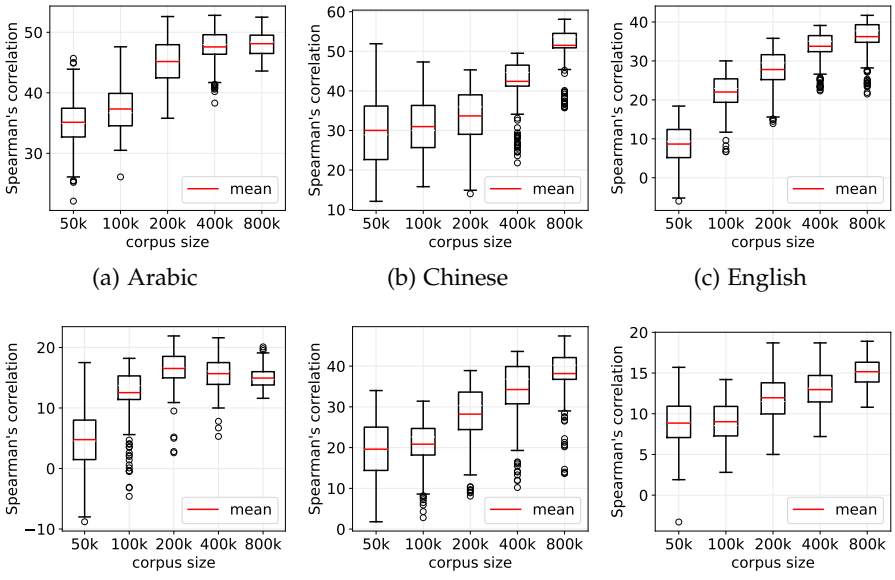
5 https://github.com/MagedSaeed/farasapy
6 https://github.com/fxsjy/jieba
7 https://github.com/SamuraiT/mecab-python3
8 https://code.google.com/archive/p/word2vec/
9 https://github.com/jungmaier/dirichlet-smoothed-word-embeddings



(a) Arabic (b) Chinese (c) English

(d) Finnish (e) French (f) Japanese

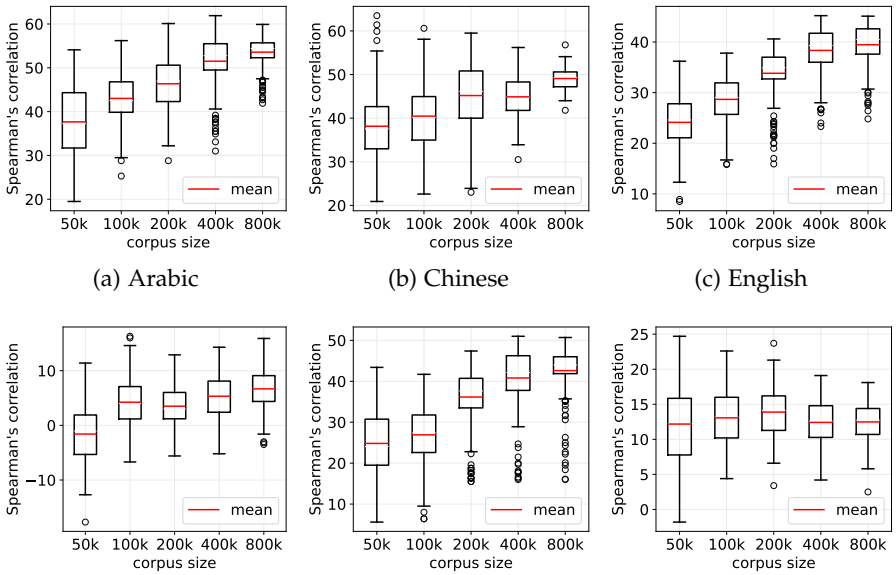
Figure 3.1: Performance of CBOW with different training parameters.



(a) Arabic (b) Chinese (c) English

(d) Finnish (e) French (f) Japanese

Figure 3.2: Performance of SG with different training parameters.



(a) Arabic (b) Chinese (c) English

(d) Finnish (e) French (f) Japanese

Figure 3.3: Performance of SVD-PPMI$_\lambda$ with different training parameters.

performance varies considerably with different parameter choices, especially when the training data is limited. For instance, in the case of Chinese embeddings trained with CBOW, the performance gap between the worst and best parameter sets can reach up to 40% when the corpus size is 50k. With an 800k corpus size, the gap is reduced to approximately 10%, which remains a significant discrepancy. This pattern is observed consistently across languages and learning algorithms.

We also find that **increasing the data scale improves embedding quality in most cases**. However, for Finnish, the performance quickly reaches a saturation point with diminishing returns, and for Japanese, SVD-PPMI$_\lambda$ yields similar results on average regardless of the training data size. Overall, the performance of Finnish and Japanese is relatively lower compared to other languages. This may be attributed to their agglutinative nature and more complex morphological structures, which pose additional challenges for embedding learning.

To get further intuition about the effectiveness of the three learning algorithms, we present their average results in Table 3.2. The results show that the performance of the two Word2Vec methods, CBOW and SG, is quite similar across all setups. **SVD-PPMI$_\lambda$ consistently demonstrates the best performances, with Finnish being the only exception**. Given our small training corpus sizes, these results highlight the effectiveness of SVD-PPMI$_\lambda$ under low-resource experimental settings.

In Figure 3.4, we show the detailed results for individual evaluation datasets in English and Japanese, **highlighting the performance discrepancies across different subsets**. For English, the performance on

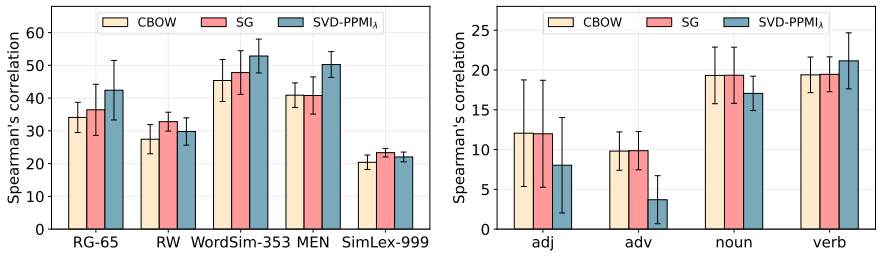
Method	Arabic	Chinese	English	Finnish	French	Japanese	Avg. ↑
50k							
CBOW	35.25	30.34	9.74	**4.80**	19.60	8.69	18.07
SG	35.13	30.02	8.65	4.77	19.61	8.85	17.84
SVD-PPMI$_\lambda$	**37.66**	**38.15**	**24.13**	-1.59	**24.82**	**12.18**	**22.56**
100k							
CBOW	37.41	30.98	20.21	12.50	20.98	9.00	21.85
SG	37.33	30.98	22.02	**12.54**	20.85	9.02	22.12
SVD-PPMI$_\lambda$	**43.02**	**40.47**	**28.66**	4.24	**26.92**	**13.06**	**26.06**
200k							
CBOW	45.20	33.59	24.82	16.50	28.07	11.85	26.67
SG	45.16	33.67	27.78	**16.53**	28.22	11.96	27.22
SVD-PPMI$_\lambda$	**46.35**	**45.18**	**33.82**	3.50	**36.16**	**13.89**	**29.82**
400k							
CBOW	47.61	42.20	31.61	**15.69**	34.19	**12.98**	30.71
SG	47.58	42.40	33.76	15.67	34.25	12.96	31.11
SVD-PPMI$_\lambda$	**51.50**	**44.89**	**38.33**	5.33	**40.81**	12.42	**32.21**
800k							
CBOW	48.18	**51.51**	33.64	**15.03**	38.21	15.14	33.62
SG	48.12	51.49	36.23	14.95	38.17	**15.16**	**34.02**
SVD-PPMI$_\lambda$	**53.57**	49.08	**39.46**	6.68	**42.61**	12.49	33.98

Table 3.2: Comparison of three embedding learning algorithms. We report the mean score over all parameter choices for each language and corpus size.

the WordSim-353 dataset is significantly higher than on SimLex-999. Notably, WordSim-353 rates how related two words are according to its annotation guidelines, whereas SimLex-999 explicitly focuses on word similarity. As an example, the word pair `(clothes, closet)` receives a human rating of 1.96 in SimLex-999 but 8.00 in WordSim-353, indicating that SimLex-999 might be more challenging – it requires models to capture similarity independently of relatedness or association, which is often overlooked by co-occurrence patterns in corpora. Meanwhile, for Japanese, models more easily learn embeddings for nouns and verbs, which carry concrete meanings, compared to adjectives and adverbs. These findings underscore the influence of dataset characteristics and linguistic features on word embedding learning.

3.5.2 *Correlation Analysis*

In the previous section, we have systematically explored the performance variations with varying parameters. To further quantify the impact of the three training parameters, in this section, we conduct a correlation analysis using Ordinary Least Square (OLS) regression,



(a) English (b) Japanese

Figure 3.4: Performance on different evaluation subsets with a corpus size of 800k. The error bar is the standard deviation over all training parameter choices.

assuming that there exists a linear relationship between the training parameters and evaluation performance. Our goal is to identify the strength and direction of this relationship and better understand how each parameter contributes to the overall performance.

Let x_c denote the context window size, x_s denote the subsampling threshold, and x_m denote the minimum word count. First, we normalize x_c and x_m using min-max normalization. For x_s, we scale it to $-\log_{10}(x_s)$, meaning that a larger normalized value corresponds to more subsampling. The OLS regression model can be formulated as:

$$y = \beta_0 + \beta_c x_c + \beta_s x_s + \beta_m x_m. \tag{3.4}$$

Here, y is the evaluation performance (i.e., Spearman's correlation score) to be fitted, and β_c, β_s, and β_m are the coefficients to be estimated for the three explanatory variables, with β_0 representing the model intercept. The goodness-of-fit of the regression model is measured by the R^2 statistic:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.5}$$

where n is the total number of observations, y_i and $\hat{y}_i$ are the observed and predicted values for the i-th observation, and $\bar{y}$ is the mean of the observed values. The closer R^2 is to 1.0, the better the model explains the variation in the dependent variable y.

We select the experimental results for a training corpus size of 800k and perform OLS regression for three algorithms across six languages, with 200 observations for each regressor. As shown in Table 3.3, English and Japanese consistently achieve high R^2 values, with $R^2 \geq 0.71$ for English and $R^2 \geq 0.65$ for Japanese, suggesting that a significant proportion of the variance in their performances can be linearly explained by these training parameters. In contrast, Arabic exhibits very low R^2 values ($R^2 \leq 0.37$) but shows a high baseline performance ($\beta_0 \geq 45.78$), indicating that these training parameters struggle to make a substantial impact on its performance.

Language	Method	β_0	β_c	β_s	β_m	$R^2 \uparrow$
English	CBOW	25.99	5.84	6.26	3.21	0.85
	SG	27.87	11.14	2.51	3.08	0.71
	SVD-PPMI$_\lambda$	30.96	9.91	3.02	4.07	0.77
Japanese	CBOW	15.96	1.99	0.01$_\times$	-3.64	0.65
	SG	15.92	2.08	-0.05$_\times$	-3.55	0.66
	SVD-PPMI$_\lambda$	8.04	1.06	2.04	5.80	0.71
French	CBOW	26.67	14.89	5.63	2.57	0.68
	SG	26.63	14.72	5.82	2.55	0.68
	SVD-PPMI$_\lambda$	33.69	11.11	4.42	2.33$_\times$	0.43
Finnish	CBOW	16.70	-3.55	-1.10	0.73	0.54
	SG	16.89	-3.55	-1.07	0.73	0.53
	SVD-PPMI$_\lambda$	-0.65$_\times$	8.45	2.07	4.14	0.67
Chinese	CBOW	44.76	12.63	-0.45$_\times$	1.31$_\times$	0.57
	SG	44.99	12.25	-0.60$_\times$	1.35$_\times$	0.56
	SVD-PPMI$_\lambda$	47.00	1.33	0.31$_\times$	2.53	0.17
Arabic	CBOW	45.90	1.30	0.60$_\times$	2.66	0.29
	SG	45.78	1.26	0.67$_\times$	2.74	0.29
	SVD-PPMI$_\lambda$	49.24	5.86	0.12$_\times$	2.68	0.37

Table 3.3: Results of regression coefficients and R^2 statistics. $\times$ indicates that the estimation is *not* significant at $p = 0.01$. The color in the R^2 column reflects its magnitude, with darker cells indicating higher values.

Looking at the coefficients, the context window size (β_c) shows a positive relationship with performance in most languages, particularly in English and French. A larger context window generally improves performance, although its impact varies by language. However, the coefficients for the subsampling threshold (β_s) and minimum word count (β_m) are not statistically significant in many cases, meaning that these two parameters have a relatively weaker impact on performance.

3.6 DISCUSSION

Our findings in the previous section demonstrate that the choice of algorithm, training corpus size, and various training parameters can largely affect the word embedding quality for many languages, particularly in low-resource settings. These findings offer practical guidance for practitioners to prioritize specific factors and guide their optimization efforts. In this section, we open up new questions and provide further insights, which are discussed below.

EVALUATION PITFALLS Evaluating word embeddings is a fundamental challenge and there is still no consensus within the community about which evaluation methods should be used (Schnabel et al., 2015; Bakarov, 2018). We adopt a widely used intrinsic evaluation method – word similarity and relatedness. However, to comprehensively assess the quality of learned embeddings, it is also important to consider other intrinsic tasks, such as word analogy and categorization, and performance on downstream tasks (i.e., extrinsic evaluation). Additionally, while we omit OOV words from our evaluation, it is important to explore strategies to address the OOV issue, which contributes to the overall expressiveness of learned word embeddings.

CONNECTING STATIC AND CONTEXTUAL EMBEDDINGS While contextual embeddings from transformer-based models have gained prominence in NLP tasks, static embeddings remain valuable in resource-constrained environments, given the high costs of training and inferencing modern language models (Bender et al., 2021). To take advantage of the strengths of both, some studies explore combining static and contextual embeddings to improve downstream performance and efficiency (Gupta and Jaggi, 2021; Alghanmi et al., 2020). In addition, techniques initially developed for static embeddings, such as interpretability methods and post-processing, have been helpful to better understand and further improve contextual embeddings (Bommasani et al., 2020; Sajjad et al., 2022).

FROM WORD EMBEDDINGS TO SENTENCE EMBEDDINGS This chapter focuses on learning embeddings for words, and a natural question that follows is how to obtain semantic representations for longer pieces of text, such as sentences. Early works derive sentence embeddings by learning composition operators that map word vectors to sentence vectors (Wieting et al., 2016; Arora et al., 2017; Pagliardini et al., 2018). More recently, Sentence-BERT (Reimers and Gurevych, 2019) adapts transformer-based models to generate meaningful sentence embeddings, marking a breakthrough in this field. Building on this, advances such as SimCSE (Gao et al., 2021b) leverage contrastive learning to produce highly effective sentence embeddings. In the next chapter, we will investigate how to further improve sentence embedding learning for transformer-based models.

3.7 CONCLUSION

In this chapter, we present a systematic study of the key factors that influence the quality of word embeddings. We explore three types of static embeddings for six languages, revealing that the choice of algorithm, corpus size, and training parameters can significantly impact performance, although to varying degrees across languages. Our

results also show that SVD-PPMI$_\lambda$ generally produces better embeddings than the other two algorithms. Moreover, the variation in performance and optimal configurations highlights the importance of customized parameter tuning, as certain settings are more effective for specific languages. We end by discussing how to properly evaluate word embeddings, the connection between different kinds of embeddings, which provides insights into a deeper understanding and further improvements.

4 MULTIMODAL SENTENCE EMBEDDING LEARNING

The previous chapter explored how to obtain good word embeddings by examining various influencing factors. A natural question that follows is how to get high quality sentence embeddings – representations that can capture the compositional structure of text and convey richer semantic meaning, which is still an open problem in NLP. While current sentence embedding learning methods rely primarily on textual data, we argue that incorporating visual information offers critical advantages: it provides grounded world knowledge that can be useful to disambiguate linguistic representations and yield more robust sentence embeddings. In this chapter, we propose a novel sentence embedding learning approach that exploits both visual and textual information via a multimodal contrastive objective. Through experiments on a variety of semantic textual similarity tasks, we demonstrate that our approach consistently improves the performance across various datasets and pre-trained encoders. In particular, combining a small amount of multimodal data with existing text-only corpus, we improve the state-of-the-art average Spearman's correlation by 1.7%. By analyzing the properties of the textual embedding space, we show that our model excels in aligning semantically similar sentences, providing an explanation for its improved performance.[1]

4.1 INTRODUCTION

Sentence embedding learning, i.e., encoding sentences into fixed-length vectors that faithfully reflect the semantic relatedness among sentences, is a fundamental challenge in NLP. Despite the tremendous success of pre-trained kanguage models (PLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), it has been shown that the off-the-shelf sentence embeddings of PLMs without fine-tuning are even inferior to averaging Glove embeddings (Pennington et al., 2014) in terms of semantic similarity measure (Reimers and Gurevych, 2019). Hence, recent research (Li et al., 2020; Zhang et al., 2020b; Su et al., 2021) focuses on adjusting the original sentence embeddings derived from PLMs in an unsupervised manner. In particular, there has been growing interest in adopting contrastive learning objectives

1 This chapter is based on Zhang et al. (2022). As the first author, Miaoran Zhang proposed the idea, implemented the algorithm, conducted the experiments, and led the paper writing. The source code for this work is available on Github: https://github.com/uds-lsv/mcse.

to achieve this goal (Carlsson et al., 2020; Kim et al., 2021; Gao et al., 2021b).

Although purely text-based models have led to impressive progress, it remains an open question to what extent they capture the deeper notion of sentence meaning beyond the statistical distribution of texts, which lies outside of the text and is grounded in the real-world (Bender and Koller, 2020; Bisk et al., 2020). As a central part of the human perceptual experience, vision has been shown to be effective in grounding language models and improving performance on various NLP tasks (Zhang et al., 2019b; Bordes et al., 2019; Zhao and Titov, 2020). We hypothesize that using vision as supplementary semantic information can further promote sentence representation learning.

In this chapter, we propose MCSE, an approach for multimodal contrastive learning of sentence embeddings. To exploit both visual and textual information, we adopt the state-of-the-art contrastive sentence embedding framework SimCSE (Gao et al., 2021b) and extend it with a multimodal contrastive objective. In addition to the textual objective in SimCSE that maximizes agreement between positive sentence pairs, the multimodal objective maximizes agreement between sentences and corresponding images in a shared space. We conduct extensive experiments on standard STS benchmarks and show the effectiveness of MCSE across various datasets and pre-trained encoders. We find that, using a small amount of multimodal data in addition to a text-only corpus yields significant improvements on STS tasks. By analyzing the alignment and uniformity properties of the embedding space (Wang and Isola, 2020), we show that MCSE better aligns the semantically similar sentences while maintaining uniformity, providing an explanation for its superior performance.

4.2 RELATED WORK

Sentence representation learning. Early works learn composition operators that map word vectors to sentence vectors (Wieting et al., 2016; Arora et al., 2017; Pagliardini et al., 2018). With the rise of transformer-based models, recent works for sentence embedding learning can be categorized into supervised (Conneau et al., 2017; Cer et al., 2018b; Reimers and Gurevych, 2019; Wieting et al., 2020) and unsupervised approaches (Li et al., 2020; Carlsson et al., 2020; Su et al., 2021; Kim et al., 2021; Gao et al., 2021b; Liu et al., 2021; Yan et al., 2021). Supervised approaches mostly utilize supervision from annotated natural language inference data (Bowman et al., 2015; Williams et al., 2018) or paraphrase data (Wieting and Gimpel, 2018). Unsupervised approaches are able to make use of the intrinsic semantic information embedded in the natural language text corpus by adjusting the training objective to STS tasks, thereby eliminating the need for a costly annotation process. In particular, contrastive learning objective (Carls-

son et al., 2020; Kim et al., 2021; Gao et al., 2021b; Liu et al., 2021; Yan et al., 2021) regularizes the embedding space by pulling positive (i.e., semantically similar) sentences closer and pushing apart negatives, showcasing great effectiveness in capturing the semantic similarity among sentences. Our approach adopts the contrastive learning framework and is built on top of the current state-of-the-art approach (Gao et al., 2021b), further pushing the frontier of STS by leveraging multimodal semantic information.

Visually grounded representation learning. There are various works showing that grounding NLP models to the visual world can improve textual representation learning. Lazaridou et al. (2015) and Zablocki et al. (2018) learn word embeddings by aligning words to the visual entity or visual context. Kiela et al. (2018) ground sentence embeddings by predicting both images and alternative captions related to the same image. Bordes et al. (2019) enhance the Skip-Thought model (Kiros et al., 2015) by learning a grounded space that preserves the structure of visual and textual spaces. Recently, Tan and Bansal (2020) and Tang et al. (2021) train large scale language models with multimodal supervision from scratch with the goal of improving general language understanding. Different from the aforementioned works, we focus on learning visually grounded sentence embeddings by fine-tuning pre-trained models in a contrastive learning framework.

4.3 METHOD

To effectively integrate both visual and textual information, we adopt SimCSE (Gao et al., 2021b) as the textual baseline and extend it with a multimodal contrastive learning objective. We begin by introducing SimCSE in Section 4.3.1, a simple yet strong method in this line of research. The original SimCSE has two variants: supervised and unsupervised. In this chapter, we focus on the unsupervised SimCSE, as it is more closely aligned with real-world applications and minimizes the dependence on human annotations. In Section 4.3.2, we present our multimodal contrastive learning objective. It is worth noting that while our method extends SimCSE, it is broadly applicable to other contrastive learning frameworks for embedding learning.

4.3.1 *Unsupervised SimCSE*

The core idea of unsupervised SimCSE is contrastive self-supervised representation learning (Chen et al., 2020a), where the model takes an input sentence and predicts itself in a contrastive objective. To create meaningful positive pairs for contrastive learning, many works (Zhang et al., 2020b; Giorgi et al., 2021; Yan et al., 2021) take different views from data augmentation or different copies of models of the same sentence. In contrast, unsupervised SimCSE uses a simple idea by

taking different outputs of the same sentence from standard dropout. More specifically, given a collection of sentences $\{x_i\}_{i=1}^m$, we construct a positive pair for each input x_i by encoding it twice using different dropout masks: $\boldsymbol{h}_i^z = g_\phi(f_\theta(x_i, z))$ and $\boldsymbol{h}_i^{z'} = g_\phi(f_\theta(x_i, z'))$, where z and z' denote different dropout masks[2], $f_\theta(\cdot)$ is the output of a pre-trained language encoder such as BERT, and $g_\phi(\cdot)$ is the output of a projection head[3] on top of the `[CLS]` token. The training objective is:

$$\ell_i^{\mathrm{S}} = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_i^{z_i'})/\tau}}{\sum_{j=1}^N e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_j^{z_j'})/\tau}} , \tag{4.1}$$

where N is the size of the mini-batch, τ is a temperature parameter and $\mathrm{sim}(\boldsymbol{h}_1, \boldsymbol{h}_2)$ is the cosine similarity $\frac{\boldsymbol{h}_1^T \boldsymbol{h}_2}{\|\boldsymbol{h}_1\| \cdot \|\boldsymbol{h}_2\|}$. After training, the `[CLS]` token outputs of the language encoder are taken as the sentence embeddings.

4.3.2 *Multimodal Contrastive Learning*

In addition to the textual objective in unsupervised SimCSE, we introduce a multimodal objective within the contrastive learning framework. An overview of our proposed MCSE model is shown in Figure 4.1. Given a collection of sentence-image pairs $D = \{x_i, y_i\}_{i=1}^m$, firstly we map sentence x_i and image y_i into a shared space:

$$\boldsymbol{s}_i^z = g_{\phi_1}(f_\theta(x_i, z)),\ \boldsymbol{v}_i = g_{\phi_2}(f^v(y_i)) , \tag{4.2}$$

where $f^v(\cdot)$ is a pre-trained image encoder such as ResNet (He et al., 2016), which is fixed during training.[4] $g_{\phi_1}(\cdot)$ and $g_{\phi_2}(\cdot)$ are distinct projection heads for text and image modality respectively. To pull semantically close image-sentence pairs together and push away non-related pairs, we define the multimodal contrastive learning objective as:

$$\ell_i^{\mathrm{M}} = -\sum_{z \in \{z_i, z_i'\}} \log \frac{e^{\mathrm{sim}(\boldsymbol{s}_i^z, \boldsymbol{v}_i)/\tau'}}{\sum_{j=1}^N e^{\mathrm{sim}(\boldsymbol{s}_i^z, \boldsymbol{v}_j)/\tau'}} , \tag{4.3}$$

where τ' is a temperature parameter. Let λ denote the trade-off hyperparameter between two objectives, we formulate the final loss as:

$$\ell_i = \ell_i^S + \lambda \ell_i^M . \tag{4.4}$$

Our method further regularizes the sentence representation in a way that aligns with the image representation in the grounded space.

2 The standard dropout masks in Transformers are used.

3 There is a MLP pooler layer over `[CLS]` in BERT's implementation. Gao et al. (2021b) use it with re-initialization.

4 Early experiments show that fine-tuning the image encoder did not yield any improvements.

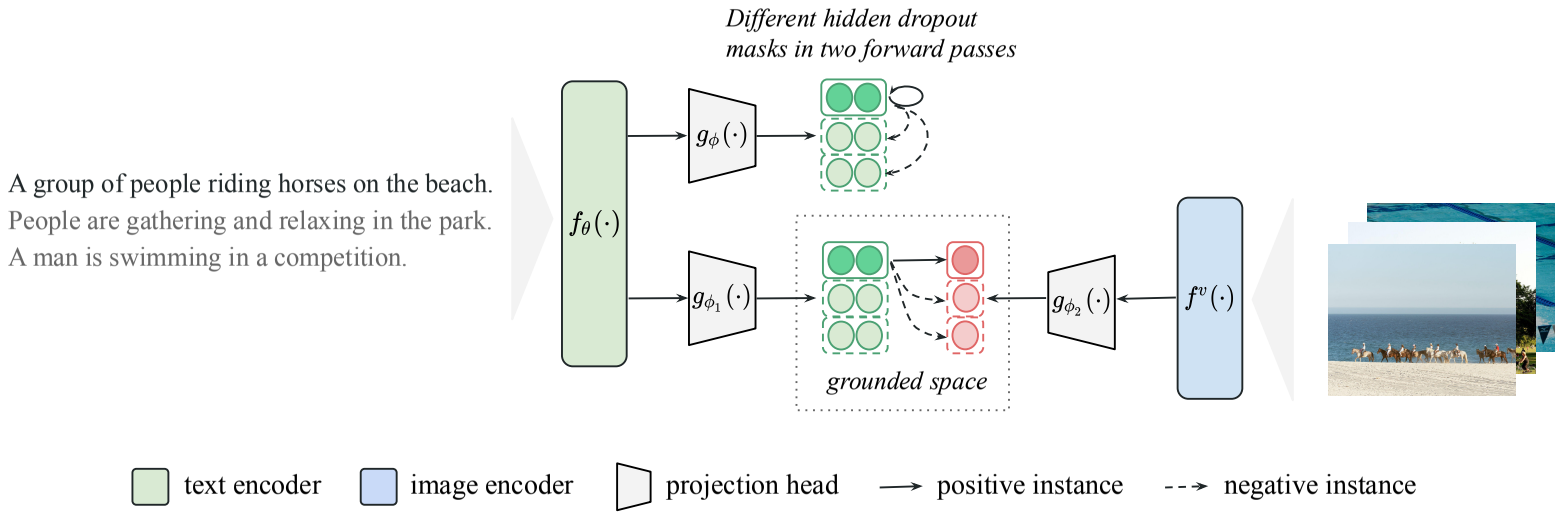


Figure 4.1: The overall architecture of MCSE. Compared to SimCSE, MCSE introduces a novel mutimodal contrastive learning objective computed in a shared space. For each input sentence, the model selects its paired image (e.g., from caption datasets) as the positive instance, while treating all other images in the same batch as negatives.

4.4 EXPERIMENTAL SETUP

4.4.1 *Training Datasets*

We use Flickr30k (Young et al., 2014) and MS-COCO (Lin et al., 2014) as our multimodal datasets. Flickr30k contains 29,783 training images and MS-COCO contains 82,783 training images. Each image is annotated with multiple captions and we randomly sample only one caption to create image-sentence pairs. Following Gao et al. (2021b), we use Wiki1M as the text-only corpus, which consists of 10^6 sentences randomly drawn from English Wikipedia. To examine the effectiveness of our approach under different data conditions, we explore five training settings: *wiki*, *wiki+flickr*, *wiki+coco*, *flickr*, and *coco*. Specifically, *wiki* denotes training on Wiki1M with the textual contrastive learning objective only. All other settings use both textual and multimodal objectives: *wiki+flickr* and *wiki+coco* refer to combined training on Wiki1M with Flickr30k and MS-COCO, where mini-batches are sampled proportionally to dataset sizes; *flickr* and *coco* refer to training solely on Flickr30k and MS-COCO.

4.4.2 *Evaluation*

The trained models are evaluated on 7 STS tasks: STS12 (Agirre et al., 2012), STS13 (Agirre et al., 2013), STS14 (Agirre et al., 2014), STS15 (Agirre et al., 2015), STS16 (Agirre et al., 2016), STS Benchmark (STS-B) (Cer et al., 2017), and SICK-Relatedness (SICK-R) (Marelli et al., 2014). Each of these datasets consists of a collection of sentence pairs and the goal is to predict a similarity score $\in [0, 1]$ for each sentence pair. Following Gao et al. (2021b), we report the Spearman's rank correlation coefficient ($\times 100$) between gold annotations and predicted scores in the "all" setting[5], i.e., for each task, we concatenate all the subsets and report the overall Spearman's correlation.

4.4.3 *Implementation Details*

LANGUAGE ENCODER We use BERT_{base} (Devlin et al., 2019) and RoBERTa_{base} (Liu et al., 2019) as language encoders. Our implementation is based on the Hugging Face Transformers library (Wolf et al., 2020).[6] We start from the checkpoints of `bert-base-uncased` and `roberta-base`, and fine-tune the pre-trained models using the contrastive objective function. We use the 768-dimensional `[CLS]` token

5 The other way is to calculate results for different subsets separately and average them, which is denoted as "mean". Since the "all" setting merges data from different topics together, it makes the evaluation closer to real-world scenarios. Therefore, Gao et al. (2021b) take the "all" setting as the default, which we also follow.

6 https://github.com/huggingface/transformers

outputs before the MLP pooler layer as sentence embeddings for evaluation.

IMAGE ENCODER We use ResNet (He et al., 2016) as the image encoder (`ResNet-50`), and extract 2048-dimensional feature vectors from its last layer. The image encoder is kept frozen, as our preliminary experiments indicated that fine-tuning it does not have a significant impact on the STS performance.

PROJECTION HEADS Distinct projection heads are used for different modalities and contrastive learning objectives, each implemented as a single-layer MLP with Tanh activation. For the textual contrastive learning objective, the extracted `[CLS]` token representations (i.e., sentence embeddings) are further projected into a 768-dimensional space. For the multimodal objective, both the sentence embeddings and image feature vectors are projected into a shared 256-dimensional space and then normalized before computing the multimodal contrastive loss.

HYPERPARAMETERS Most hyperparameter settings are adopted from Gao et al. (2021b). The temperature parameters τ and τ' are both set to 0.05. Other key hyperparameters including learning rate, batch size, training epoch, and the trade-off parameter λ are reported in Table 4.1. To determine the best λ, we perform grid search over $\{0.005, 0.01, 0.05, 0.1, 0.5\}$ using the STS-B development set, and results in Table 4.2 show that models often achieve their best performance at moderate values of λ. During training, we evaluate checkpoints every 125 steps on the STS-B development set, and select the best checkpoint for final evaluation.

Model	**Parameter**	**Training Setting**				
		wiki	*wiki+flickr*	*wiki+coco*	*flickr*	*coco*
BERT	learning rate	3e-5				
	batch size	64				
	λ	–	0.01	0.01	0.05	0.05
	epochs	3	3	3	6	3
RoBERTa	learning rate	1e-5				
	batch size	128				
	λ	–	0.01	0.01	0.01	0.01
	epochs	3	3	3	6	3

Table 4.1: Parameter setups for different pre-trained models and training settings.

λ	0.001	0.01	0.05	0.1	0.5
MCSE-BERT	78.38	79.95	**80.41**	80.35	80.01
MCSE-RoBERTa	80.60	**81.48**	81.08	80.73	79.85

Table 4.2: STS-B performance of MCSE models trained on Flickr30k with different trade-off parameters. Both MCSE-BERT and MCSE-RoBERTa achieve their best results at moderate values of λ.

4.5 RESULTS AND ANALYSIS

4.5.1 *Main Results*

To fully utilize different types of data resources, we first conduct experiments with a text-only corpus (i.e., Wiki1M) and multimodal data (i.e., Flickr30k and MS-COCO). SimCSE is trained on sentences and captions only, while MCSE additionally computes the multimodal objective for image-caption pairs. As shown in Table 4.3, averaging the standard BERT and RoBERTa embeddings[7] yields poor performance on STS tasks. SimCSE models significantly outperform the average embeddings. MCSE models, which have access to auxiliary visual information, further achieve noticeable improvements even if the amount of multimodal data is relatively small. For example, when MCSE is applied to the combination of Wiki1M and Flickr30k (i.e., *wiki+flickr*), it improves the state-of-the-art result for BERT (76.3 $\rightarrow$ 77.3) and RoBERTa (76.6 $\rightarrow$ 78.3) by a decent margin. Our results demonstrate that **augmenting text-only corpus with small scale multimodal data can lead to significant improvements.**

Meanwhile, we also train models solely on multimodal data and report results in Table 4.4. We observe that, without the large text-only corpus, the performances decrease considerably compared to results in Table 4.3. Still, MCSE models consistently surpass SimCSE models (0.9 – 3.8 points improvement). Moreover, replacing the paired images with shuffled images before training MCSE leads to 0.8 – 5.0 points reduction in terms of average Spearman's correlation. This degradation validates that the observed improvements stem from the model's alignment with meaningful visual semantics.

To delve into the performance gap between MCSE and SimCSE, we calculate the Spearman's correlation for different subsets of each year's STS challenge separately. The Spearman's correlation improvements of MCSE-BERT over SimCSE-BERT are shown in Figure 4.2 and 4.3. In STS12, "MSRvid" subset achieves the largest improvement, which is a corpus of video descriptions. "Image" subsets in STS14 and STS15 also get considerable improvements. On the other hand, the performance of

7 Following (Gao et al., 2021b), we take the average embeddings from the first and last layers, which is better than using only the last layer.

	Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.↑
	$\text{BERT}_{\text{first-last avg.}}$	39.7	59.4	49.7	66.0	66.2	53.9	62.1	56.7
	$\text{RoBERTa}_{\text{first-last avg.}}$	40.9	58.7	49.1	65.6	61.5	58.6	61.6	56.6
wiki	SimCSE-BERT$^{\diamondsuit}$	68.4	82.4	74.4	80.9	78.6	76.9	72.2	76.3
wiki	SimCSE-RoBERTa$^{\diamondsuit}$	70.2	81.8	73.2	81.4	80.7	80.2	68.6	76.6
wiki	SimCSE-BERT	$67.8_{\pm 1.6}$	$80.0_{\pm 2.1}$	$72.5_{\pm 1.7}$	$80.1_{\pm 0.8}$	$77.6_{\pm 0.8}$	$76.5_{\pm 0.8}$	$70.1_{\pm 0.9}$	$74.9_{\pm 1.1}$
wiki	SimCSE-RoBERTa	$68.7_{\pm 1.0}$	$82.0_{\pm 0.5}$	$74.0_{\pm 1.0}$	$82.1_{\pm 0.4}$	$81.1_{\pm 0.4}$	$80.6_{\pm 0.3}$	$69.2_{\pm 0.2}$	$76.8_{\pm 0.5}$
wiki+flickr	SimCSE-BERT	$69.9_{\pm 1.7}$	$79.8_{\pm 1.5}$	$72.9_{\pm 0.9}$	$81.9_{\pm 0.8}$	$\mathbf{77.8}_{\pm 0.9}$	$76.6_{\pm 1.1}$	$68.4_{\pm 0.8}$	$75.3_{\pm 0.9}$
wiki+flickr	MCSE-BERT	$\mathbf{71.4}_{\pm 0.9}$	$\mathbf{81.8}^{*}_{\pm 1.3}$	$\mathbf{74.8}^{*}_{\pm 0.9}$	$\mathbf{83.6}_{\pm 0.9}$	$77.5_{\pm 0.8}$	$\mathbf{79.5}^{*}_{\pm 0.5}$	$\mathbf{72.6}^{*}_{\pm 1.4}$	$\mathbf{77.3}^{*}_{\pm 0.5}$
wiki+flickr	SimCSE-RoBERTa	$69.5_{\pm 0.9}$	$81.6_{\pm 0.5}$	$74.1_{\pm 0.6}$	$82.4_{\pm 0.3}$	$80.9_{\pm 0.5}$	$79.9_{\pm 0.3}$	$67.3_{\pm 0.5}$	$76.5_{\pm 0.4}$
wiki+flickr	MCSE-RoBERTa	$\mathbf{71.7}^{*}_{\pm 0.2}$	$\mathbf{82.7}^{*}_{\pm 0.4}$	$\mathbf{75.9}^{*}_{\pm 0.3}$	$\mathbf{84.0}^{*}_{\pm 0.4}$	$\mathbf{81.3}_{\pm 0.3}$	$\mathbf{82.3}^{*}_{\pm 0.5}$	$\mathbf{70.3}^{*}_{\pm 1.3}$	$\mathbf{78.3}^{*}_{\pm 0.1}$
wiki+coco	SimCSE-BERT	$69.1_{\pm 1.0}$	$80.4_{\pm 0.9}$	$72.7_{\pm 0.7}$	$81.1_{\pm 0.3}$	$\mathbf{78.2}_{\pm 0.9}$	$73.9_{\pm 0.6}$	$66.6_{\pm 1.2}$	$74.6_{\pm 0.2}$
wiki+coco	MCSE-BERT	$\mathbf{71.2}^{*}_{\pm 1.3}$	$\mathbf{79.7}_{\pm 0.9}$	$\mathbf{73.8}_{\pm 0.9}$	$\mathbf{83.0}^{*}_{\pm 0.4}$	$77.8_{\pm 0.9}$	$\mathbf{78.5}^{*}_{\pm 0.4}$	$\mathbf{72.1}^{*}_{\pm 1.4}$	$\mathbf{76.6}^{*}_{\pm 0.5}$
wiki+coco	SimCSE-RoBERTa	$66.4_{\pm 0.9}$	$80.7_{\pm 0.7}$	$72.7_{\pm 1.1}$	$81.3_{\pm 0.9}$	$80.2_{\pm 0.8}$	$76.8_{\pm 0.6}$	$65.7_{\pm 0.7}$	$74.8_{\pm 0.5}$
wiki+coco	MCSE-RoBERTa	$\mathbf{70.2}^{*}_{\pm 1.7}$	$\mathbf{82.0}^{*}_{\pm 0.7}$	$\mathbf{75.5}^{*}_{\pm 1.2}$	$\mathbf{83.0}^{*}_{\pm 0.6}$	$\mathbf{81.5}^{*}_{\pm 0.7}$	$\mathbf{80.8}^{*}_{\pm 1.0}$	$\mathbf{69.9}^{*}_{\pm 0.6}$	$\mathbf{77.6}^{*}_{\pm 0.8}$

Table 4.3: Performance comparison on STS tasks. STS-B: STS Benchmark, SICK-R: SICK-Relatedness, Avg.: average across 7 tasks. $\diamondsuit$: single seed results from Gao et al., 2021b. All other results are from our implementation. Models are trained with 5 random seeds and we report the means and standard deviations. $*$: difference between SimCSE and MCSE is significant at $\alpha = 0.05$ according to an independent t-test.

"answers-students" subset in STS15 drops extensively, and none of the subsets in STS16 get noticeable improvement by MCSE. These results indicate that the gains from MCSE are domain-sensitive: subsets that are semantically or stylistically closer to the multimodal training data, such as those involving descriptions of scenes, actions, or images, show the clearest improvements. In contrast, subsets drawn from more abstract or conversational may not align well with the visual grounding signal, and in some cases, performance can even degrade.

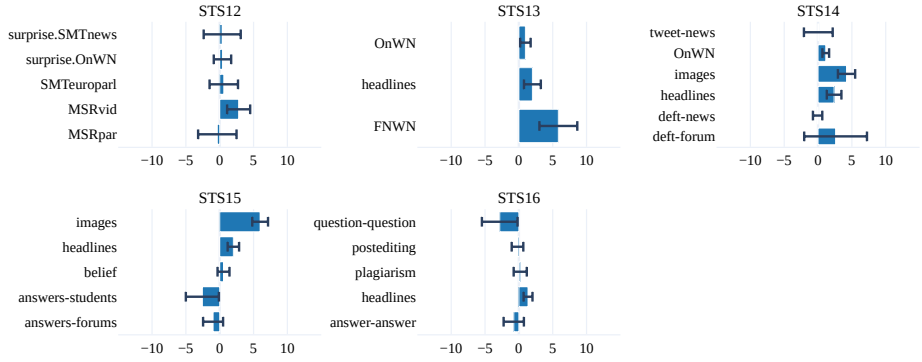


Figure 4.2: Performance improvements over different subsets in the training setting of *wiki+flickr*.

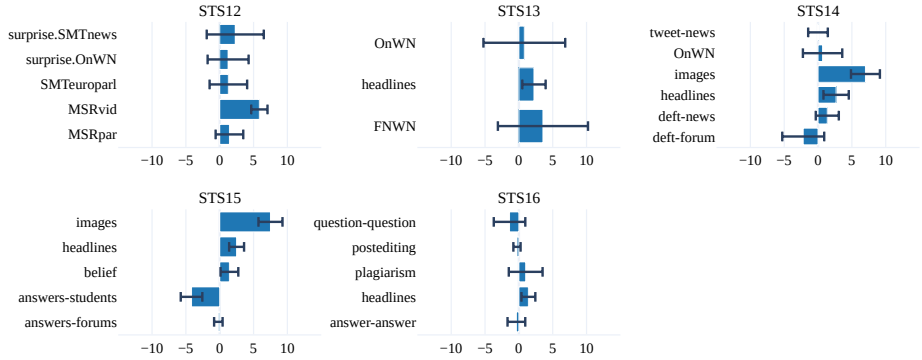


Figure 4.3: Performance improvements over different subsets in the training setting of *wiki+coco*.

4.5.2 *Ablation Studies*

IMPACT OF DATA SCALE We take BERT-based models trained merely on caption datasets and investigate the impact of training data scales. We limit the number of training samples to 100, 500, 1,000, 5,000 and 10,000, and compare their performance with the full set performance. In all of these settings, we optimize the models for same number of training steps as the full set setting. The results are shown in Figure 4.4. We find that SimCSE achieves better performance than MCSE under low-data conditions, while MCSE starts to outperform

Model	Training Setting	
	flickr	*coco*
SimCSE-BERT	$68.8_{\pm 0.7}$	$67.8_{\pm 0.4}$
MCSE-BERT	$\mathbf{70.6}^{*}_{\pm 0.5}$	$\mathbf{71.6}^{*}_{\pm 0.2}$
w/ shuffling	$67.9_{\pm 0.6}\downarrow$	$66.6_{\pm 0.3}\downarrow$
SimCSE-RoBERTa	$72.9_{\pm 0.3}$	$72.8_{\pm 0.3}$
MCSE-RoBERTa	$\mathbf{73.8}^{*}_{\pm 0.2}$	$\mathbf{74.3}^{*}_{\pm 0.3}$
w/ shuffling	$73.0_{\pm 0.4}\downarrow$	$72.8_{\pm 0.3}\downarrow$

Table 4.4: Comparison of the average Spearman's correlation on 7 STS tasks. We report the means and standard deviations over 5 seeds. $*$: difference between SimCSE and MCSE is significant.

SimCSE as the amount of training data increases. We attribute this trend to the progressive adaptation of the multimodal projection heads, which require sufficient training signals to effectively align textual and visual representations.

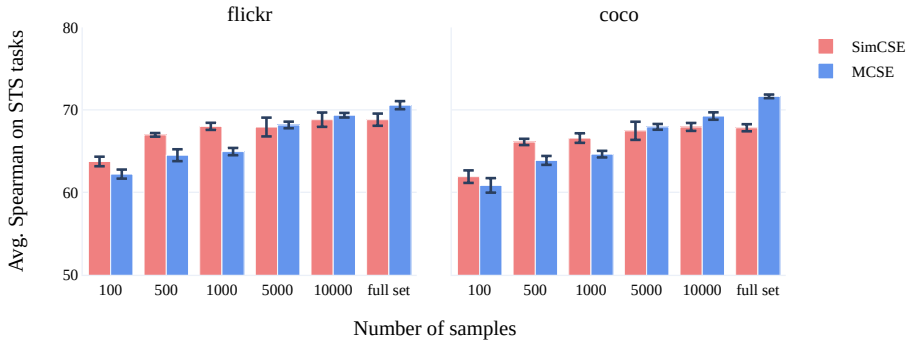


Figure 4.4: Performances of different data scales. The full Flickr30k and MS-COCO datasets contain about 30K and 87K samples, respectively.

IMPACT OF IMAGE ENCODER We further replace the ResNet encoder with CLIP (Radford et al., 2021) to examine the impact of different image encoders. Our implementation is based on the Sentence Transformer library[8] (Reimers and Gurevych, 2019) and we use the checkpoint `clip-ViT-B-32` to extract 512-dimensional feature vectors. As shown in Table 4.5, different image encoders yield very similar performances across tasks.

COMBINING WIKI1M, FLICKR30K, AND MS-COCO We apply the same parameter setting as *wiki+flickr* and *wiki+coco*, and train models on the combination of Wiki1M, Flickr30k, and MS-COCO datasets. As shown in Table 4.6, MCSE models outperform SimCSE baselines by 1.9

8 https://github.com/UKPLab/sentence-transformers

	Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.↑
flickr	SimCSE-BERT	$62.1_{\pm 0.5}$	$73.8_{\pm 0.9}$	$64.2_{\pm 0.6}$	$74.2_{\pm 0.8}$	$\mathbf{74.8}^{*}_{\pm 0.6}$	$67.1_{\pm 1.1}$	$65.4_{\pm 1.1}$	$68.8_{\pm 0.7}$
	MCSE-ResNet-BERT	$\mathbf{63.6}^{*}_{\pm 0.7}$	$\mathbf{74.0}_{\pm 0.9}$	$65.5_{\pm 1.1}$	$75.5_{\pm 0.2}$	$71.6_{\pm 0.4}$	$74.0_{\pm 0.4}$	$69.8_{\pm 0.3}$	$70.6_{\pm 0.5}$
	MCSE-CLIP-BERT	$63.1_{\pm 0.7}$	$73.9_{\pm 1.0}$	$\mathbf{65.8}^{*}_{\pm 0.9}$	$\mathbf{76.0}^{*}_{\pm 0.7}$	$70.7_{\pm 0.3}$	$\mathbf{74.9}^{*}_{\pm 0.5}$	$\mathbf{70.7}^{*}_{\pm 0.3}$	$\mathbf{70.7}^{*}_{\pm 0.2}$
	SimCSE-RoBERTa	$66.6_{\pm 0.5}$	$78.3_{\pm 0.5}$	$69.7_{\pm 0.6}$	$77.7_{\pm 0.5}$	$\mathbf{76.3}^{*}_{\pm 0.5}$	$75.8_{\pm 0.3}$	$66.2_{\pm 0.4}$	$72.9_{\pm 0.3}$
	MCSE-ResNet-RoBERTa	$\mathbf{67.6}^{*}_{\pm 0.5}$	$\mathbf{78.8}_{\pm 0.4}$	$\mathbf{70.1}_{\pm 0.3}$	$78.5_{\pm 0.2}$	$75.4_{\pm 0.5}$	$\mathbf{77.4}^{*}_{\pm 0.3}$	$68.6_{\pm 0.3}$	$\mathbf{73.8}^{*}_{\pm 0.2}$
	MCSE-CLIP-RoBERTa	$67.0_{\pm 0.5}$	$78.6_{\pm 0.4}$	$69.8_{\pm 0.5}$	$\mathbf{78.7}^{*}_{\pm 0.8}$	$74.9_{\pm 0.5}$	$\mathbf{77.4}^{*}_{\pm 0.4}$	$\mathbf{69.5}^{*}_{\pm 0.5}$	$73.7_{\pm 0.2}$
coco	SimCSE-BERT	$59.3_{\pm 0.9}$	$73.0_{\pm 1.2}$	$62.7_{\pm 0.6}$	$74.7_{\pm 0.7}$	$\mathbf{74.4}^{*}_{\pm 0.4}$	$65.3_{\pm 0.7}$	$65.4_{\pm 0.5}$	$67.8_{\pm 0.4}$
	MCSE-ResNet-BERT	$\mathbf{64.9}^{*}_{\pm 0.5}$	$\mathbf{74.8}^{*}_{\pm 0.9}$	$\mathbf{68.1}^{*}_{\pm 0.6}$	$\mathbf{76.8}^{*}_{\pm 0.6}$	$72.7_{\pm 0.8}$	$\mathbf{74.5}^{*}_{\pm 0.4}$	$69.7_{\pm 0.4}$	$\mathbf{71.6}^{*}_{\pm 0.2}$
	MCSE-CLIP-BERT	$64.8_{\pm 0.6}$	$74.1_{\pm 0.6}$	$68.0_{\pm 0.2}$	$76.2_{\pm 0.5}$	$71.6_{\pm 0.4}$	$\mathbf{74.5}^{*}_{\pm 0.3}$	$\mathbf{70.3}^{*}_{\pm 0.6}$	$71.4_{\pm 0.1}$
	SimCSE-RoBERTa	$64.7_{\pm 0.6}$	$79.2_{\pm 0.4}$	$70.2_{\pm 0.4}$	$79.0_{\pm 0.6}$	$\mathbf{78.2}_{\pm 0.5}$	$73.8_{\pm 0.5}$	$64.6_{\pm 0.3}$	$72.8_{\pm 0.3}$
	MCSE-ResNet-RoBERTa	$\mathbf{67.0}^{*}_{\pm 0.8}$	$\mathbf{79.4}_{\pm 0.4}$	$\mathbf{70.9}^{*}_{\pm 0.4}$	$\mathbf{80.0}^{*}_{\pm 0.4}$	$77.8_{\pm 0.5}$	$\mathbf{76.9}^{*}_{\pm 0.4}$	$67.9_{\pm 0.7}$	$\mathbf{74.3}^{*}_{\pm 0.3}$
	MCSE-CLIP-RoBERTa	$66.0_{\pm 1.0}$	$79.0_{\pm 0.7}$	$70.6_{\pm 0.6}$	$\mathbf{80.0}^{*}_{\pm 0.8}$	$77.6_{\pm 0.5}$	$76.5_{\pm 0.4}$	$\mathbf{68.4}^{*}_{\pm 0.8}$	$74.0_{\pm 0.2}$

Table 4.5: Performance comparison on STS tasks. STS-B: STS Benchmark, SICK-R: SICK-Relatedness, Avg.: average across 7 tasks. Models are trained with 5 random seeds and we report means and standard deviations. $*$: difference between SimCSE and MCSE (using ResNet or CLIP) is significant at $\alpha = 0.05$ according to an independent t-test.

and 2.6 points when using BERT and RoBERTa, respectively. However, this combined training yields lower performance than training on Flickr30k alone. We attribute this to the MS-COCO caption style, which prioritizes object-level details over the global scene descriptions as in Flickr30k. When learning sentence embeddings, this shift in textual style may introduce noise rather than a useful signal, thereby offsetting the benefits of a larger training set.

Model	**Training Setting** *wiki+flickr+coco*
SimCSE-BERT	74.3±1.0
MCSE-BERT	**76.2**±0.3
SimCSE-RoBERTa	75.3±0.7
MCSE-RoBERTa	**77.9**±0.6

Table 4.6: Comparison of the average Spearman's correlation of 7 STS tasks. We report the means and standard deviations over 5 random seeds.

Model	**image → text**		**text → image**	
	R@1	R@5	R@1	R@5
MCSE-BERT$_{wiki+flickr}$	16.7	43.5	22.5	50.4
MCSE-BERT$_{flickr}$	20.4	50.2	23.8	52.5
MCSE-BERT$_{wiki+coco}$	8.8	26.6	10.9	31.2
MCSE-BERT$_{coco}$	8.2	25.2	9.0	27.1

Table 4.7: Cros-modal retrieval results on Flickr30k test set and MS-COCO minival set.

4.5.3 *Retrieval Performance*

CROSS-MODAL RETRIEVAL We evaluate BERT-based MCSE models (using the same random seed) on cross-modal retrieval tasks. Retrieval performance is measured using Recall@K, which checks whether the ground-truth caption or image for a query appears among the top-K retrieved items. As shown in Table 4.7, MCSE models achieve a decent level of retrieval performance as a by-product of multimodal contrastive learning.

SENTENCE RETRIEVAL We take BERT-based models trained on the Flickr30k train set and conduct a sentence retrieval experiment on Flickr30k test set. For each input sentence, the nearest neighbor is retrieved based on cosine similarity. Selected retrieval examples are presented in Table 4.8. We observe two key patterns: (1) SimCSE tends

to retrieve sentences with similar syntax, whereas MCSE retrieves sentences that vary in syntax while preserving semantic content (e.g., Q1, Q3, Q6). (2) MCSE is more effective at recognizing similar event scenes and capturing the number of entities in a sentence (e.g., Q2, Q4, Q5).

Model	**Sentence**
Query 1: A young girl is washing her teddy bear in the kitchen sink.	
SimCSE:	A middle-aged woman is vacuuming her kitchen floor with a canister vac.
MCSE:	A young girl, blond and wearing a polka-dot shirt, washes a stuffed animal in a vanity sink.
Query 2: Three chefs , wearing white hats and black aprons , are preparing food in a crowded kitchen.	
SimCSE:	Numerous workers with blue shirts and white aprons are preparing fish for sale.
MCSE:	Three men are preparing food in a kitchen setting.
Query 3: A couple kisses in a shady walkway.	
SimCSE:	A couple strolls down a path near benches and water.
MCSE:	Couple kissing outside on street.
Query 4: A man is standing on the streets taking photographs.	
SimCSE:	People run a marathon on a city street with a crowd watching.
MCSE:	A guy wearing a white shirt is taking a picture.
Query 5: Two boys are playing in pool filled with sparkling blue water.	
SimCSE:	A little girl is swimming under the crystal blue water.
MCSE:	Two children are swimming in a pool.
Query 6: An old man sitting on a bench staring at the ocean.	
SimCSE:	A man sitting on a bench by the ocean.
MCSE:	An old man sits on a bench overlooking the water.

Table 4.8: Sentence retrieval examples from Flickr30k test set.

4.5.4 *Representation Analysis*

To dissect the inner workings of MCSE, we use two quantifiable metrics proposed in Wang and Isola (2020): *alignment* and *uniformity*, as measurements of representation quality. Let p_{pos} denote the positive pairs distribution and p_{data} denote the data distribution. The *alignment*

loss prefers encoders that assign similar features to semantically similar instances (assuming features have been normalized):

$$\mathcal{L}_{align} \triangleq \underset{(x,x^+) \sim p_{\text{pos}}}{\mathbb{E}} \left\| f(x) - f(x^+) \right\|_2^2 . \tag{4.5}$$

And the *uniformity loss* prefers a uniform distribution in the hypersphere:

$$\mathcal{L}_{uniform} \triangleq \log \underset{x,y \overset{i.i.d.}{\sim} p_{\text{data}}}{\mathbb{E}} e^{-2\|f(x)-f(y)\|_2^2} . \tag{4.6}$$

Gao et al. (2021b) empirically showed that sentence embedding models with both lower alignment and uniformity tend to achieve better overall performance. Similarly, we compute the two losses on STS-B[9] and results are presented in Figure 4.5. It shows that MCSE models achieve better alignment scores compared to SimCSE while also maintaining uniformity. This analysis provides evidence that visual grounding serves as a powerful regularizer, enhancing sentence representation learning by refining the geometric structure of the textual embedding space.

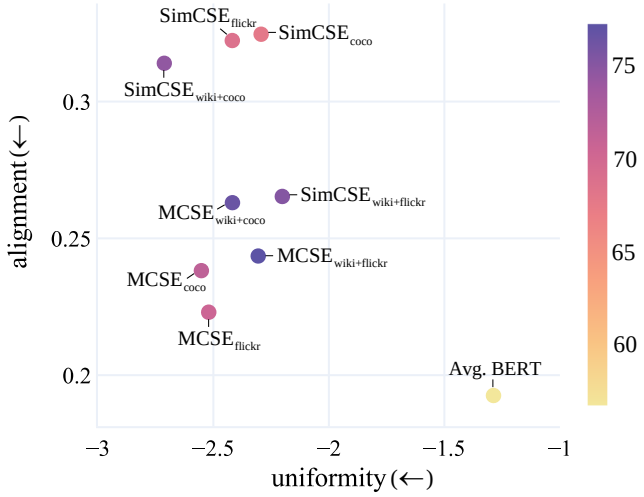


Figure 4.5: The alignment-uniformity plot of SimCSE and MCSE models using BERT. Dot color represents the average Spearman's correlation.

4.6 DISCUSSION

LIMITATIONS Despite showing performance improvements on STS benchmarks, MCSE has its limitations as well. First, we take caption

9 We take STS-B pairs with a score higher than 4.0 as p_{pos} and the full STS-B as p_{data}. Since Gao et al. (2021b) did not release the code for calculating these two losses, the absolute values we obtained might be different from theirs. We make sure our calculation across different models is consistent.

datasets as the source of multimodal information, while these datasets are collected and curated with non-negligible human efforts. In practice, it would be valuable to explore whether noisy or weakly aligned image-sentence pairs can be leveraged effectively, or even whether explicit alignments between images and sentences can be eliminated entirely. Second, we observe that performance improvements are often limited to subsets of data from related domains, whereas other domains suffer from distribution shifts. This highlights the importance of mitigating domain gaps in order to develop sentence embeddings that generalize broadly across contexts. Third, the definition of *semantic similarity* is inherently task-dependent. While STS benchmarks have been a dominant evaluation framework, they only capture a narrow view of similarity. A more comprehensive evaluation should be considered, encompassing diverse applications such as retrieval, paraphrase detection, and clustering.

FUTURE WORKS Rethinking the problem of sentence embedding learning opens up follow-up questions:

- **Explainability**: what factors truly drive the similarity between two sentences? Beyond surface form, can abstract semantic representations provide interpretable explanations for model behavior? A work in this line of research (Opitz and Frank, 2022) propose S^3BERT embeddings, which decompose sentence representations into explainable sub-embeddings that emphasize various semantic features (e.g., semantic roles, negation, or quantification), thereby making the contribution of each aspect to sentence similarity more transparent.
- **Decoder-only models**: while decoder-only models have outperformed encoder-only models on a large variety of tasks, their role in sentence representation learning remains underexplored, mainly due to the architectural limitation of the causal attention mechanism. Early work, such as LLM2Vec (BehnamGhader et al., 2024), provides pilot efforts to transform a decoder-only model into a strong text encoder, but leaves open questions like how to optimize the training and inference latency with the ever-increasing model scale.

4.7 CONCLUSION

In this chapter, we introduce MCSE, a novel approach for sentence embedding learning that leverages a multimodal contrastive objective to align sentences and corresponding images in a grounded space. Through extensive experiments, we demonstrate that MCSE consistently improves the performance on STS tasks, outperforming strong text-only baselines. To better understand the inner workings of MCSE,

we analyze the alignment and uniformity properties of the embedding space, showing that visual grounding improves the alignment score while maintaining uniformity. Importantly, the multimodal objective is generic and can be potentially incorporated into other sentence embedding methods to boost their performance.

5

MULTILINGUAL SEMANTIC TEXTUAL RELATEDNESS

Building on the exploration of word-level and sentence-level semantic embeddings in the previous chapters, this chapter turns to the challenge of modeling semantic textual relatedness, a task that becomes particularly difficult for under-represented languages with limited data, such as African languages. To address this problem, we propose a framework based on the cross-encoder architecture and apply a set of techniques in low-resource settings, including data augmentation via machine translation, task-adaptive pre-training to better align pre-trained models with the downstream objective, and adapter-based parameter-efficient tuning. We further investigate source language selection strategies for zero-shot cross-lingual transfer. Our approach achieves top performance on the SemEval-2024 leaderboard (Ousidhoum et al., 2024a) in both supervised and cross-lingual transfer setups. In addition, we also perform a fine-trained error analysis that reveals the specific weakness of our method and highlights promising directions for future research.[1]

5.1 INTRODUCTION

Semantic textual relatedness (STR) measures the degree to which two linguistic units, such as a pair of words or sentences, are meaningfully connected (Budanitsky, 1999; Mohammad and Hirst, 2012). For example, one can easily tell that *"I like playing games"* is more semantically related to *"The game is fun"* than *"The weather is good"*, which largely depends on their lexical semantic relation and topic consistency. Semantic Textual Similarity (STS), a closely related concept, indicates whether two units have a paraphrasing relation. The difference between these two concepts is clarified in Abdalla et al. (2023) that while similar pairs are also related, the reverse is not necessarily true. For example, *"The economy is slowing down"* and *"Unemployment rates are rising"* are clearly related, yet not strictly similar, as they provide different information about the same domain.

In stark contrast to the extensive research on STS with embedding-based methods (Gao et al., 2021b; Chuang et al., 2022; Zhang et al., 2022; Seonwoo et al., 2023), exploration of STR lags behind and focuses

1 This chapter is based on Zhang et al. (2024b). As the first author, Miaoran Zhang led the project, conducted model training and evaluation experiments, and was the main writer of this paper. The source code for this work is available on Github: https://github.com/uds-lsv/AAdaM.

primarily on English (Marelli et al., 2014; Abdalla et al., 2023). The main bottleneck lies in the lack of high-quality datasets for a broader set of languages. To close this gap, SemEval-2024 Task 1: Semantic Textual Relatedness (Ousidhoum et al., 2024a) is proposed to encourage STR research on 14 representative African and Asian languages, including three setups with different data availability: supervised learning, unsupervised learning, and zero-shot cross-lingual transfer.

In this chapter, we present AAdaM (**A**ugmentation and **Ada**ptation for **M**ultilingual STR), a framework developed to advance multilingual semantic relatedness modeling, focusing on supervised learning and zero-shot cross-lingual transfer setups. Our framework adopts a cross-encoder architecture that takes the concatenation of a pair of sentences as input and predicts the relatedness score through a regression head (Devlin et al., 2019). We perform data augmentation via machine translation to address the limited data for non-English languages. To better adapt a pre-trained model to the STR task, we apply task-adaptive pre-training (Gururangan et al., 2020) which has shown effectiveness on many tasks (Xue et al., 2021; Wang et al., 2023c). For supervised training, we explore full parameter fine-tuning and adapter-based tuning (Houlsby et al., 2019), and for cross-lingual transfer, we use the MAD-X adapter framework (Pfeiffer et al., 2020), enabling efficient modular transfer across languages.

We evaluate AAdaM on the SemEval-2024 STR benchmarks, and results show that it demonstrates superior performance: AAdaM improves over the baseline model by 3.8% and 7.6% points in supervised and cross-lingual transfer settings, respectively. In particular, AAdaM achieves first-place ranking on average among the participating systems in both scenarios. These results highlight the effectiveness of our proposed framework, which combines data augmentation, task-adaptive pre-training, and adapter-based tuning for robust multilingual semantic relatedness modeling.

5.2 RELATED WORK

Multilingual pre-trained language models such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) have revolutionized multilingual learning in NLP. These models excel at learning cross-lingual representations, due to their ability to capture language-neutral features that go beyond linguistic differences (Pires et al., 2019; Libovický et al., 2020; Xie et al., 2022b; Chang et al., 2022). However, their performance is uneven across languages due to varying data availability, and thus efforts have been made to improve these models for better cross-lingual transfer (Pfeiffer et al., 2020; Parović et al., 2022; Alabi et al., 2022). Some multilingual models focus on modeling semantic relationships between texts, such as Reimers and Gurevych (2020) and Wang et al. (2024b). These embedding models often adopt

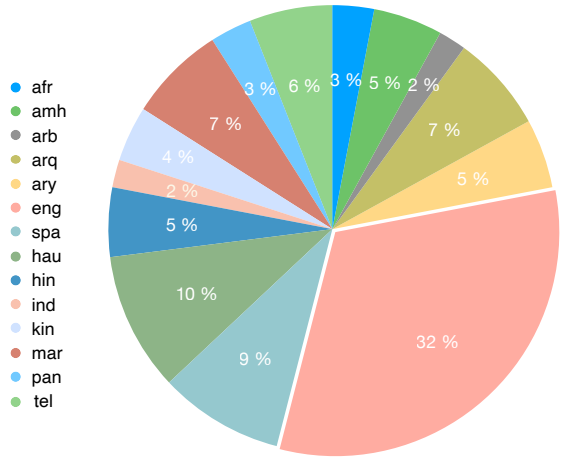


Figure 5.1: SemRel data distribution across languages. Languages: `afr` = Afrikaans, `amh` = Amharic, `arb` = Modern Standard Arabic, `arq` = Algerian Arabic, `ary` = Moroccan Arabic, `eng` = English, `spa` = Spanish, `hau` = Hausa, `hin` = Hindi, `ind` = Indonesian, `kin` = Kinyarwanda, `mar` = Marathi, `pan` = Punjabi, `tel` = Telugu.

a bi-encoder architecture, which encodes texts independently and enables efficient large-scale semantic search. A complementary approach is to use cross-encoders (Gao et al., 2021a), which differ from bi-encoders by jointly processing text pairs rather than encoding them separately. Because cross-encoders allow full interaction (e.g., via cross-attention) between the two texts during encoding, they tend to produce more accurate results at the cost of inference speed (Humeau et al., 2020). Recent work (Liang et al., 2024) illustrates a hybrid pipeline: first training an embedding model and then refining it via knowledge distillation from a cross-encoder teacher, achieving better embedding quality without prohibitive inference costs.

5.3 SEMREL DATASET

In this section, we describe the dataset used in this chapter. SemRel, introduced by Ousidhoum et al. (2024b), is a newly proposed STR dataset annotated by native speakers. It is composed of sentence pairs, each assigned a relatedness score between 0 (completely unrelated) and 1 (maximally related). It covers 14 languages from five distinct language families, most of which are spoken in Africa and Asia and remain under-represented in NLP resources. Table 5.1 presents the number of samples for each language across the training, development, and test sets. As illustrated in Figure 5.1, the overall data sizes vary widely from language to language constrained by the availability of resources. Notably, English data comprises 32% of the whole dataset and surpasses other languages by a large margin.

Language	Train	Dev	Test	Total
afr	-	375	375	700
amh	992	95	171	1,258
arb	-	32	595	627
arq	1,261	97	583	1,941
ary	924	71	426	1,421
eng	5,500	250	2,600	8,350
spa	1,562	140	600	2,302
hau	1,736	212	603	2,551
hin	-	288	968	1,256
ind	-	144	360	504
kin	778	102	222	1,102
mar	1,200	293	298	1,791
pan	-	242	634	876
tel	1,170	130	297	1,597

Table 5.1: Number of samples in the training, dev, and test sets for different languages. Languages with no training data (`afr`, `arb`, `hin`, `ind`) are only used in the cross-lingual transfer evaluation.

5.4 METHOD

This section presents the core components of our proposed framework. We begin with **model selection** (Section 5.4.1) from a range of multilingual pre-trained models and examine two architectures: bi-encoders and cross-encoders, which forms the foundation for the subsequent discussions. Next, we introduce our **data augmentation** (Section 5.4.2) and **task-adaptive pre-training** (Section 5.4.3) techniques to address the low-resource challenge. We then describe the **supervised training** paradigms (Section 5.4.4), including full parameter fine-tuning and adapter-based tuning. Finally, we explore **cross-lingual transfer** strategies (Section 5.4.5), focusing on the choice of source languages and using modularized adapters to effectively transfer knowledge from the source to target languages.

5.4.1 *Model Selection*

Our first step is to determine a suitable backbone model and architecture for semantic relatedness modeling. There are two commonly used architectures: bi-encoders (i.e., embedding models) and cross-encoders. Bi-encoders, such as Sentence-BERT (Reimers and Gurevych, 2019), encode sentences independently and compute their relatedness based on cosine distance. In contrast, cross-encoders take a concatenated sentence pair as input and predict a relatedness score through a regression head. Previous work has shown that cross-encoders generally

achieve better performance by allowing full interaction between sentences, although at the cost of inference latency (Humeau et al., 2020). In our study, we evaluate the capability of a range of multilingual pre-trained models in both architectures, including two categories of models:

- **sentence transformers**: `mpnet-base-v2`[2] and `LaBSE` (Feng et al., 2022)
- **general-purpose models**: `XLMR-large` (Conneau et al., 2020), `AfroXLMR-large` (Alabi et al., 2022), `AfriBERTa-large` (Ogueji et al., 2021), `AfroXLMR-large-61L` and `AfroXLMR-large-75L` (Adelani et al., 2024)

We first assess their out-of-the-box capabilities by extracting contextual embeddings for sentence pairs and using cosine similarity to predict semantic relatedness scores. We then evaluate their full potential by fine-tuning them as bi-encoders and cross-encoders on the task data. We also include two simple baselines: word overlap[3] and fastText (Mikolov et al., 2018). For both fastText embeddings and contextual embeddings extracted from pre-trained models, we apply mean pooling to derive sentence-level embeddings from token-level embeddings. Performance is measured by the Spearman's rank correlation coefficient ($\times 100$) between the gold annotations and the predicted scores.

In Table 5.2, we can see that sentence transformers achieve superior performance in most languages without additional training. This observation is not surprising, as these models are explicitly optimized to produce meaningful sentence embeddings that capture semantic relationships effectively. However, this trend changes once the models are fine-tuned on the task data using either a bi-encoder or a cross-encoder architecture. Specifically, we pick up three models for 10-fold cross-validation on the SemRel training sets: `mpnet-base-v2` and `LaBSE`, the two best-performing sentence transformers, and `AfroXLMR-large-61L`, the strongest general-purpose model based on the out-of-the-box performance. Results show that with the cross-encoder setup, `AfroXLMR-large-61L` attains performance comparable to that of LaBSE. To follow the requirements of SemEval-2024 Task 1, that models pre-trained on relatedness or similarity data are prohibited, we adopt `AfroXLMR-large-61L` as our backbone model and exclude LaBSE from consideration. Moreover, since the cross-encoder architecture yields stronger performance on average, we use it for all subsequent experiments, leaving the exploration of more efficient alternatives for future work.

2 https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

3 https://github.com/semantic-textual-relatedness/Semantic_Relatedness_SemEval2024/blob/main/STR_Baseline.ipynb

Model	eng	amh	arq	ary	spa	hau	mar	tel	Avg.↑
Baselines w/o training:									
Overlap	56.57	63.28	44.00	**53.76**	58.67	38.86	57.52	60.61	54.16
FastText	55.69	60.64	**44.27**	22.12	57.47	9.19	59.23	69.39	47.25
mpnet-base-v2	**81.94**	69.94	26.35	34.40	56.58	30.86	72.43	56.33	53.60
LaBSE	72.14	**76.49**	40.80	38.58	**63.11**	**41.51**	**73.83**	**75.99**	**60.31**
XLMR-large	39.53	42.07	27.91	4.15	47.59	7.34	40.51	56.36	33.18
AfroXLMR-large	16.55	39.82	20.30	-0.46	30.42	8.13	35.94	30.74	22.68
AfriBERTa-large	53.12	69.23	16.04	13.36	56.68	35.14	20.84	9.73	34.27
AfroXLMR-large-61L	44.10	52.96	32.15	0.35	51.07	17.62	37.66	47.17	35.39
AfroXLMR-large-75L	22.61	37.93	29.38	-2.39	43.58	13.86	32.13	40.42	27.19
Bi-encoders w/ supervised training:									
mpnet-base-v2	85.07	80.43	56.73	75.51	65.29	58.62	81.53	74.49	72.21
LaBSE	84.45	82.59	59.49	78.29	69.02	68.94	**83.97**	76.35	75.39
AfroXLMR-large-61L	82.81	74.61	40.02	66.58	66.65	66.51	38.51	65.73	62.68
Cross-encoders w/ supervised training:									
mpnet-base-v2	80.26	75.04	60.25	80.31	64.92	53.66	65.36	68.54	68.54
LaBSE	86.13	84.75	**60.75**	**82.55**	67.23	69.31	81.10	77.25	**76.13**
AfroXLMR-large-61L	**86.65**	**84.88**	46.61	81.56	**69.08**	**74.65**	75.55	**80.94**	74.99

Table 5.2: Performance of 10-fold cross-validation on training sets (Spearman's correlation ×100). For each language, the best performance is boldfaced in *w/o training* and *w/ supervised training* settings.

5.4.2 *Data Augmentation*

Data augmentation serves as a widely used strategy to mitigate data scarcity in low-resource languages (Hedderich et al., 2021; Feng et al., 2021). Inspired by work on machine translation (Hu et al., 2020; Amjad et al., 2020), we create additional training data for non-English languages by translating from various English sources, as illustrated below.

SEMREL TRANSLATION As discussed in Section 5.3, English data occupies a significant portion of the entire SemRel dataset. To augment the data for other languages, we translate the English subset of SemRel into all target languages.

STS-B TRANSLATION STS-B (Cer et al., 2017), a semantic similarity dataset closely related to STR, is also translated from English into all target languages to provide additional training examples.

It is worth noting that using translations as data augmentation can result in varying data quality, since the translation process may introduce noises. Furthermore, while the concepts of "similarity" and "relatedness" are related, they are not identical, which can lead to discrepancies in their annotated scores. To effectively leverage data

of mixed quality, Zhu et al. (2023) shows that a two-phase approach is beneficial: first training the model on noisy data, and then fine-tuning it on clean data. Our training procedure follows this two-phase scheme: (1) a *warmup* phase training on the augmented data, and (2) a subsequent phase training on the original task data, ensuring the model benefits from both the quantity of augmented data and the quality of original annotations.

5.4.3 *Task-Adaptive Pre-Training*

PLMs are trained on massive text corpora with self-supervision objectives for general purposes (Devlin et al., 2019; Liu et al., 2019). To better adapt PLMs to downstream tasks, Gururangan et al. (2020) propose task-adaptive pre-training (TAPT), which involves continued pre-training on task-specific unlabeled data. This intermediate step helps bridge the distribution gap between the model's general pre-training domain and the target task, which has been shown to effectively improve downstream performance. We integrate this strategy into our system by performing masked language modeling (MLM) on unlabeled task data for a given target language before initiating any supervised training.

5.4.4 *Supervised Training Paradigms*

Fine-tuning is the conventional approach to adapt general-purpose PLMs to downstream tasks. Vanilla fine-tuning (i.e., full parameter fine-tuning) updates all model parameters for each task, leading to inefficiency with the ever-increasing model scales and number of tasks. Recently, many works propose lightweight alternatives to improve efficiency (Lester et al., 2021; Hu et al., 2022; He et al., 2022). For example, adapter-based tuning (Houlsby et al., 2019) only updates small modules known as adapters inserted between the layers of PLMs while keeping the remaining parameters frozen. In particular, it has shown impressive performance in cross-lingual transfer (Pfeiffer et al., 2020; Ansell et al., 2021; Pfeiffer et al., 2022).

In this study, we explore both full parameter fine-tuning and adapter-based tuning to compare their effectiveness on multilingual STR. For full parameter fine-tuning, we update all model parameters at each stage, namely the TAPT stage with unlabeled task data (Section 5.4.3), the warmup stage with augmented data (Section 5.4.2), and the final supervised training stage using the annotated task data. For adapter-based tuning, we adopt the MAD-X framework (Pfeiffer et al., 2020) which consists of language-specific adapters and task-specific adapters. The language adapters are pre-trained with an MLM objective on unlabeled monolingual corpora. We collect open-source data from the Leipzig Corpus Collection (Goldhahn et al., 2012) and use the recent

Language	Family / Subfamily	Domain	Corpus Size
English (eng)	Indo-Europoean / Germanic	News, Wikipedia	1.2M
Afrikaans (afr)	Indo-Europoean / Germanic	News, Wikipedia	68k
Amharic (amh)	Afro-Asiatic / Semitic	Community, Wikipedia	250k
Modern Standard Arabic (arb)	Afro-Asiatic / Semitic	News, Wikipedia	110k
Algerian Arabic (arq)	Afro-Asiatic / Semitic	News	244k
Moroccan Arabic (ary)	Afro-Asiatic / Semitic	News	564k
Spanish (spa)	Indo-Europoean / Italic	News, Wikipedia	444k
Hausa (hau)	Afro-Asiatic / Chadic	Community, Wikipedia	564k
Hindi (hin)	Indo-European / Indo-Iranian	News, Wikipedia	472k
Indonesian (ind)	Austronesian / Malayic	News, Wikipedia	92k
Kinyarwanda (kin)	Niger-Congo / Atlantic–Congo	Community	320k
Punjabi (pan)	Indo-European / Indo-Iranian	Wikipedia	412k
Marathi (mar)	Indo-European / Indo-Iranian	News, Wikipedia	856k
Telugu (tel)	Dravidian / South-Central	News, Wikipedia	756k

Table 5.3: Data statistics for pre-training corpora collected from the Leipzig Corpus Collection.

data derived from news and Wikipedia domains[4] as the pre-training corpora. Data statistics are shown in Table 5.3. The task adapters are trained on labeled task-specific data (augmented or original), while keeping the language adapters fixed. Note that when applying TAPT, only language adapters are updated.

5.4.5 *Cross-Lingual Transfer*

In cross-lingual transfer, the model is trained on a source language without access to labeled data in the target language. We adopt the MAD-X framework for this setup. During inference, we simply replace the source language adapter with the *target language adapter* while retaining the *source task adapter*. This task adapter has been trained on labeled data from the source language.[5] A crucial challenge for cross-lingual transfer lies in source language selection, as inappropriate sources may lead to negative results (Lange et al., 2021). To determine the best source language, we explore three metrics to estimate the transfer performance:

LINGUISTIC DISTANCE. We use the average of six distances obtained from the URIEL database (Littell et al., 2017) to measure the similarity between a pair of languages. These distances include syntactic, phonological, inventory, geographic, genetic, and featural distances. A lower distance indicates that the two languages are more similar, potentially facilitating more effective transfer.

TOKEN OVERLAP. We follow Wu and Dredze (2019) to measure how many tokens are shared in the source language training set and the target language test set. A higher token overlap indicates that more tokens were encountered during training in the source language, potentially transferring more supervision from the source to the target.

DEVELOPMENT SET PERFORMANCE. As small development sets are available in the SemEval task, we use their performance as an indicator of the transfer performance on test sets, assuming that they share a similar data distribution.

These metrics operate at different levels of dependency: linguistic distance is a heuristic metric; token overlap is a data-driven metric based on data characteristics; development set performance captures the combined effects of the data and model behavior. In our follow-up experiments, we find that development set performance proves to

4 As the SemRel data spans diverse domains, there is a potential risk of domain mismatch between the pre-training data and task data, which needs further investigation.

5 Note that when transferring from any other language to English, we ensure that the source task adapter has not been trained on augmented data translated from English resources, thereby eliminating the effect of data leakage.

be the most reliable predictor of cross-lingual transfer effectiveness. Notably, it might be more beneficial to directly use small development sets for training rather than source selection (Zhu et al., 2023), if training is allowed, which needs to be further explored.

5.5 EXPERIMENTAL SETUP

MODEL. Our backbone model is `AfroXLMR-large-61L`, adapted from XLM-R (Conneau et al., 2020) through multilingual adaptive fine-tuning (Alabi et al., 2022) for 61 languages. We use the sigmoid function as our regression head. For English data translation, we use NLLB (Team et al., 2022)[6] and GoogleTranslate[7].

IMPLEMENTATION. We conduct our experiments using a single NVIDIA A100 GPU with a batch size 16. For MLM, we set the learning rate to 5e-5 and train models for 10 epoch. For fine-tuning, we select the optimal learning rate from {2e-5, 5e-5} and train models for 6 epochs. For adapter-based tuning, we select the optimal learning rate from {1e-4, 2e-4, 5e-5} and train adapters for 15 epochs.

5.6 RESULTS AND ANALYSIS

5.6.1 *Supervised Learning Results*

In this setup, labeled data is available for target languages for model training. We first compare the performance on development sets using fine-tuning and adapter-based tuning, combined with various techniques, as reported in Table 5.4. Overall, fine-tuning achieves the highest performance in most languages (6 out of 9), which is not surprising given that it optimizes the entire parameter space. Notably, adapter-based tuning demonstrates comparable performance to fine-tuning in Hausa (`hau`) and Telugu (`tel`), and even surpasses it in Kinyarwanda (`kin`), Marathi (`mar`), and Spanish (`spa`), highlighting its efficiency and competitiveness.

Looking at the effectiveness of TAPT and warmup techniques in fine-tuning, we find that they generally improve performance compared to using no techniques at all. For example, applying warmup with STS-B and TAPT improves the performance of Algerian Arabic (`arq`) from 52.96 to 68.25 – a gain of over 15 points. However, the improvements are sometimes marginal, particularly for languages such as Amharic (`amh`), English (`eng`), and Moroccan Arabic (`ary`), where baseline performance is already relatively high compared to other languages. This suggests that TAPT and warmup techniques are especially beneficial for languages with more challenging linguistic characteristics.

6 https://huggingface.co/facebook/nllb-200-distilled-600M
7 https://pypi.org/project/pyGoogleTranslate/

Model Tuning	TAPT	Warmup	arq	amh	eng	hau	kin	mar	ary	spa	tel
FINE-TUNING	✗	✗	52.96	87.70	83.07	78.91	68.59	85.23	88.26	73.83	84.90
	✗	SemRel	55.96	87.86	/	79.87	70.06	85.51	**88.59**	72.93	85.38
	✗	STS-B	62.05	88.50	**84.31**	79.86	69.78	86.48	86.97	73.33	85.15
	✓	✗	65.70	88.03	82.79	79.41	67.03	84.88	88.50	70.47	83.84
	✓	SemRel	66.74	85.58	/	**80.73**	71.29	85.74	87.01	73.37	**85.77**
	✓	STS-B	**68.25**	**88.72**	83.01	78.95	69.38	85.26	87.07	73.50	84.66
ADAPTER TUNING	✗	✗	55.44	87.01	82.96	78.23	70.45	84.62	86.43	72.62	84.51
	✗	SemRel	59.58	87.66	/	79.15	70.56	86.54	86.88	74.90	84.88
	✗	STS-B	62.83	87.63	82.97	80.29	82.01	87.18	87.53	74.18	84.17
	✓	✗	58.81	85.61	82.74	78.40	70.48	84.56	85.78	72.15	84.34
	✓	SemRel	58.47	87.57	/	79.78	71.67	**87.24**	87.35	**76.65**	85.69
	✓	STS-B	59.58	87.40	82.32	79.22	**73.04**	87.12	87.22	73.22	83.70

Table 5.4: Supervised learning performance on development sets (Spearman's correlation $\times 100$). SemRel: warmup by training on SemRel translations; STS-B: warmup by training on STS-B translations. For each language, we underline the best performance of fine-tuning and adapter-based tuning, and **bold** the best performance across all variants.

Model	arq	amh	eng	hau	kin	mar	ary	spa	tel	Avg.↑
Overlap♢	40.	63.	67.	31.	33.	62.	63.	67.	70.	55.11
LaBSE♢	60.	85.	83.	69.	72.	88.	77.	70.	82.	76.22
NRK	67.36	86.42	83.29	67.20	75.69	87.93	82.70	68.99	83.42	78.11
PEAR	46.33	83.42	84.79	69.41	77.22	85.60	81.53	71.01	82.75	75.78
AAdaM (Ours)	66.23	86.71	84.84	72.36	77.91	89.43	83.50	74.04	84.77	79.98

Table 5.5: Supervised learning performance on test sets (Spearman's correlation ×100). ♢: baseline results from Ousidhoum et al. (2024b).

For the test set evaluation on the SemEval leaderboard, we selected the best-performing model for each language based on results on development sets. Table 5.5 compares our approach, AAdaM, with two official baselines (Ousidhoum et al., 2024b) and other top-performing submissions: NRK (Kiet and Thin, 2024) and PEAR (Jørgensen, 2024). NRK ensembles BERT-based models and uses a weighted voting technique to improve the performance. PEAR focuses on hyperparameter optimization and data sampling using multilingual bi-encoders. Our model demonstrates substantial improvements over all baselines, achieving particularly strong gains for Hausa (`hau`), Moroccan Arabic (`ary`), and Spanish (`spa`).

5.6.2 *Cross-lingual Transfer Results*

In this setup, the labeled task data of the target language is not accessible. Therefore, we replace the source language adapter with the target language adapter, while keeping the task adapter from the source language trained in Section 5.6.1. As demonstrated in Section 5.4.5, we explore three methods for source language selection: linguistic distance, token overlap, and development set performance. Figure 5.2 and Figure 5.3 present metrics scores for each language selection method, along with the corresponding best source languages identified. For development set performance, we consider two types of language adapters: base language adapters trained only on the Leipzig corpora and TAPT language adapters further trained on unlabeled task data.

When looking at the performance on the development sets from Figure 5.3, we first observe a discrepancy in the optimal source languages selected by two types of adapters, indicating a change in behavior after applying TAPT. Interestingly, English (`eng`) – often considered the default source language in cross-lingual transfer – is not consistently the most effective choice. Instead, the performance for target languages proves to be highly sensitive to the selection of the source language. For example, using Spanish (`spa`) as the source for Indonesian (`ind`) yields a substantially higher score (52.5 with the base adapter and 52.8

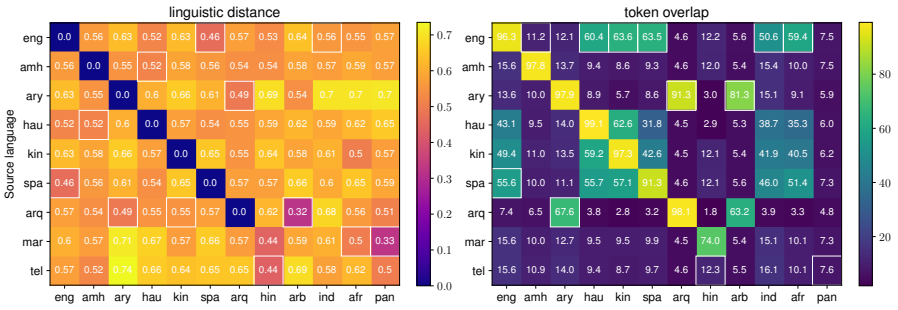


Figure 5.2: **Left**: Linguistic distances between source and target languages. The *smallest* distance for each target language is highlighted with a box. **Right**: Token overlaps between source and target languages. The *highest* overlap for each target language is highlighted with a box.

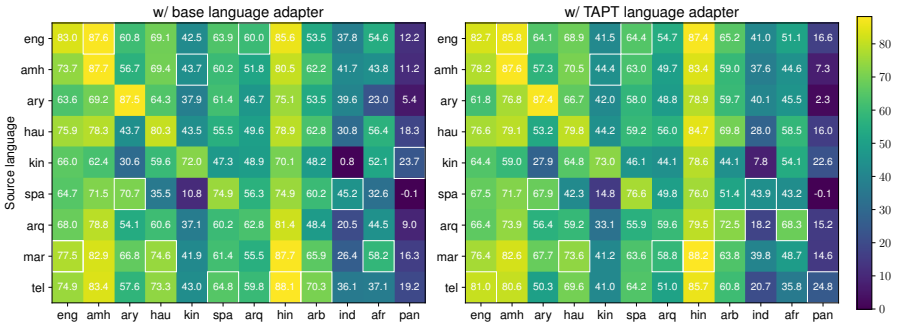


Figure 5.3: Performance on development sets (Spearman's correlation $\times 100$) using different types of language adapters. Note that when the target language is English, we use task adapters from different source languages that have not been trained on augmented datasets (translated from English). The *highest* performance for each target language is highlighted with a box.

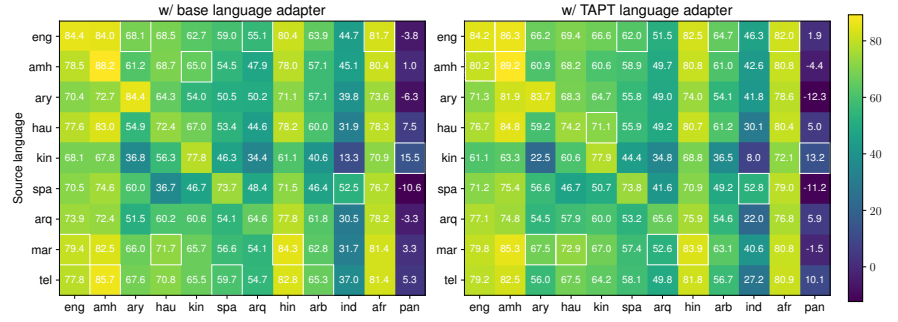


Figure 5.4: Performance on test sets (Spearman's correlation $\times 100$) using different types of language adapters. The *highest* performance for each target language is highlighted with a box, serving as the "ground-truth" selection.

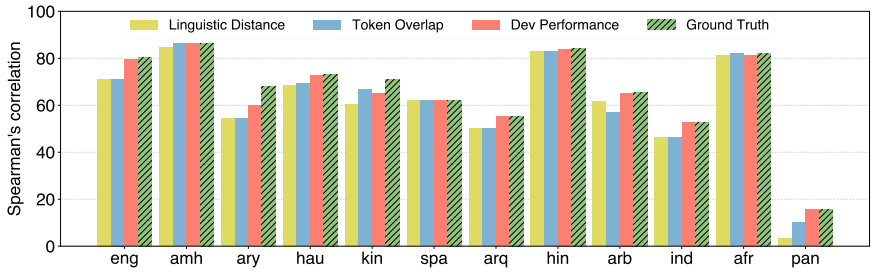


Figure 5.5: Test set performance across languages using different selection methods. For each setup, we report the best performance obtained across the two adapter types.

with the TAPT adapter) compared to using Kinyarwanda (kin) (13.3 with the base adapter and 8.0 with the TAPT adapter), highlighting the critical need for thoughtful source language selection. Among the target languages, Amharic (amh) achieves cross-lingual transfer performance comparable to its supervised learning performance. In contrast, languages like Indonesian (ind) and Punjabi (pan) remain challenging in the cross-lingual transfer setting, with their best performances reaching only 52.8 and 15.5, respectively, even when paired with the most suitable source language.

We use development set performance as the primary indicator for source language selection, as it reflects the combined effects of both the model and the data, providing a more robust measure. Following the release of the SemEval test sets, a post-hoc evaluation (Figure 5.4) confirms that this metric is indeed the most reliable indicator: the optimal source languages it identifies closely align with those determined from the ground-truth test results, and achieve the best performance in most cases as shown in Figure 5.5. However, this approach is not feasible in a true zero-shot scenario. This highlights the need for future work to develop more effective methods for estimating transfer performance a priori (Lin et al., 2019; Pruksachatkun et al., 2020).

In Table 5.6, we compare our test set performance with two baselines and several top-performing competitors. UAlberta (Shi et al., 2024) applies its English model to translations of non-English test sets. USTCCTSU (Li et al., 2024) employs a data filtering strategy to retain only training languages that positively impact the target language. UMBCLU (Roy Dipta and Vallurupalli, 2024) relies on English and Spanish fine-tuned models for all other target languages. Compared to LaBSE, a multilingual sentence embedding model, AAdaM achieves superior performance on most languages, particularly Algerian Arabic (arq), Hausa (hau), Moroccan Arabic (ary), and Punjabi (pan). However, our system is slightly outperformed by the simple word overlap baseline on Indonesian (ind), Moroccan Arabic (ary), and Spanish (spa), indicating the need for a more nuanced investigation. Additionally, AAdaM achieves consistently strong results across most languages,

Model	afr	arq	amh	eng	hau	hin	ind	kin	arb	ary	pan	spa	Avg.↑
Overlap♢	71.	40.	63.	67.	31.	53.	55.	33.	32.	63.	-27.	67.	45.67
LaBSE♢	79.	46.	84.	80.	62.	76.	47.	57.	61.	40.	-5.	62.	57.42
UAlberta	80.57	44.13	81.60	-	67.85	82.78	44.90	63.58	**67.15**	60.22	-1.74	57.16	-
USTCCTSU	74.87	41.44	70.90	78.40	47.63	65.80	46.02	45.41	46.87	61.32	-24.79	68.51	51.87
UMBCLU	82.23	12.63	4.30	78.75	45.69	15.52	51.53	48.36	3.54	-3.75	-7.75	60.89	32.66
AAdaM (ours)	81.39	55.07	86.29	79.37	72.88	83.86	52.80	64.99	65.32	60.03	15.53	62.05	64.97

Table 5.6: Cross-lingual transfer performance on test sets (Spearman's correlation ×100). ♢: baseline results from Ousidhoum et al. (2024b).

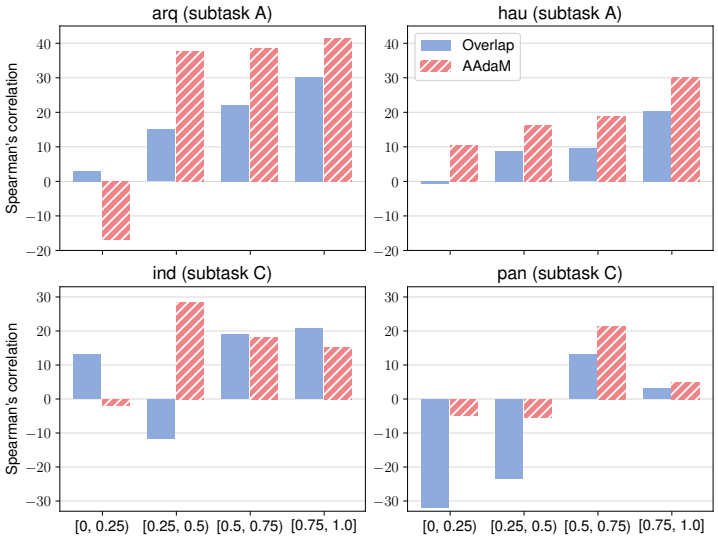


Figure 5.6: Performance on test sets (Spearman's correlation $\times 100$) in different relatedness levels. *subtask A* refers to the supervised learning setting, and *subtask C* refers to the cross-lingual transfer setting.

whereas other competitors show more uneven performance: UMBCLU performs poorly on several languages, such as Amharic (`amh`) and Moroccan Arabic (`ary`), and all baselines struggle on Punjabi (`pan`). Overall, AAdaM achieves the highest performance, with an average score of 64.97.

5.6.3 *Analysis*

For a more fine-grained analysis, we partition ground-truth relatedness scores, which range from 0 to 1, into different levels. Figure 5.6 presents the detailed model performance on several under-performing languages. Although our evaluation scores on the entire test sets remain positive, certain subsets, particularly those with lower relatedness scores, exhibit negative correlations. We can see that AAdaM falls behind the simple word overlap baseline for Algerian Arabic (`arq`) and Indonesian (`ind`) within the 0 to 0.25 score range. These observations highlight the challenge of capturing subtle semantic relationships in specific categories. The observed performance patterns may be influenced by the data annotation process, imbalanced distribution of examples across relatedness levels, or limited exposure to low-relatedness pairs during training. This suggests that further improvements may require more targeted strategies to handle low-relatedness instances and under-represented categories.

5.7 CONCLUSION

This chapter introduces AAdaM, a cross-encoder framework for multilingual semantic relatedness that achieves strong performance across both supervised learning and cross-lingual transfer settings. Our findings reveal that strategic data augmentation and task-adaptive pre-training yield substantial performance gains, while adapter-based tuning emerges as a particularly efficient and flexible approach for zero-shot cross-lingual transfer.

Although our system has demonstrated impressive performance, relying on development sets for source language selection undermines its practical value in the true zero-shot condition. While linguistic (dis)similarity (Littell et al., 2017) is a commonly used estimator for cross-lingual transfer performance, it alone does not explain many transfer results (Lauscher et al., 2020). Furthermore, it is still unclear how languages interfere with each other during pre-training and task learning phases. The challenge remains how to determine the most beneficial sources without post-hoc evaluation.

6 MULTILINGUAL IN-CONTEXT LEARNING

In this chapter, we shift to an analytical perspective, aiming to better understand how few-shot demonstrations influence in-context learning across a wide range of languages and tasks requiring nuanced semantic understanding. In-context learning is a popular inference strategy where large language models solve a task using only a few labeled demonstrations without needing any parameter updates. Although there have been extensive studies on English in-context learning, multilingual in-context learning remains under-explored, and we lack an in-depth understanding of the role of demonstrations in this context. To address this gap, we conduct a multidimensional analysis of multilingual in-context learning, experimenting with 5 models from different model families, 9 datasets covering classification and generation tasks, and 56 typologically diverse languages. Our results reveal that the effectiveness of demonstrations varies significantly across models, tasks, and languages. We also find that strong instruction-following models including Llama 2-Chat, GPT-3.5, and GPT-4 are largely insensitive to the quality of demonstrations. Instead, a carefully crafted template often eliminates the benefits of demonstrations for some tasks and languages altogether. These findings show that the importance of demonstrations might be overestimated, highlighting the need for granular evaluation across multiple axes towards a better understanding of in-context learning.[1]

6.1 INTRODUCTION

An intriguing property of large language models (LLMs) is their ability to perform in-context learning (Brown et al., 2020), i.e., solve a task conditioned on a few demonstrations at inference time, without updating the model parameters. It has been shown to be an efficient alternative to fine-tuning when adapting models to diverse tasks and domains (Dong et al., 2022; Min et al., 2022b; Si et al., 2023, *inter alia*). In light of the success of in-context learning, there has been increased interest in better understanding the factors that influence its success, such as demonstration selection (Liu et al., 2022b; Rubin et al., 2022; Wang et al., 2023d), prompt design (Min et al., 2022a; Wei et al., 2022), and more generally on understanding how and why in-context learn-

1 This chapter is based on (Zhang et al., 2024a). As the first author, Miaoran Zhang led the project, conducted the experiments, and was the main writer of the paper. The source code for this work is available on Github: https://github.com/uds-lsv/multilingual-icl-analysis.

ing works (Xie et al., 2022a; Bansal et al., 2023; Hendel et al., 2023; Pan et al., 2023; Wang et al., 2023b).

However, most recent work on in-context learning predominantly focuses on English, and the exploration of multilingual in-context learning generally lags behind. This is problematic, as results that apply to English might not hold for other languages, especially those that are less represented in the training data. While there have been a few studies on in-context learning that go beyond English, they either focus on benchmarking LLMs on multilingual tasks without in-depth exploration, e.g., MEGA (Ahuja et al., 2023) and BUFFET (Asai et al., 2024), or zoom in on specific capabilities such as mathematical reasoning (Shi et al., 2023b), machine translation (Zhu et al., 2024b; Agrawal et al., 2023), or code-switching (Zhang et al., 2023b).

In this chapter, we take a multidimensional approach (Ruder et al., 2022) that unifies these strands of research and comprehensively evaluate the multilingual in-context learning abilities of LLMs. We focus on dissecting the *actual* impact of in-context demonstrations, which is crucial for understanding model behaviour. Our research covers various models, tasks, and languages, and we seek to answer the following research questions:

1. Does multilingual performance benefit from demonstrations?
2. Does demonstration quality matter?
3. What is the interplay between demonstrations and templates?
4. How do the answers to these questions vary across languages and models?

Specifically, we address our research questions by evaluating 5 LLMs including base models that are only pre-trained on unlabeled text corpora (XGLM and Llama 2), and chat models that are further refined with instruction tuning and reinforcement learning (Llama 2-Chat, GPT-3.5, and GPT-4). We evaluate on 9 multilingual datasets that include both classification and generation tasks, covering 56 typologically different languages.

Our main findings are: (1) The effectiveness of demonstrations varies widely depending on the model, task, and language used. For base models, in-context learning barely outperforms zero-shot learning on many tasks. In general, in-context learning matters more for generation tasks with loosely-specified prompts; (2) Even with sophisticated demonstration selection methods, in-context learning is not always beneficial and can sometimes be worse than using no demonstrations at all; (3) Chat models are less sensitive to seeing correctly-labeled demonstrations than base models, suggesting that for the former, demonstrations primarily help the model understand the task format, while for the latter, demonstrations also impart task-specific knowledge; (4) Using a formatting-focused template can even

eliminate the need for demonstrations with chat models. The relative significance of demonstrations versus prompt templates varies based on inherent model capabilities.

In sum, we suggest that the benefits of adding demonstrations may be overestimated. Future work on in-context learning should carefully compare their results with zero-shot learning and on multiple templates to faithfully represent its effectiveness. Given the vast variance across models, tasks, and languages, it is also important to cautiously frame claims about in-context learning.

6.2 RELATED WORK

Prompt-based learning. Prompt-based learning has emerged as a new paradigm for adapting LLMs to new scenarios without updating model parameters. It leverages task-specific instructions, or prompts, to enable zero-shot learning (Radford et al., 2019) or few-shot learning (i.e., in-context learning) with few labeled examples (Brown et al., 2020; Mishra et al., 2022a). Chain-of-Thought (CoT) prompting further enhances reasoning by encouraging models to generate intermediate steps (Wei et al., 2022; Kojima et al., 2022), inspiring more structured variants such as Tree-of-Thought (Yao et al., 2023) and Graph-of-Thought reasoning (Yao et al., 2024). In this work, we focus on in-context learning, where labeled demonstrations are provided to the model as conditions for answer generation.

Multilingual in-context learning. Most multilingual in-context learning studies focus on benchmarking LLMs on diverse tasks and comparing them with smaller fine-tuned models (Ahuja et al., 2023; Asai et al., 2024; Zhang et al., 2023b; Zhu et al., 2024b). As these works focus on benchmarking, their analysis of the role of demonstrations is limited. Ahuja et al. (2023) explore different prompting strategies by adjusting the language of templates and demonstrations. Zhang et al. (2023b) find that demonstrations sometimes do not contribute to or even degrade model performance on code-switching. Zhu et al. (2024b) look at machine translation and analyze the effects of template and demonstration selection with XGLM. In the context of cross-lingual transfer, Shi et al. (2022), Tanwar et al. (2023), and Agrawal et al. (2023) investigate demonstration selection for specific applications. In contrast, we take a much broader perspective and investigate the actual impact of demonstrations across a wide range of models, tasks and languages.

Demonstration analysis. Most of the current demonstration analysis literature focuses on English: Lu et al. (2022) analyze the sensitivity of in-context learning to the order of demonstrations, Min et al. (2022c) and Yoo et al. (2022) explore whether the ground truth labels matter for classification tasks, and Wei et al. (2023) investigate the sensitivity of various model families to different input-label mappings. Similarly,

Pan et al. (2023) disentangle task recognition and task learning by manipulating the label space. Beyond this, Shi et al. (2023a) and Wang et al. (2023a) modify the validity of CoT reasoning steps in demonstrations and explore the impact of this modification on mathematical reasoning. Also focusing on CoT, Chen et al. (2023) investigate how varying the number of demonstrations affects performance.

6.3 BACKGROUND

6.3.1 *In-Context Learning*

In-context learning (ICL) is a popular inference strategy where models solve[2] a task without any parameter updates (Brown et al., 2020). Instead, the model performs the task by conditioning on *labeled demonstrations*. Demonstrations are typically formatted using "pattern-verbalizer pairs," as this has been shown to be effective in eliciting good task performance (Schick and Schütze, 2021b; Bach et al., 2022). Here, a *pattern* is used to format the input for the model, and a *verbalizer* maps the label to a textual representation. Additionally for instruction-tuned LLMs, a *task instruction* is often added to provide information about the task beyond individual demonstrations (Mishra et al., 2022b; Wang et al., 2022; Ouyang et al., 2022).

Formally, given a test sample x_t, k demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$, a pattern $\mathcal{P}$, a verbalizer $\mathcal{V}$ and a task instruction $\mathcal{I}$, the model (parameterized by θ) makes its prediction as follows:

$$y_t \sim p_\theta(y|\mathcal{I}, \{(\mathcal{P}(x_i), \mathcal{V}(y_i))\}_{i=1}^{k}, \mathcal{P}(x_t)). \tag{6.1}$$

Taken together, the pattern, the verbalizer and the optional task instruction comprise the *template* with which demonstrations and the test sample are formatted as the input prompt for model inference. The effectiveness of demonstrations is thus linked with the template used to present them to the model.

6.3.2 *Multilingual Prompting*

Previous studies highlight that the selection of demonstrations and prompt templates can significantly influence model performance (Liu et al., 2022b; Fu et al., 2023b; Sclar et al., 2024). In multilingual in-context learning, the variation in input prompts is further complicated by the *language* of demonstrations, templates and test samples, all of which are important design choices.

For the template language, Lin et al. (2022) and Ahuja et al. (2023) found that English templates generally perform better than native

2 The extent to which models actually "solve" tasks is an open question as ICL, similar to fine-tuning, has generalization issues despite its impressive results (Mosbach et al., 2023). Regardless, we use the word "solve" in the rest of this paper for simplicity.

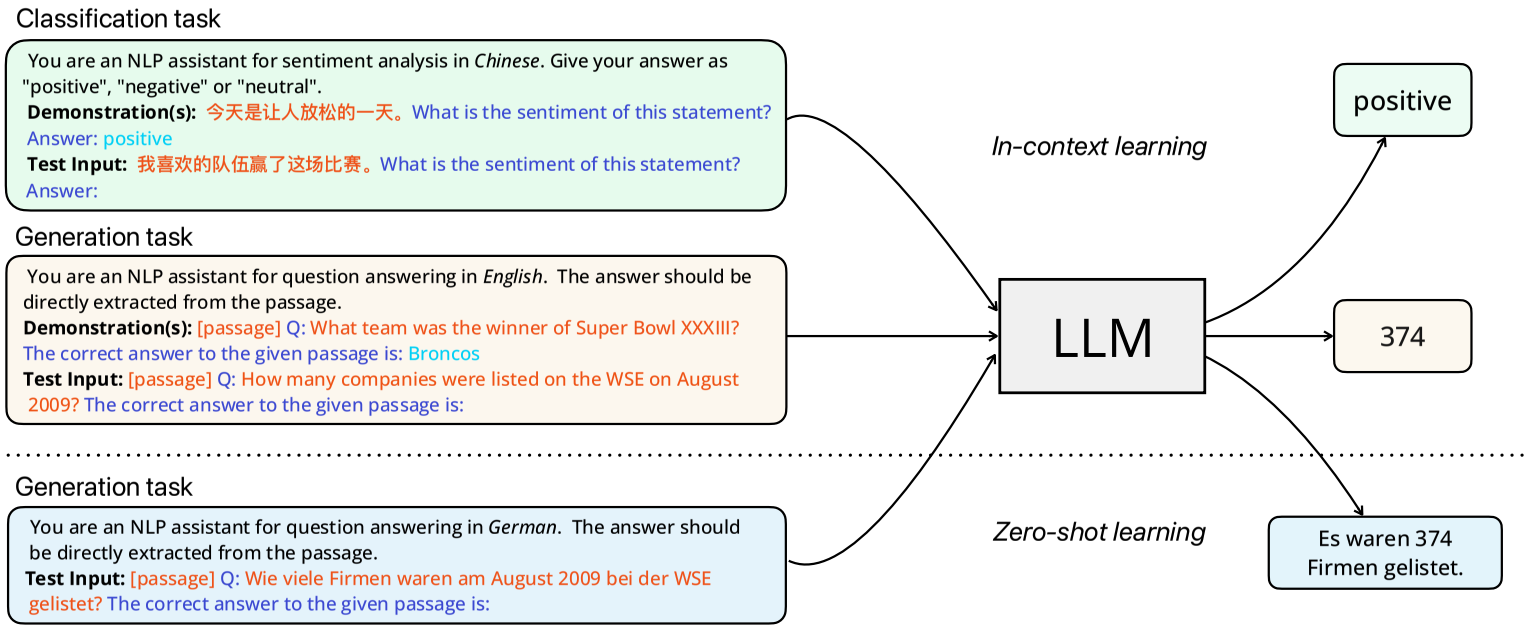


Figure 6.1: An overview of the components of multilingual in-context learning with a comparison to zero-shot learning. Sources of variation include tasks, languages, models, and the template, i.e., the task instruction, patterns for formatting inputs, and verbalized labels.

language templates, possibly due to superior instruction-following abilities on existing LLMs on English compared to other languages. Following this, we use English templates in our study.

For the language of few-shot demonstrations and test samples, there are three popular settings. Given a test sample in a certain language, the most straightforward approach is to use demonstrations in the same language (referred to as *in-language demonstrations*). This setting directly measures the model's inherent ability to solve problems in that language. Another choice is to use *English demonstrations* regardless of the language of the test sample. This is a cross-lingual transfer setup, where the goal is to transfer knowledge from a pivot language to a target language via in-context learning. As highlighted in Shi et al. (2023b) and Ahuja et al. (2023), in-language demonstrations often outperform English demonstrations on diverse multilingual tasks. Yet another option is to translate the test sample into English – an approach called *translate-test*, where the demonstrations are also in English. While translate-test leads to strong performance (Ahuja et al., 2023), this approach heavily relies on a translation system for data processing and centers the English proficiency of LLMs. In this work, we are interested in dissecting the intrinsic multilingual capabilities of LLMs, therefore we choose to use *in-language demonstrations*.

All these design choices are represented visually in Figure 6.1, which gives an overview of multilingual in-context learning. Detailed setup information is provided in the next section.

6.4 EXPERIMENTAL SETUP

6.4.1 *Models*

We evaluate two types of LLMs: pre-trained base models and chat models. Our base models include XGLM (Lin et al., 2022) and Llama 2 (Touvron et al., 2023). Our chat models are Llama 2-Chat, GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI et al., 2023). Specifically, we use `xglm-7.5B`, `Llama-2-13b`, and `Llama-2-13b-chat` on Huggingface (Wolf et al., 2020), and get access to `gpt-3.5-turbo-16k` and `gpt-4-32k` APIs via Microsoft Azure. Additionally, we performed exploratory experiments for BLOOMZ (`bloomz-7b1`) and mT0 (`mt0-xxl`) (Muennighoff et al., 2023b), which are instruction-tuned models, but exclude them from our main comparative framework due to their training scheme and the performance patterns demonstrated in Section 6.5.1.

6.4.2 *Tasks and Datasets*

We experiment on a diverse range of multilingual classification and generation tasks, using 9 datasets covering 56 languages in total. Our dataset selection largely follows MEGA, but we add datasets for ex-

Dataset	Task	Languages	|Lang.|	Release Date
XNLI	natural language inference	English, German, Russian, French, Spanish, Chinese, Vietnamese, Turkish, Arabic, Greek, Thai,Bulgarian, Hindi, Urdu, Swahili	15	2019.09
IndicXNLI	natural language inference	Hindi, Bengali, Tamil, Marathi, Malayalam, Telugu, Kannada, Punjabi, Oriya, Assamese, Gujarati	11	2022.04
PAWS-X	paraphrase identification	English, German, Japanese, French, Spanish, Chinese, Korean	7	2019.08
XCOPA	commonsense reasoning	Chinese, Italian, Vietnamese, Indonesian, Turkish, Thai, Estonian, Tamil, Swahili, Haitian, Quechua	11	2020.04
XStoryCloze	commonsense reasoning	English, Russian, Spanish, Chinese, Indonesian, Arabic, Hindi, Basque, Telugu, Burmese, Swahili	11	2023.05
AfriSenti	sentiment analysis	Swahili, Amharic, Hausa, Kinyarwanda, Yoruba, Tigrinya, Igbo, Oromo, Moroccan Arabic, Algerian Arabic, Nigerian Pidgin, Mozambican Portuguese, Tsonga, Twi	14	2023.05
XQuAD	extractive QA	English, German, Russian, Spanish, Chinese, Vietnamese, Turkish, Greek, Romanian, Thai, Hindi	12	2019.10
TyDiQA-GoldP	extractive QA	English, Russian, Indonesian, Korean, Arabic, Finnish, Bengali, Telugu, Swahili	9	2020.02
MAFAND	machine translation	Amharic, Hausa, Kinyarwanda, Luganda, Luo, Chichewa, Nigerian Pidgin, Shona, Swahili, Setswana, Twi, Xhosa,Yoruba, Zulu	14	2022.06

Table 6.1: Multilingual benchmarking datasets. As the black-box training data of OpenAI APIs that we used is up to September 2021, we include the dataset release date in the table, which can serve as an indicator of possible dataset contamination.

tremely under-represented African languages. Our classification tasks include natural language inference (NLI), paraphrase identification, commonsense reasoning and sentiment analysis, with the following datasets: XNLI (Conneau et al., 2018), IndicXNLI (Aggarwal et al., 2022), PAWS-X (Yang et al., 2019), XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2022) and AfriSenti (Muhammad et al., 2023). Our generation tasks are extractive question answering (QA) and machine translation (MT), for which we use XQuAD (Artetxe et al., 2020), TyDiQA-GoldP (Clark et al., 2020), and MAFAND (Adelani et al., 2022). For the machine translation dataset MAFAND, English serves as the pivot language and there are two translation directions: en-xx (i.e., translating from English to another language) and xx-en (i.e., translating from another language to English). See Table 6.1 for dataset details.

6.4.3 *In-Context Learning*

All in-context demonstrations are in the same language as the test sample, and all templates are in English. The demonstrations are randomly sampled by default, and we explore two other types of demonstrations in Section 6.5.2. More specifically, for each test sample, we select $k \in \{0, 2, 4, 8\}$ different demonstrations, and for QA datasets, we select a maximum of 4 demonstrations due to context size limitations. These few-shot demonstrations are sampled from the validation set, and the test set is used only for evaluation. For datasets without a test data split (XStoryCloze and TyDiQA), we sample few-shots from the train set and evaluate the validation set. Since XQuAD only has a validation data split, we utilize it for both demonstration sampling and evaluation, ensuring that the test sample itself is not included in its demonstrations. For chat models (Llama 2-Chat, GPT-3.5, and GPT-4), we limit the test sample size to a maximum of 200 in order to reduce inference expenses and ensure a fair comparison.

Task	Pattern	Verbalizer
NLI	`{premise}`, right? `{label}`, `{hypothesis}`	Yes || Also || No
PAWS-X	`{sentence1}`, right? `{label}`, `{sentence2}`	No || Yes
XCOPA	`{premise}` {% if question == "cause" %}because{% else %} so{% endif %} `{label}`	`{choice1}` || `{choice2}`
XStoryCloze	`{input_sentence_1}` `{input_sentence_2}` `{input_sentence_3}` `{input_sentence_4}` `{label}`	`{sentence_quiz_1}` || `{sentence_quiz_2}`
AfriSenti	`{tweet}` The sentiment of the previous sentence is `{label}`	positive || neutral || negative
QA	`{context}\nQ:{question}\nA:{answer}`	`{answer}`
MT	`{source_sentence}` = `{target_sentence}`	`{target_sentence}`

Table 6.2: Prompting templates for XGLM and Llama 2.

We use appropriate task-specific templates for different model types. For XGLM and Llama 2, we follow Brown et al. (2020) and Lin et al.

Task	Pattern	Verbalizer
NLI	{premise} Based on the previous passage, is it true that {hypothesis}? Yes, No, or Maybe? {label}	Yes || Maybe || No
PAWS-X	Sentence 1: {sentence1}\n Sentence 2: {sentence2}\n Question: Can we rewrite Sentence 1 to Sentence 2? Yes or No? {label}	No || Yes
XCOPA	{premise} {% if question == "cause" %}This happened because... {% else %} As a consequence...{% endif %}\n Help me pick the more plausible option:\n - {choice1}\n - {choice2}\n {label}	{choice1} || {choice2}
XStoryCloze	{input_sentence_1} {input_sentence_2} {input_sentence_3} {input_sentence_4}\n What is a possible continuation for the story given the following options?\n - {sentence_quiz_1}\n - {sentence_quiz_2}\n {label}	{sentence_quiz_1} || {sentence_quiz_2}
AfriSenti	{tweet} Would you rate the previous sentence as positive, neutral or negative? {label}	positive || neutral || negative
QA	{context}\nQ:{question}\nReferring to the passage above, the correct answer to the given question is:{answer}	{answer}
MT	Translate the following {src_language} text to {tgt_language}:\n {src_sentence}\n{tgt_sentence}	{tgt_sentence}

Table 6.3: Prompting templates for BLOOMZ and mT0.

(2022) to use GPT-3 style prompting templates, i.e. cloze tests, as shown in Table 6.2. The templates used for BLOOMZ and mT0 are shown in Table 6.3, following the design choices in Bach et al. (2022) and Muennighoff et al. (2023b). For chat models, including Llama 2-Chat, GPT-3.5 and GPT-4, the default templates are shown in Table 6.4, for which we follow Ahuja et al. (2023) and Ojo et al. (2023) and add language identifiers in task instructions as it is an effective strategy for improving multilingual prompting (Huang et al., 2023). We also design formatting-focused templates to reinforce LLM to generate formatted outputs, as shown in Table 6.5. The effect of this modification on model performance is analyzed in Section 6.5.3.

6.4.4 *Evaluation Metrics*

For classification tasks, we report the rank classification accuracy[3] for open-source base models (Muennighoff et al., 2023b; Lin et al., 2022). For chat models, we measure the exact match between generated out-

3 The scoring function is the average of per-token log probabilities (ignoring the common prefix of different candidates). The candidate with the highest score is chosen as the prediction.

Task	Template
NLI	**task instruction**: You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems in <EVALUATION_LANGUAGE>. NLI is the task of determining the inference relation between two (short, ordered) texts: entailment, contradiction, or neutral. Answer as concisely as possible in the same format as the examples below: **pattern**: `{premise}\nQuestion: {hypothesis}\nTrue, False, or Neither?` **verbalizer**: True || Neither || False
PAWS-X	**task instruction**: You are an NLP assistant whose purpose is to perform Paraphrase Identification in <EVALUATION_LANGUAGE>. The goal of Paraphrase Identification is to determine whether a pair of sentences have the same meaning. Answer as concisely as possible in the same format as the examples below: **pattern**: `{sentence1}\nQuestion: {sentence2}\nTrue or False?` **verbalizer**: False || True
XCOPA	**task instruction**: You are an NLP assistant whose purpose is to perform open-domain commonsense causal reasoning in <EVALUATION_LANGUAGE>. You will be provided a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. Answer as concisely as possible in the same format as the examples below: **pattern**: Premise: `{premise}\n`What is the `{question}`? Pick the more plausible option:`\n` 1: `{choice1}\n`2: `{choice2}\n` You should tell me the choice number in this format 'Choice number:' **verbalizer**: Choice number: 1 || Choice number: 2
XStoryCloze	**task instruction**: You are an NLP assistant whose purpose is to perform open-domain commonsense causal reasoning in <EVALUATION_LANGUAGE>. You will be provided a four-sentence story and two continuations, where the task is to select the correct ending. Answer as concisely as possible in the same format as the examples below: **pattern**: Story: `{input_sentence_1} {input_sentence_2} {input_sentence_3} {input_sentence_4}\n` What is a possible continuation for the story? Pick the more plausible option:`\n` 1: `{sentence_quiz1}\n`2: `{sentence_quiz2}\n` You should tell me the choice number in this format 'Choice number:' **verbalizer**: Choice number: 1 || Choice number: 2
AfriSenti	**task instruction**: You are an NLP assistant whose purpose is to perform Sentiment Analysis in <EVALUATION_LANGUAGE>. Sentiment Analysis is the task of determining the sentiment, opinion or emotion expressed in a textual data. Give your answer as a single word, "positive", "neutral" or "negative". **pattern**: Does this statement "`{tweet}`" have a {positive neutral or negative} sentiment? Labels only **verbalizer**: positive || neutral || negative
QA	**task instruction**: You are an NLP assistant whose purpose is to solve reading comprehension problems in <EVALUATION_LANGUAGE>. You will be provided questions on a set of passages and you will need to provide the answer as it appears in the passage. The answer should be in the same language as the question and the passage. **pattern**: `{context}\nQ: {question}\n`Referring to the passage above, the correct answer to the given question is: **verbalizer**: `{answer}`
MT	**pattern**: Translate the following `{src_language}` text to `{tgt_language}`: `{src_sentence}` **verbalizer**: `{tgt_sentence}`

Table 6.4: Prompting templates for Llama 2-Chat, GPT-3.5, and GPT-4. Task instructions are used to assign a system role to the model.

Task	Template
XCOPA	**task instruction**: You are an NLP assistant whose purpose is to perform open-domain commonsense causal reasoning in <EVALUATION_LANGUAGE>. You will be provided a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. Answer as concisely as possible in the same format as the examples below: **pattern**: Premise: `{premise}`\nWhat is the `{question}`? Pick the more plausible option:\n 1: `{choice1}`\n2: `{choice2}`\n This is very important: Do not repeat the question and no explanation. You should tell me the choice number in this format 'Choice number:' **verbalizer**: Choice number: 1 || Choice number: 2
AfriSenti	**task instruction**: You are an NLP assistant whose purpose is to perform Sentiment Analysis in <EVALUATION_LANGUAGE>. Sentiment Analysis is the task of determining the sentiment, opinion or emotion expressed in a textual data. Give your answer as a single word, "positive", "neutral" or "negative". **pattern**: Does this statement "`{tweet}`" have a {positive neutral or negative} sentiment? This is very important: Do not repeat the question and no explanation. Labels only **verbalizer**: positive || neutral || negative
QA	**task instruction**: You are an NLP assistant whose purpose is to solve reading comprehension problems in <EVALUATION_LANGUAGE>. Answer the question from the given passage. Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. **pattern**: `{context}`\nQ: `{question}`\nThis is very important: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. **verbalizer**: `{answer}`

Table 6.5: Formatting-focused templates for chat models. We augmented the original templates in Table 6.4 with formatting-focused instructions.

puts[4] and verbalized labels (Ahuja et al., 2023). As for generation tasks, we report the F1 score for QA datasets and ChrF++ score (Popović, 2017) for MT.

6.4.5 *Implementation Details*

Our codebase is adapted from OpenICL (Wu et al., 2023). We use int8 model quantization for all models except OpenAI APIs. In our preliminary experiments, we found that int8 quantization led to a performance degradation of 1% – 2% on a few classification datasets with Llama 2 and XGLM. Since this degradation is consistent across different setups, we believe that it would not affect our overall findings. Our experiments are conducted using a single NVIDIA A100-80GB GPU. As models have a maximum context length, we preserve complete demonstrations that can fit within the context window. We employ greedy decoding for model generation. For chat models, the maximum new token is set to 50, while for machine translation, it is set to 100. For other models, the maximum new token is set to 20, while for machine translation, it is set to 50. We use three random seeds $(0, 33, 42)$ in our experiments to ensure the reliability of our results.

6.5 MULTIDIMENSIONAL ANALYSIS

6.5.1 *Number of Demonstrations*

In this section, we systematically compare ICL and zero-shot learning as this question is under-explored in previous studies of multilingual ICL (Ahuja et al., 2023; Asai et al., 2024). We examine model performance on diverse multilingual tasks while varying the number of demonstrations, and show the results for classification tasks and generation tasks in Figure 6.2.

We begin with the overall trends across models and datasets. OpenAI's GPT-3.5 and GPT-4 models achieve the best multilingual in-context learning performance on all our datasets, which is unsurprising as they are currently the state-of-the-art on a large suite of NLP benchmarks. The next best models are Llama 2 and Llama 2-Chat, which demonstrate comparable or superior performance to the multilingual XGLM model despite being trained primarily on English corpora (Touvron et al., 2023). This indicates that their task-solving abilities can transfer across languages. Regardless of the model, however, performance on the AfriSenti and MAFAND datasets, particularly when translating English to African languages, lags significantly behind other tasks, showing that language discrepancies remain even in the best models.

4 We extract verbalized labels from the generated outputs using regular expressions before calculating the exact match.

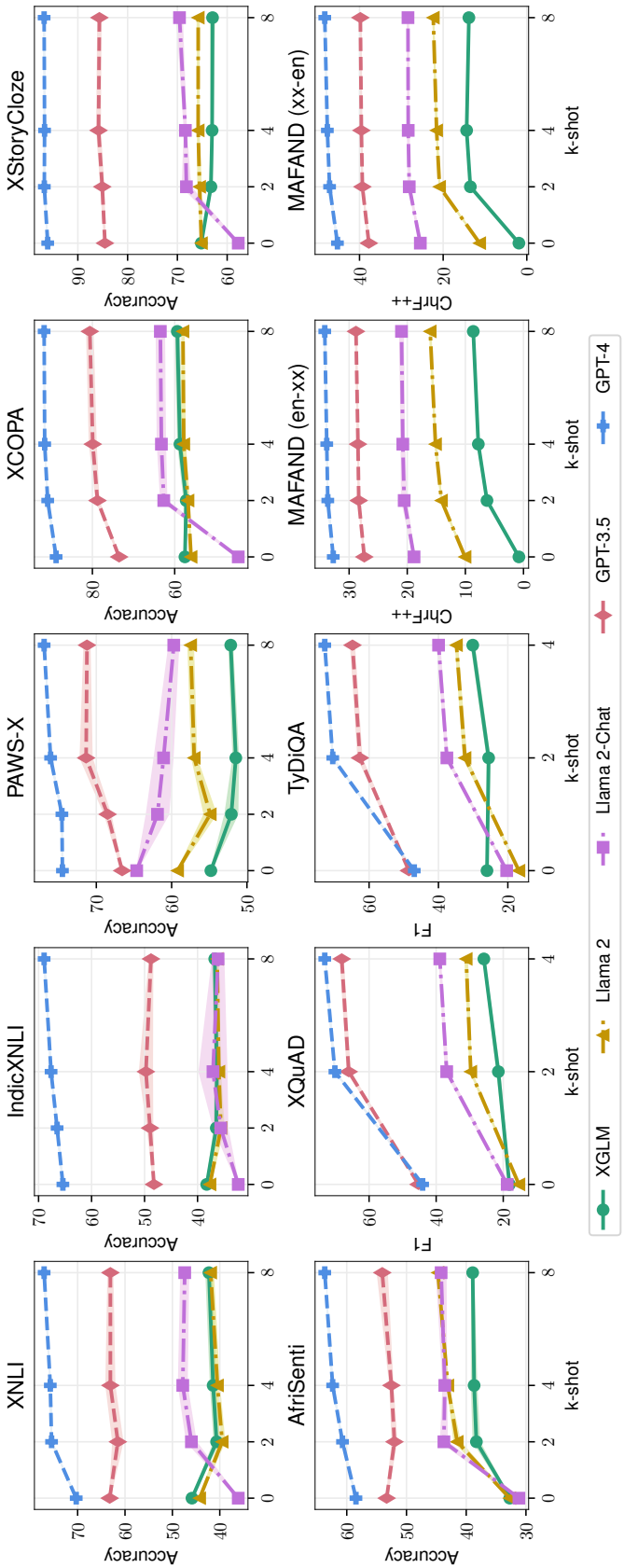


Figure 6.2: Average performance across languages with different numbers of demonstrations. We average and report standard deviations over 3 seeds for all models except GPT-4. Note that the standard deviations are relatively small, possibly because of averaging over languages. en-xx: translating from English to another language, xx-en: translating from another language to English.

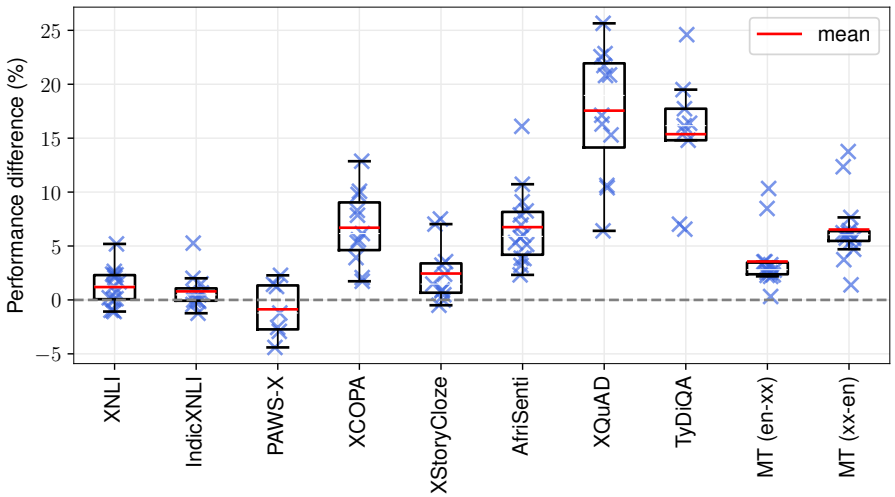


Figure 6.3: Performance difference between 4-shot and 0-shot. Each marker represents the average performance across models for each language in a given task. MT refers to the MAFAND dataset.

An important pattern across datasets and models is that **in-context learning does not always improve over zero-shot learning** – in particular, it helps with generation tasks, but results on classification tasks are mixed. For the AfriSenti dataset, many models show noticeable improvements with ICL. However, with other tasks such as IndicXNLI, XNLI and PAWS-X, the same models, especially base models, perform much worse compared to the zero-shot setting. We also see marginal improvements in some cases, e.g., XGLM and Llama 2 on XCOPA. In comparison to chat models, the addition of demonstrations typically reduces the performance of base models across many tasks. When examining the cases where ICL improves performance, we see that **improvements saturate quickly with 2 to 4 demonstrations**. This aligns with Chen et al. (2023), who found that reducing the number of demonstrations to one does not significantly deteriorate chain-of-thought reasoning.

Looking at the improvements over zero-shot performance (for all models and languages combined) across tasks in Figure 6.3, we observe that there are large fluctuations between individual languages that are not captured by the average. The PAWS-X dataset in particular shows an average degradation, but in fact some languages benefit from ICL while others degrade. For a more nuanced understanding of language-specific differences within a task, we zoom into this dataset in Figure 6.4 to inspect these language-specific differences.[5] We see that languages and models can behave very differently even on just one dataset, and a pattern which holds for one language with one model does not necessarily apply to a different language. For example, the ICL performance of Llama 2 outperforms its zero-shot performance by 2.3

5 Plots for other datasets are provided in Appendix A.1.

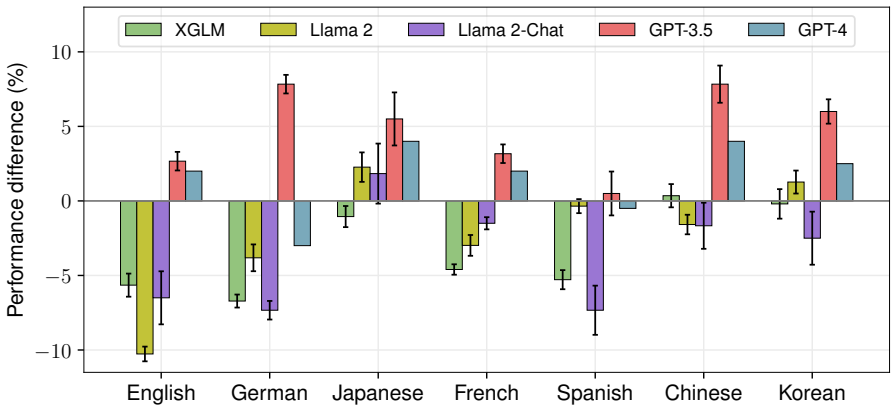


Figure 6.4: Performance difference between 4-shot and 0-shot for individual languages in PAWS-X. Error bars represent standard deviations calculated over 3 seeds.

points on Japanese and 1.3 points on Korean. However, demonstrations degrade performance for other languages, e.g., English performance degrades by 10.3 points. In sum, **the effectiveness of demonstrations varies widely depending on the model, task, and language.**

In addition to the 5 base and chat models we discussed above, we also experiment with two instruction-tuned models: BLOOMZ and mT0. The results for varying numbers of random demonstrations are shown in Figure 6.5. In line with findings from Asai et al. (2024), we observe significant performance degradation when using demonstrations compared to zero-shot learning in all cases. This decline can be attributed to their training scheme, where models are fine-tuned on a large collection of existing datasets in a zero-shot manner. In contrast, several studies (Chen et al., 2022b; Wang et al., 2022) focus on enhancing the ICL ability of LLMs by incorporating demonstrations into their training process. These results highlight the importance of considering the training strategy of a model when evaluating or comparing its learning behavior.

6.5.2 *Demonstration Quality*

Our previous experiments evaluated ICL using randomly selected demonstrations. To ablate for the effects of demonstration quality, this section experiments with the choice of demonstrations as well as the importance of ground truth labels, i.e., the input-label mapping. Inspired by work on demonstration selection (Liu et al., 2022b; Rubin et al., 2022) and input-label mapping (Min et al., 2022c; Yoo et al., 2022) in English, we compare the following three types of demonstrations:

- RANDOM: demonstrations are randomly selected from clean data

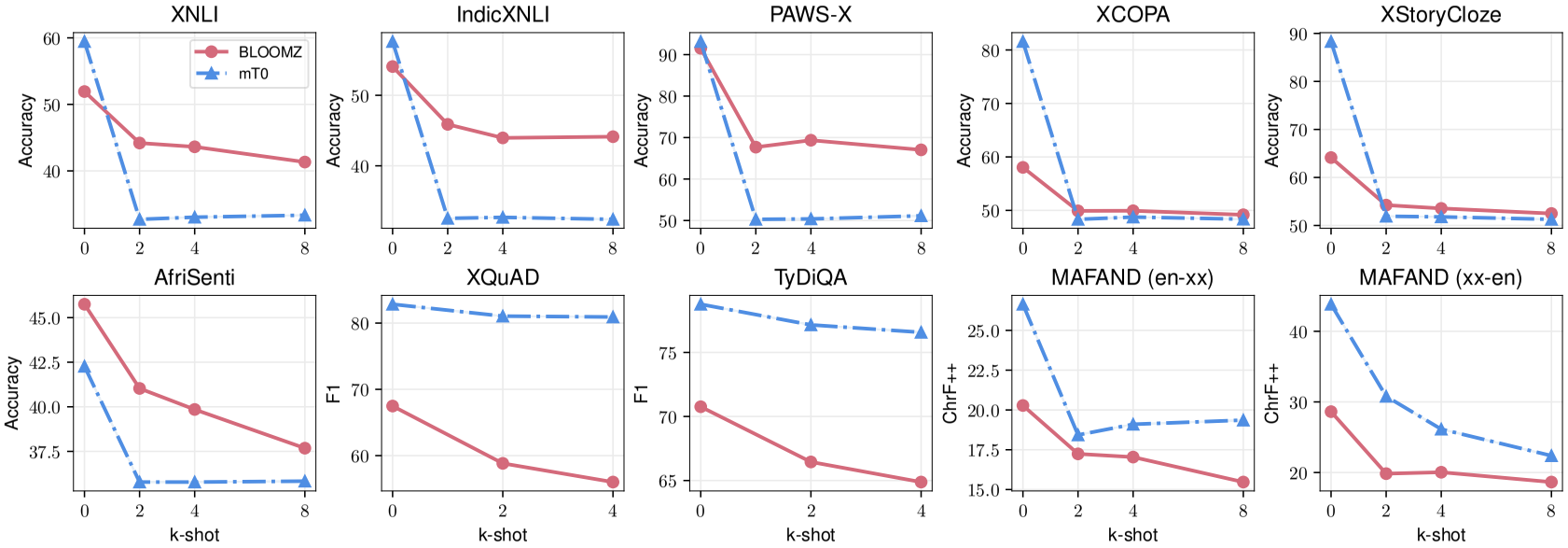


Figure 6.5: Average performance across languages for BLOOMZ and mT0 with different numbers of demonstrations. The results are obtained with a single random seed. Note that PAWS-X, XQuAD and TyDiQA are included in the instruction-tuning datasets of BLOOMZ and mT0. The single-seed results are obtained with the seed 0.

Model	XNLI	IndicXNLI	PAWS-X	XCOPA	XStoryCloze	AfriSenti
XGLM	4.59	2.49	$0.24_{\triangledown}$	0.03	$0.97_{\triangledown}$	5.62
Llama 2	6.61	4.17	2.35	−0.11	0.33	4.17
Llama 2-Chat	−0.28	−1.36	$-1.71_{\triangledown}$	0.32	0.43	2.17
GPT-3.5	0.18	0.71	−2.07	0.86	−0.61	$-0.66_{\triangledown}$
GPT-4	0.76	−0.19	0.07	−0.36	0.05	−0.68

Table 6.6: Performance difference of 4-shot ICL with TOP-K vs. RANDOM selection for **classification tasks**. Positive numbers show that TOP-K is better than RANDOM (expected), and highlighted cells show where top-k is even worse than random selection. ▽: TOP-K performance is even worse than zero-shot learning. For RANDOM, we average over 3 seeds (except for GPT-4).

Model	XQuAD	TyDiQA	MAFAND (en-xx)	MAFAND (xx-en)
XGLM	1.77	4.21	1.31	0.66
Llama 2	1.32	0.54	2.15	1.35
Llama 2-Chat	1.02	2.42	0.74	0.66
GPT-3.5	−0.34	2.98	0.72	0.43
GPT-4	−0.77	1.88	1.21	0.65

Table 6.7: Performance difference of 4-shot ICL with TOP-K vs. RANDOM selection for **generation tasks**. Positive numbers show that TOP-K is better than RANDOM (expected), and highlighted cells show where top-k is even worse than random selection. ▽: TOP-K performance is even worse than zero-shot learning. For RANDOM, we average over 3 seeds (except for GPT-4).

- TOP-K: the k most semantically similar[6] examples to a given test sample are selected (Liu et al., 2022b)
- RANDOM-CORRUPTED: demonstrations are randomly selected but the labels are corrupted by replacement with random labels (Min et al., 2022c)[7]

Table 6.6 and Table 6.7 shows that top-k selection performs better than random selection in many cases, especially for the base models XGLM and Llama 2. For chat models, the largest improvements are on generation tasks. For example, GPT-3.5 achieves a 2.98 point improvement on TyDiQA. Nevertheless, top-k selection often degrades performance on many other tasks, e.g., GPT-3.5 is 2.07 points worse on

6 We quantify semantic similarity using LaBSE (Feng et al., 2022), a multilingual sentence embedding model trained on 109+ languages.

7 For classification tasks, we randomly choose a label from the fixed label set. For generation tasks, we randomly choose a label from the label space of the entire demonstration data.

Model	XNLI	IndicXNLI	PAWS-X	XCOPA	XStoryCloze	AfriSenti
XGLM	0.46	−0.05	0.44	0.51	0.62*	3.78*
Llama 2	0.96*	0.43	1.16	0.61*	1.12*	2.27*
Llama 2-Chat	−0.34	0.04	1.48	0.03	−0.23	0.77*
GPT-3.5	0.39	1.02	0.64	0.26	0.58*	−0.62
GPT-4	−0.86	−0.04	0.57	0.86	1.13	0.90

Table 6.8: Performance difference of 4-shot ICL with RANDOM vs. RANDOM-CORRUPTED demonstrations for **classification tasks**. Positive numbers show that RANDOM is better than RANDOM-CORRUPTED (expected), and highlighted cells show where corrupted labels perform even better than ground-truth labels. We average over 3 seeds (except for GPT-4). ∗: a significant difference ($p = 0.05$).

Model	XQuAD	TyDiQA	MAFAND (en-xx)	MAFAND (xx-en)
XGLM	24.56*	26.64*	3.18*	6.73*
Llama 2	26.68*	29.20*	4.79*	8.34*
Llama 2-Chat	5.94*	4.37*	1.13*	1.53*
GPT-3.5	5.46*	5.61*	1.39*	0.48*
GPT-4	9.60	6.97	1.24	0.64

Table 6.9: Performance difference of 4-shot ICL with RANDOM vs. RANDOM-CORRUPTED demonstrations for **generation tasks**. Positive numbers show that RANDOM is better than RANDOM-CORRUPTED (expected), and highlighted cells show where corrupted labels perform even better than ground-truth labels. We average over 3 seeds (except for GPT-4). ∗: a significant difference ($p = 0.05$).

PAWS-X compared to random selection. When compared to zero-shot performance, ICL with top-k selection is even *worse* in some cases, such as XGLM on PAWS-X and XStoryCloze. In cases where random selection performs worse than zero-shot, even top-k selection gives only marginal improvements. These findings indicate that **sophisticated demonstration selection methods are not always beneficial and can sometimes be worse than using no demonstrations at all**.

Exploring this further, in Table 6.8 and Table 6.9, we compare randomly selected demonstrations with ground truth labels and corrupted labels. We find that using corrupted labels does not hurt performance on multilingual classification tasks much, which is consistent with previous research on English (Min et al., 2022c). On generation tasks, however, all models perform worse with corrupted labels, but to vastly different extents. XGLM and Llama 2 perform significantly worse with corrupted labels, especially on the machine translation task, whereas **chat models do not rely as much on correct labels**.

This might be explained by ICL helping the model understand the task format and activating prior knowledge acquired by the model, rather than the model learning the task from demonstrations. The observed model insensitivity to correct labels on certain tasks implies that random labels can serve as a strong baseline for demonstration generation before exploring more complex methods (Lyu et al., 2023; Wan et al., 2023). Detailed results for the three types of demonstrations are shown in Table 6.10.

To investigate how these patterns split up across languages, Figure 6.6 shows language-specific results on AfriSenti and XQuAD with Llama 2 and GPT-3.5.[8] On AfriSenti, top-k selection outperforms random selection with Llama 2 across most languages; however, in the case of Swahili and Tsonga, there is a performance drop of 3.2 and 1.2 points, respectively. With GPT-3.5, top-k selection does not help across most languages, but it does help with Mozambican Portuguese and Twi. Similarly, the impact of corrupted labels varies. Llama 2 is affected dramatically by corrupted labels on all languages in XQuAD, whereas GPT-3.5 is much less affected, although to varying degrees across different languages. We urge NLP practitioners to **attend to these discrepancies when creating language-specific applications**, and leave it to future work to explore where they come from.

6.5.3 *Templates vs. Demonstrations*

In-context learning performance depends not only on the demonstrations, which we have varied so far, but also on how they are formatted using templates. Previous work (Gonen et al., 2023; Mizrahi et al., 2024) has shown that modifying the template changes task performance. This section thus seeks to examine the interplay between template choice and demonstrations.

TEMPLATE DESIGN In the zero-shot setting, we observe that chat models tend to generate verbose responses (e.g., "Sure! I can help you with that") or explanations (e.g., "The reason is that ...") that pose a challenge for automatic evaluation. We observe a reduction in this behaviour with ICL, which leads us to question whether demonstrations are merely a means to format model responses. To see if we can achieve the same effect with minor template engineering, we augment the original templates with instructions that focus on output formatting. We call these *formatting-focused templates* which are shown in Table 6.5.

In this section, we focus on XCOPA, AfriSenti, XQuAD, and TyDiQA, as these are the classification and generation tasks that seem to benefit most from in-context demonstrations (see Section 6.5.1). However, as

8 See Appendix A.2 for other models and datasets.

Model	Demonstration	XNLI	IndicXNLI	PAWS-X	XCOPA	XStoryCloze	AfriSenti	XQuAD	TyDiQA	MAFAND (en-xx)	MAFAND (xx-en)
XGLM	Zero-shot	45.87	38.27	**54.79**	57.51	**65.19**	32.71	18.16	26.01	0.79	1.89
	Top-k	**45.99**	**38.85**	51.72	**58.76**	63.99	**44.30**	**27.54**	**34.32**	**9.08**	**15.05**
	Random	$41.40_{0.50}$	$36.36_{0.33}$	$51.48_{0.33}$	$58.73_{0.43}$	$63.02_{0.09}$	$38.68_{0.39}$	$25.77_{0.06}$	$30.11_{0.36}$	$7.77_{0.09}$	$14.39_{0.02}$
	Random-corrupted	$40.94_{0.42}$	$36.41_{0.35}$	$51.04_{0.28}$	$58.21_{0.30}$	$62.40_{0.21}$	$34.90_{0.42}$	$1.21_{0.03}$	$3.47_{0.07}$	$4.59_{0.01}$	$7.66_{0.04}$
Llama 2	Zero-shot	44.25	37.66	59.21	56.02	65.17	32.71	15.33	16.81	10.06	11.27
	Top-k	**47.10**	**40.15**	**59.35**	57.69	**66.16**	**47.25**	**32.37**	**35.36**	**17.29**	**22.92**
	Random	$40.49_{0.35}$	$35.98_{0.24}$	$57.00_{0.29}$	$57.80_{0.32}$	$65.83_{0.08}$	$43.08_{0.02}$	$31.05_{0.28}$	$34.82_{0.21}$	$15.14_{0.01}$	$21.57_{0.02}$
	Random-corrupted	$39.53_{0.20}$	$35.55_{0.44}$	$55.85_{0.92}$	$57.19_{0.06}$	$64.71_{0.25}$	$40.81_{0.36}$	$4.36_{0.25}$	$5.62_{0.27}$	$10.35_{0.04}$	$13.23_{0.04}$
Llama 2-Chat	Zero-shot	36.10	32.32	**64.64**	44.55	57.77	31.18	18.82	20.33	18.83	25.46
	Top-k	47.53	35.73	59.36	**63.55**	**68.82**	**45.75**	**39.94**	**42.38**	**21.50**	**29.02**
	Random	$47.81_{0.85}$	$\mathbf{37.09}_{2.57}$	$61.07_{1.2}$	$63.23_{0.91}$	$68.39_{0.14}$	$43.58_{0.23}$	$38.92_{0.09}$	$39.96_{0.30}$	$20.76_{0.23}$	$28.36_{0.12}$
	Random-corrupted	$\mathbf{48.15}_{1.22}$	$37.05_{3.09}$	$59.59_{1.13}$	$63.20_{0.33}$	$68.62_{0.93}$	$42.81_{0.11}$	$32.98_{0.39}$	$35.59_{0.08}$	$19.63_{0.04}$	$26.82_{0.11}$
GPT-3.5	Zero-shot	63.23	48.23	66.57	73.50	84.55	**53.32**	45.25	48.52	27.39	37.77
	Top-k	**63.27**	**50.45**	69.29	**80.77**	85.23	51.86	67.82	**67.76**	**29.20**	**39.99**
	Random	$63.09_{0.88}$	$49.74_{1.17}$	$\mathbf{71.36}_{0.75}$	$79.91_{0.75}$	$\mathbf{85.84}_{0.30}$	$52.52_{0.21}$	$\mathbf{68.16}_{0.36}$	$64.78_{0.47}$	$28.48_{0.01}$	$39.56_{0.03}$
	Random-corrupted	$62.70_{1.05}$	$48.73_{0.51}$	$70.71_{0.66}$	$79.65_{0.74}$	$85.26_{0.07}$	$53.14_{0.47}$	$62.70_{0.19}$	$59.17_{0.27}$	$27.09_{0.18}$	$39.08_{0.08}$
GPT-4	Zero-shot	70.30	65.41	74.50	88.82	96.05	58.46	44.03	46.97	32.73	45.28
	Top-k	76.53	67.45	**76.14**	91.23	**96.73**	61.68	72.44	**74.65**	**35.06**	**48.34**
	Random	75.77	67.64	76.07	**91.59**	96.68	**62.36**	**73.21**	72.77	33.85	47.69
	Random-corrupted	**76.63**	**67.68**	75.50	90.73	95.55	61.46	63.61	65.80	32.61	47.05

Table 6.10: Performance of different types of demonstrations. For random and random-corrupted, we report the mean and standard deviation across 3 seeds except for GPT-4. Best results for each model and dataset are boldfaced.

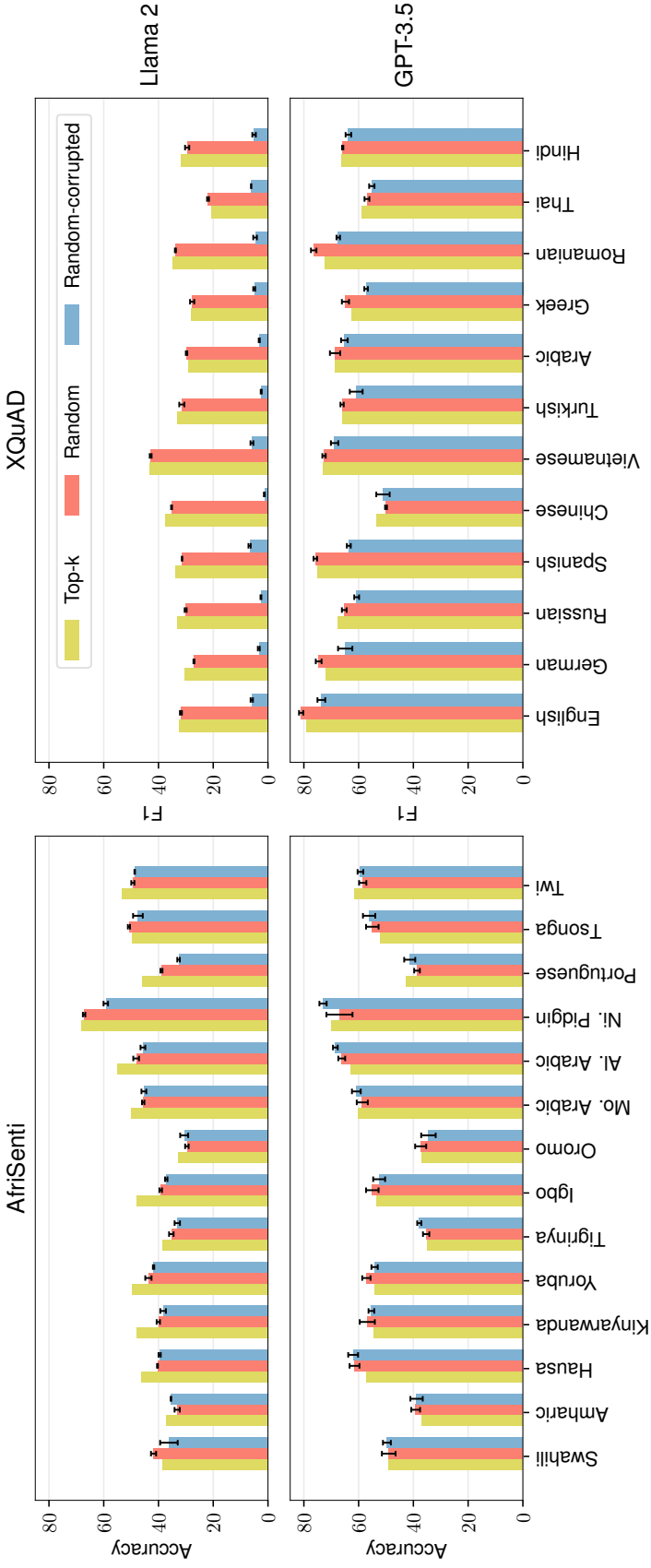


Figure 6.6: Performance of 4-shot ICL using different types of demonstrations for individual languages on AfriSenti and XQuAD. The top row shows Llama 2 results, and the bottom row shows GPT-3.5 results.

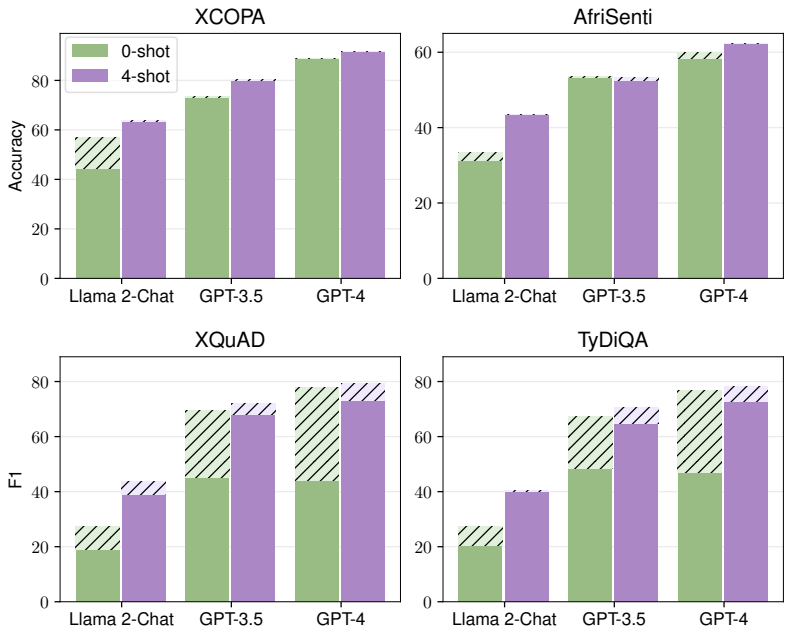


Figure 6.7: Effect of using different templates on 0-shot and 4-shot performance. Formatting-focused templates (with hatching) improve 0-shot performance over original templates (solid colours), and reduce the gap between 0-shot and 4-shot performance. Few-shot results are averaged across 3 seeds except for GPT-4.

Figure 6.7 shows, **the performance gap between zero-shot and in-context learning diminishes with formatting-focused templates**. The gap reduction is more substantial for QA datasets (i.e., the generation tasks) than for XCOPA and AfriSenti (i.e., the classification tasks). We speculate that it is simpler for the model to generate label words for classification tasks with a pre-defined label space than to answer questions in a way that is easy to evaluate automatically. In the latter case, formatting-focused templates can teach output styling, largely eliminating the benefits of demonstrations.

Compared to GPT-3.5 and GPT-4, Llama 2-Chat performs worse in both zero-shot and few-shot settings, and formatting-focused templates have a less pronounced impact. On QA datasets, GPT-3.5 and GPT-4 even achieve better zero-shot performance with formatting-focused templates than ICL with original templates, a pattern that is not observed with Llama 2-Chat. This suggests that **the relative significance of demonstrations and templates varies based on the inherent abilities of models** at solving tasks and following instructions.

With our new formatting-focused templates, we revisit the impact of the input-label mapping discussed in Section 6.5.2. As Table 6.11 shows, all models perform worse with corrupted labels, but formatting-focused templates largely mitigate this degradation. Notably, **GPT-4 using corrupted labels performs on par with ground truth labels**. This strengthens our finding that the correct input-label mapping

Model	Demo. Label	XQuAD		TyDiQA	
		O	F	O	F
Llama 2-Chat	Original	$38.9_{\pm 0.1}$	$43.8_{\pm 0.7}$	$40.0_{\pm 0.3}$	$40.6_{\pm 0.8}$
	Corrupted	$33.0_{\pm 0.4}$	$38.6_{\pm 0.3}$	$35.6_{\pm 0.1}$	$36.3_{\pm 0.5}$
	Δ	5.9	5.2	4.4	4.3
GPT-3.5	Original	$68.2_{\pm 0.4}$	$72.2_{\pm 0.4}$	$64.8_{\pm 0.5}$	$70.5_{\pm 0.5}$
	Corrupted	$62.7_{\pm 0.2}$	$69.9_{\pm 0.2}$	$59.2_{\pm 0.3}$	$67.1_{\pm 0.7}$
	Δ	5.5	2.3	5.6	3.4
GPT-4	Original	73.2	79.3	72.8	78.3
	Corrupted	63.6	79.8	65.8	77.6
	Δ	9.6	-0.5	7.0	0.7

Table 6.11: Effect of using different templates on 4-shot performance with RANDOM and RANDOM-CORRUPTED demonstrations. When using formatting-focused templates (F) over the original templates (O), the performance gap (Δ) between original and corrupted labels decreases. We average and report standard deviations over 3 seeds for all models except GPT-4.

is not that important, while also highlighting the crucial role that templates play in in-context learning.

Figure 6.8 shows the language-specific effects of formatting-focused templates on XQuAD. For Llama 2-Chat, demonstrations remain essential even with a formatting-focused template for most languages, but not Greek and Hindi. GPT-3.5 and GPT-4 also show variance across languages. Moreover, for most languages, zero-shot learning with minor template engineering can match and even exceed in-context learning performance, aligning with previous work on GPT-3 (Reynolds and McDonell, 2021). The fact that we can achieve the same effects through template engineering or demonstrations reinforces our hypothesis that models are not actually learning tasks on the fly. Instead, some combination of demonstrations and templates serves to activate prior knowledge of a task and encourage a consistent output format for automatic evaluation.

We also examine the effect of templates for XCOPA, AfriSenti, and TyDiQA, and show language-specific results in Figure 6.9, Figure 6.10, and Figure 6.11. In a few cases, we found that formatting-focused templates lead to a decline in performance compared to original templates (e.g., Igbo and Mozambican Portuguese in AfriSenti with GPT-3.5). This can be attributed to the model's sensitivity to prompts, highlighting the potential of automatic prompt engineering. Still, formatting-

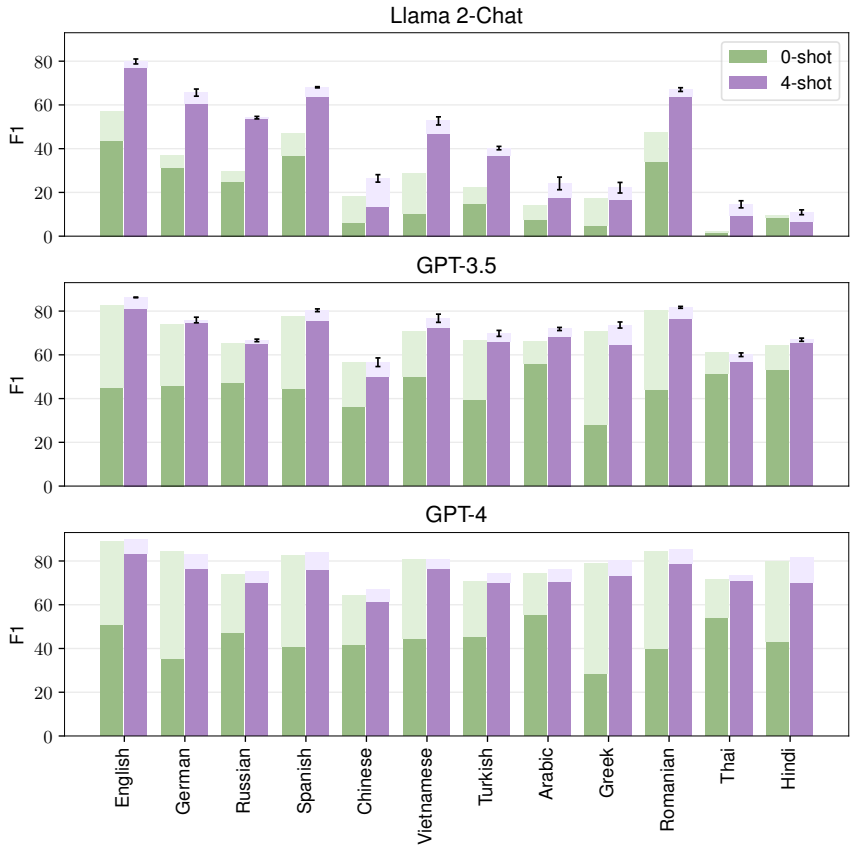


Figure 6.8: Effect of using different templates on 0-shot and 4-shot XQuAD performance. Formatting-focused templates (with hatching) improve 0-shot performance over original templates (solid colours), and reduce the gap between 0-shot and 4-shot performance. Few-shot results are averaged across 3 seeds except for GPT-4.

focused template can largely narrow the performance gap between 0-shot and 4-shot in a broad context.

6.6 DISCUSSION

Our systematic study provides strong evidence that the importance of in-context demonstrations on existing multilingual datasets might be overestimated, as it highly depends on the model, task, and language used. For strong instruction-following models, the effect of demonstrations is *superficial* and can be eliminated with minor template engineering. These findings open up new questions, which we discuss below.

UNDERSTANDING THE FAILURES OF ICL There has been a surge of research interest in understanding the underlying mechanisms of ICL (Xie et al., 2022a; Von Oswald et al., 2023; Wang et al., 2023b; Hendel et al., 2023), motivated by its successes. Our results show that

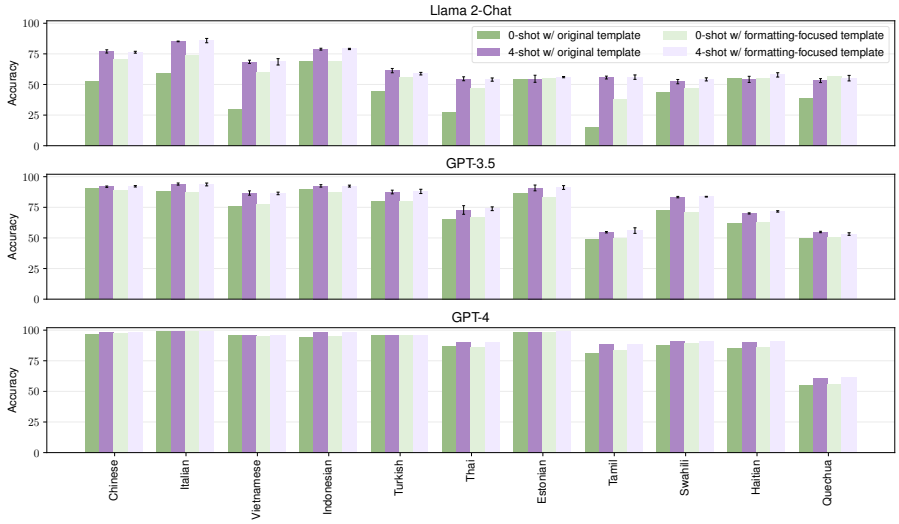


Figure 6.9: Language-specific 0-shot and 4-shot performance for XCOPA with different templates.

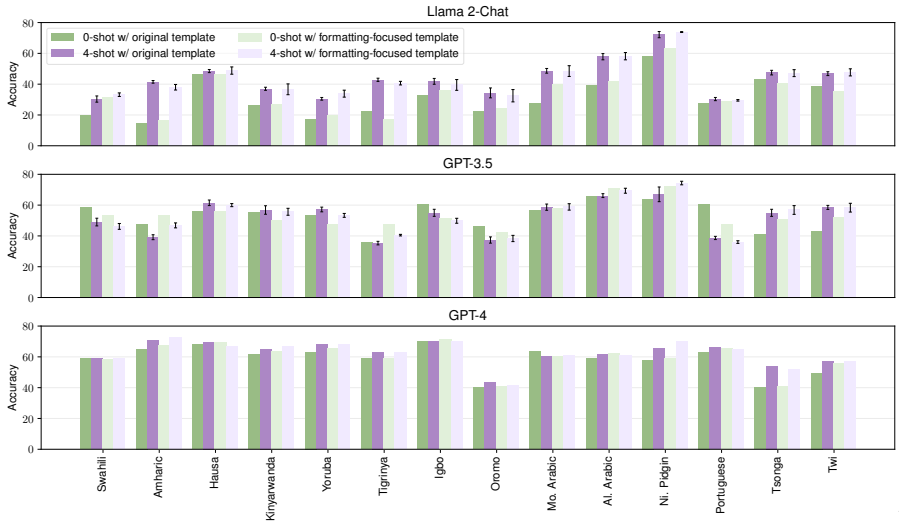


Figure 6.10: Language-specific 0-shot and 4-shot performance for AfriSenti with different templates.

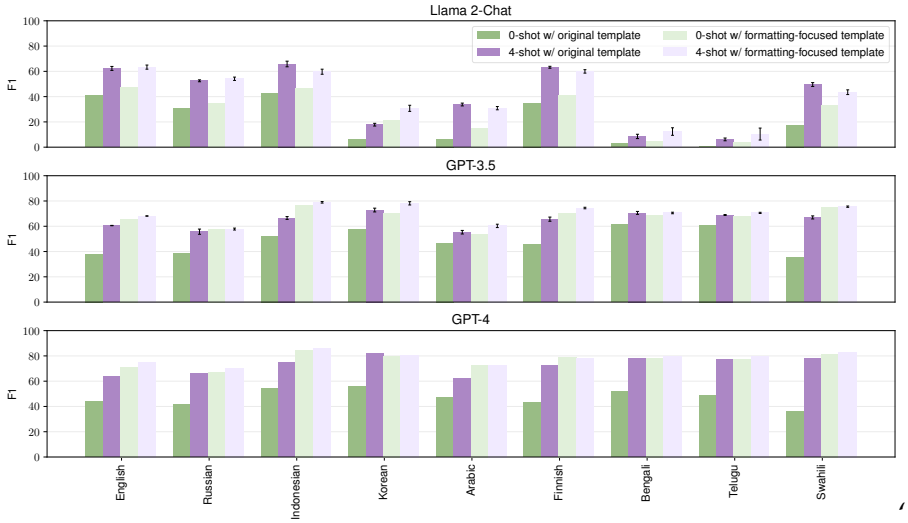


Figure 6.11: Language-specific 0-shot and 4-shot performance for TyDiQA with different templates.

ICL is *not* always effective, and that its performance changes depending on multiple factors including the choice of model, task and language. The failures of ICL need as much scrutiny as its successes for a more fundamental understanding of the learning mechanisms of LLMs.

OPTIMIZING DEMONSTRATIONS OR TEMPLATES With the increasing popularity of research on demonstration selection (Liu et al., 2022b; Rubin et al., 2022; Li et al., 2023a) and prompt engineering (Mishra et al., 2022a; White et al., 2023; Khattab et al., 2023), it is important to understand the interplay of the two. We show that good demonstrations help base models perform better on certain tasks, but that formatting-focused prompting has a much bigger impact on chat models. These results show that the impact of demonstrations cannot be fairly evaluated in isolation from the choice of prompt. These findings have implications both for researchers interested in fairly evaluating ICL, and for practitioners to choose to spend time optimizing demonstrations, templates or both.

EVALUATING MULTILINGUAL ICL Compared to the extensive research on ICL in English (Zhao et al., 2021; Dong et al., 2022; Min et al., 2022b; Mosbach et al., 2023), multilingual ICL remains underexplored. There is no widely accepted setup to robustly evaluate the effectiveness of ICL across languages, since the choice of multilingual models and tasks is limited. Based on our findings, we have some recommendations for the nascent field of multilingual ICL. First, critical evaluation is important. We need to compare ICL strategies to zero-shot learning, and ablate them with multiple templates. Second, as there is so much variance across models, tasks and languages, it is important to carefully scope claims about ICL. Last but not least, every

language is different, so granular per-language analysis is a must in multilingual research.

6.7 LIMITATIONS AND FUTURE WORK

DATA CONTAMINATION Since LLMs are trained with a vast amount of data scraped from the internet, this might result in data contamination, i.e., when the training data includes test datasets. Ahuja et al. (2023) suspect that many multilingual datasets appear in the training data of GPT-4, which might lead to an overestimation of the model's capabilities. In the context of our work, our prompt might just be reminding LLMs of a task they have already seen, whereas on an unseen task, the impact of demonstrations might be different. We do not examine the impact of potential data contamination in our paper and leave an exploration of this to future work.

OTHER DEMONSTRATION CHOICES In this work, we choose to use demonstrations that are in the same languages as the test sample, due to our focus on evaluating inherent multilingual abilities of LLMs, as explained in Section 6.3.2. However, using English demonstrations for cross-lingual transfer or translating test samples into English has its own practical value for NLP applications. Additionally, it is worth exploring selecting demonstrations from a mixture of languages. Expanding our study to more setups would provide additional insights into multilingual and cross-lingual LLM abilities.

OTHER PROMPTING METHODS In Section 6.5.3, we only experiment with manually augmented templates to illustrate how the choice of template can reduce the effectiveness of demonstrations. There is a broad literature on prompt engineering and prompt sensitivity (White et al., 2023; Gonen et al., 2023), suggesting that it is plausible that another prompt could reduce the gap between few-shot and zero-shot performance even further. Chain-of-thought (CoT) prompting is another approach with promising multilingual abilities (Shi et al., 2023b; Huang et al., 2023) that might affect our findings. Our manually-augmented templates are intended only as a starting point for further analysis, which we leave to future work.

BEYOND AUTOMATIC EVALUATION When examining model responses, we noticed some cases where a correct answer as evaluated by a human was not fully captured by automatic evaluation metrics. Human evaluation is time-consuming, expensive, and hard to source for the wide range of languages that we explore in our work. Another option is LLM evaluation, which is becoming increasingly popular (Fu et al., 2023a; Chan et al., 2024), but is also an expensive approach. More importantly, we have no guarantees about LLMs' multilingual

capabilities. As a trade-off between cost and evaluation quality, we stick to automatic evaluation in our work for all tasks and languages.

6.8 CONCLUSION

In this chapter, we conduct an in-depth multidimensional analysis on the impact of demonstrations in multilingual in-context learning. We find that the use of demonstrations does not always provide benefits compared to zero-shot learning, and that there is a large variance in performance across models, datasets and languages. While the quality of demonstrations influences the performance of base LLMs on certain tasks, the impact is significantly reduced for LLMs tuned with alignment techniques. We also examine the interplay between demonstrations and templates, finding that a carefully crafted template can further decrease the benefits of demonstrations. Our granular analysis contributes novel insights with nuance and paves the way for a more thoughtful multilingual ICL evaluation.

7 CONCLUSION AND FUTURE WORK

In this chapter, we summarize the main contributions of this dissertation and discuss perspectives for future research.

7.1 SUMMARY OF CONTRIBUTIONS

This dissertation advances the semantic understanding capabilities of language models by addressing two key challenges: encoding linguistic units into meaningful representations, and modeling semantic relationships in the context of linguistic diversity and multimodal richness. Through a series of studies at both the word and sentence level, it offers methodological innovations and analytical frameworks that clarify and enhance how models acquire and use semantic information. Below, we summarize our main contributions.

1. **Understanding and optimizing word embeddings across languages.** In Chapter 3, we present a systematic analysis of the key factors shaping word embedding quality across diverse languages. Specifically, we examine the effects of the learning algorithm, corpus size, and training parameters. It shows that embedding quality is highly sensitive to these choices, particularly with limited data, with performance differences of up to 40% between the best and worst choice combinations. Correlation analysis indicates that increasing context window size generally improves performance, although the effect varies across languages, while subsampling threshold and minimum word count often have a weaker impact. These findings provide practical guidance for practitioners, emphasizing the factors that most strongly influence word embedding optimization.

2. **Enhancing sentence embedding learning with multimodal signals.** We propose a novel approach in Chapter 4 to incorporate visual and textual information for sentence embedding learning. The training objective not only maximizes agreement between sentence pairs, as in standard textual contrastive learning, but is also extended with a multimodal objective that aligns sentences with corresponding images in a shared embedding space. By augmenting a large text-only corpus with a small amount of multimodal data, our method consistently improves embedding quality. Further analysis of the embedding space shows that visual grounding enhances the alignment of semantically similar sentences while maintaining uniformity of the space, providing

a plausible explanation for the observed gains. Importantly, the multimodal objective is general and can be integrated into other text-only contrastive learning frameworks, offering a simple way to boost their effectiveness.

3. **Modeling semantic textual relatedness for low-resource languages.** In Chapter 5, we introduce a cross-encoder framework for modeling semantic relatedness across 14 African and Asian languages. The framework combines data augmentation and task-adaptive pre-training to improve performance, and employs adapter-based tuning to allow flexible and efficient cross-lingual transfer. We further analyze source language selection for zero-shot cross-lingual transfer, demonstrating that target language performance is highly sensitive to the choice of source language, and that using development set performance as an indicator reliably identifies the optimal source languages. The framework achieves top performance on the SemEval leaderboard and outperforms competing systems, highlighting its effectiveness and practical impact.

4. **In-depth analysis of multilingual in-context learning.** In Chapter 6, we provide a multidimensional analysis of the impact of few-shot demonstrations on ICL, covering 5 models, 9 datasets, and 56 typologically diverse languages. Our study examines three aspects: the number of demonstrations, the quality of demonstrations, and the interplay between demonstrations and templates. We find that demonstration effectiveness varies widely across models, tasks, and languages. Notably, strong instruction-following models are largely insensitive to demonstration quality. We also observe that carefully designed templates can entirely eliminate the benefits of labeled demonstrations. These results suggest that the importance of demonstrations may be overestimated and highlight the need for nuanced evaluation to better understand ICL.

7.2 FUTURE DIRECTIONS

One promising direction is to transform decoder-only LLMs – the dominant architecture of today's most capable models – into strong embedding models. While recent studies have demonstrated that such models can be converted into competitive embedders (BehnamGhader et al., 2024; Springer et al., 2025; Muennighoff et al., 2025), an open challenge is extending this capability beyond English to support a wide range of languages. In theory, these models should be able to produce multilingual embeddings given the multilingual content present in their pre-training corpora. Nevertheless, substantial performance gains on non-English tasks are very likely achievable through improved data

mixes, architectural or training objective modifications that explicitly target multilinguality.

Additionally, we need comprehensive evaluation frameworks that go beyond semantic similarity to cover diverse tasks, languages, and efficiency trade-offs for assessing embedding models. Recent benchmarks such as MTEB (Muennighoff et al., 2023a) provide valuable multi-task evaluation. Expansions like MMTEB (Enevoldsen et al., 2025) significantly broaden language coverage, and regional efforts such as AfriMTEB (Uemura et al., 2025) support under-represented African languages. However, many languages remain uncovered or unevenly distributed across tasks, and efficiency metrics are rarely included. When benchmarks themselves are biased, for example, claiming to be multilingual while largely favoring certain languages over others, assessing the multilingual capabilities of embedding models becomes unreliable. Developing next-generation, well-rounded benchmarks is essential for faithfully evaluating and guiding the development of universal embedding models.

Future work should also draw more attention to the interpretability of representations, as this can directly enable representation-level steering of LLM behavior. Wu et al. (2024) demonstrate that low-rank interventions on hidden representations can efficiently adapt models to new tasks without full re-training, while Arditi et al. (2024) show that controlling a single refusal direction in the embedding space can effectively modulate refusal behavior. These results suggest that much of the task knowledge and alignment signals are already embedded in specific regions of the latent space. Identifying subspaces that encode semantic roles, syntactic structures, factual knowledge, stylistic features, or aspects of persona could substantially facilitate modular and targeted control over model behavior.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

NLP	Natural Language Processing
PLM	Pre-trained Language Model
LLM	Large Language Model
MLM	Masked Language Modeling
CLM	Causal Language Modeling
CBOW	Continuous Bag-of-Words
SG	Skip-Gram
PPMI	Positive Pointwise Mutual Information
SVD	Singular Value Decomposition
OOV	Out-of-Vocabulary
OLS	Ordinary Least Square
STS	Semantic Textual Similarity
STR	Semantic Textual Relatedness
TAPT	Task-Adaptive Pre-Training
PEFT	Parameter-Efficient Fine-Tuning
ICL	In-Context Learning
CoT	Chain-of-Thought
NLI	Natural Language Inference
QA	Question Answering
MT	Machine Translation

BIBLIOGRAPHY

Abdalla, Mohamed, Krishnapriya Vishnubhotla, and Saif Mohammad (May 2023). "What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 782–796. DOI: 10.18653/v1/2023.eacl-main.55. URL: https://aclanthology.org/2023.eacl-main.55.

Adelani, David, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee (2024). "SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects." In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 226–245. URL: https://aclanthology.org/2024.eacl-long.14.

Adelani, David, Miaoran Zhang, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow (Nov. 2021). "Preventing Author Profiling through Zero-Shot Multilingual Back-Translation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 8687–8695. DOI: 10.18653/v1/2021.emnlp-main.684. URL: https://aclanthology.org/2021.emnlp-main.684.

Adelani, David et al. (July 2022). "A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 3053–3070. DOI: 10.18653/v1/2022.naacl-main.223. URL: https://aclanthology.org/2022.naacl-main.223.

Aggarwal, Divyanshu, Vivek Gupta, and Anoop Kunchukuttan (Dec. 2022). "IndicXNLI: Evaluating Multilingual Inference for Indian Languages." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10994–11006. DOI: 10.18653/v1/2022.emnlp-main.755. URL: https://aclanthology.org/2022.emnlp-main.755.

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe (Aug. 2014). "SemEval-2014 Task 10: Multilingual

Semantic Textual Similarity." In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, pp. 81–91. DOI: 10.3115/v1/S14-2010. URL: https://aclanthology.org/S14-2010.

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe (June 2016). "SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation." In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, pp. 497–511. DOI: 10.18653/v1/S16-1081. URL: https://aclanthology.org/S16-1081.

Agirre, Eneko, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre (2012). "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity." In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, pp. 385–393. URL: https://aclanthology.org/S12-1051.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo (June 2013). "*SEM 2013 shared task: Semantic Textual Similarity." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 32–43. URL: https://aclanthology.org/S13-1004.

Agirre, Eneko et al. (June 2015). "SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability." In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pp. 252–263. DOI: 10.18653/v1/S15-2045. URL: https://aclanthology.org/S15-2045.

Agrawal, Sweta, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad (July 2023). "In-context Examples Selection for Machine Translation." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 8857–8873. DOI: 10.18653/v1/2023.findings-acl.564. URL: https://aclanthology.org/2023.findings-acl.564.

Ahuja, Kabir et al. (2023). "MEGA: Multilingual Evaluation of Generative AI." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 4232–4267. URL: https://aclanthology.org/2023.emnlp-main.258.

Alabi, Jesujoba O., David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow (Oct. 2022). "Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4336–4349. URL: https://aclanthology.org/2022.coling-1.382.

Alabi, Jesujoba Oluwadara et al. (Nov. 2025). "AFRIDOC-MT: Document-level MT Corpus for African Languages." In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng. Suzhou, China: Association for Computational Linguistics, pp. 27758–27794. ISBN: 979-8-89176-332-6. DOI: 10.18653/v1/2025.emnlp-main.1413. URL: https://aclanthology.org/2025.emnlp-main.1413/.

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. (2022). "Flamingo: a Visual Language Model for Few-Shot Learning." In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 23716–23736. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Alghanmi, Israa, Luis Espinosa Anke, and Steven Schockaert (Nov. 2020). "Combining BERT with Static Word Embeddings for Categorizing Social Media." In: *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. Online: Association for Computational Linguistics, pp. 28–33. DOI: 10.18653/v1/2020.wnut-1.5. URL: https://aclanthology.org/2020.wnut-1.5.

Almeida, Felipe and Geraldo Xexéo (2023). *Word Embeddings: A Survey*. arXiv: 1901.09069 [cs.CL]. URL: https://arxiv.org/abs/1901.09069.

Amjad, Maaz, Grigori Sidorov, and Alisa Zhila (May 2020). "Data Augmentation using Machine Translation for Fake News Detection in the Urdu Language." English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2537–2542. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.309.

Ansell, Alan, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen (Nov. 2021). "MAD-G: Multilingual Adapter Generation for Efficient Cross-Lingual Transfer." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4762–4781. DOI: 10.18653/v1/2021

.findings-emnlp.410. URL: https://aclanthology.org/2021.findings-emnlp.410.

Ansell, Alan, Edoardo Ponti, Anna Korhonen, and Ivan Vulić (May 2022). "Composable Sparse Fine-Tuning for Cross-Lingual Transfer." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 1778–1796. DOI: 10.18653/v1/2022.acl-long.125. URL: https://aclanthology.org/2022.acl-long.125.

Arditi, Andy, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda (2024). "Refusal in Language Models Is Mediated by a Single Direction." In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. Curran Associates, Inc., pp. 136037–136083. DOI: 10.52202/079017-4322. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma (2017). "A Simple but Tough-to-Beat Baseline for Sentence Embeddings." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SyK00v5xx.

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama (July 2020). "On the Cross-lingual Transferability of Monolingual Representations." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4623–4637. DOI: 10.18653/v1/2020.acl-main.421. URL: https://aclanthology.org/2020.acl-main.421.

Asai, Akari, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi (2024). "BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer." In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 1771–1800. DOI: 10.18653/v1/2024.naacl-long.100. URL: https://aclanthology.org/2024.naacl-long.100.

Bach, Stephen et al. (May 2022). "PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Dublin, Ireland: Association for Computational Linguistics, pp. 93–104. DOI: 10.18653/v1/2022.acl-demo.9. URL: https://aclanthology.org/2022.acl-demo.9.

Bakarov, Amir (2018). *A Survey of Word Embeddings Evaluation Methods*. arXiv: 1801.09536 [cs.CL]. URL: https://arxiv.org/abs/1801.09536.

Bansal, Hritik, Karthik Gopalakrishnan, Saket Dingliwal, Sravan Bodapati, Katrin Kirchhoff, and Dan Roth (July 2023). "Rethinking the Role of Scale for In-Context Learning: An Interpretability-based Case Study at 66 Billion Scale." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 11833–11856. DOI: 10.18653/v1/2023.acl-long.660. URL: https://aclanthology.org/2023.acl-long.660.

Barzegar, Siamak, Brian Davis, Manel Zarrouk, Siegfried Handschuh, and Andre Freitas (May 2018). "SemR-11: A Multi-Lingual Gold-Standard for Semantic Similarity and Relatedness for Eleven Languages." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1618.

BehnamGhader, Parishad, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy (2024). "LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders." In: *First Conference on Language Modeling*. URL: https://openreview.net/forum?id=IW1PR7vEBf.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the dangers of stochastic parrots: Can language models be too big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463.

Bisk, Yonatan et al. (Nov. 2020). "Experience Grounds Language." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8718–8735. DOI: 10.18653/v1/2020.emnlp-main.703. URL: https://aclanthology.org/2020.emnlp-main.703.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information." In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010.

Bommasani, Rishi, Kelly Davis, and Claire Cardie (July 2020). "Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings." In: *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 4758–4781. DOI: 10.18653/v1/2020.acl-main.431. URL: https://aclanthology.org/2020.acl-main.431.

Bordes, Patrick, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari (Nov. 2019). "Incorporating Visual Semantics into Sentence Representations within a Grounded Space." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 696–707. DOI: 10.18653/v1/D19-1064. URL: https://aclanthology.org/D19-1064.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (Sept. 2015). "A large annotated corpus for learning natural language inference." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://aclanthology.org/D15-1075.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language models are few-shot learners." In: *Advances in neural information processing systems* 33, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran (July 2012). "Distributional Semantics in Technicolor." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 136–145. URL: https://aclanthology.org/P12-1015.

Budanitsky, Alexander (1999). *Lexical semantic relatedness and its application in natural language processing*. Tech. rep. technical report CSRG-390, Department of Computer Science, University of Toronto. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c2d95e890ee904f70701fa27326d31980424d5dd.

Bullinaria, John A and Joseph P Levy (2007). "Extracting semantic representations from word co-occurrence statistics: A computational study." In: *Behavior research methods* 39, pp. 510–526.

Carlsson, Fredrik, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren (2020). "Semantic re-tuning with contrastive tension." In: *International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=Ov_sMNau-PF.

Caselles-Dupré, Hugo, Florian Lesaint, and Jimena Royo-Letelier (2018). "Word2vec applied to recommendation: Hyperparameters

matter." In: *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 352–356.

Castro, Santiago, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria (July 2019). "Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4619–4629. DOI: 10.18653/v1/P19-1455. URL: https://aclanthology.org/P19-1455.

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia (Aug. 2017). "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation." In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1–14. DOI: 10.18653/v1/S17-2001. URL: https://aclanthology.org/S17-2001.

Cer, Daniel et al. (Nov. 2018b). "Universal Sentence Encoder for English." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 169–174. DOI: 10.18653/v1/D18-2029. URL: https://aclanthology.org/D18-2029.

Cer, Daniel et al. (2018a). *Universal Sentence Encoder*. arXiv: 1803.11175 [cs.CL]. URL: https://arxiv.org/abs/1803.11175.

Chan, Chi-Min, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu (2024). "ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate." In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=FQepisCUWu.

Chang, Ting-Yun and Robin Jia (July 2023). "Data Curation Alone Can Stabilize In-context Learning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 8123–8144. DOI: 10.18653/v1/2023.acl-long.452. URL: https://aclanthology.org/2023.acl-long.452.

Chang, Tyler, Zhuowen Tu, and Benjamin Bergen (Dec. 2022). "The Geometry of Multilingual Language Model Representations." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 119–136. DOI: 10.18653/v1/2022.emnlp-main.9. URL: https://aclanthology.org/2022.emnlp-main.9.

Chen, Guanzheng, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang (Dec. 2022a). "Revisiting Parameter-Efficient Tuning: Are We Really There Yet?" In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates:

Association for Computational Linguistics, pp. 2612–2626. DOI: 10.18653/v1/2022.emnlp-main.168. URL: https://aclanthology.org/2022.emnlp-main.168.

Chen, Jiuhai, Lichang Chen, Chen Zhu, and Tianyi Zhou (Dec. 2023). "How Many Demonstrations Do You Need for In-context Learning?" In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 11149–11159. DOI: 10.18653/v1/2023.findings-emnlp.745. URL: https://aclanthology.org/2023.findings-emnlp.745.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020a). "A simple framework for contrastive learning of visual representations." In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607. URL: http://proceedings.mlr.press/v119/chen20j.html.

Chen, Xinghao et al. (July 2025). "Unveiling the Key Factors for Distilling Chain-of-Thought Reasoning." In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 15094–15119. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.782. URL: https://aclanthology.org/2025.findings-acl.782/.

Chen, Yanda, Ruiqi Zhong, Sheng Zha, George Karypis, and He He (May 2022b). "Meta-learning via Language Model In-context Tuning." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 719–730. DOI: 10.18653/v1/2022.acl-long.53. URL: https://aclanthology.org/2022.acl-long.53.

Chen, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu (2020b). "Uniter: Universal image-text representation learning." In: *European conference on computer vision*. Springer, pp. 104–120.

Chingacham, Anupama, Miaoran Zhang, Vera Demberg, and Dietrich Klakow (Aug. 2024). "Human Speech Perception in Noise: Can Large Language Models Paraphrase to Improve It?" In: *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*. Ed. by Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H. Andrew Schwartz. TBD: ACL, pp. 1–15. DOI: 10.18653/v1/2024.hucllm-1.1. URL: https://aclanthology.org/2024.hucllm-1.1/.

Choenni, Rochelle, Dan Garrette, and Ekaterina Shutova (Dec. 2023). "How do languages influence each other? Studying cross-lingual data sharing during LM fine-tuning." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association

for Computational Linguistics, pp. 13244–13257. DOI: 10.18653/v1/2023.emnlp-main.818. URL: https://aclanthology.org/2023.emnlp-main.818/.

Chuang, Yung-Sung, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass (July 2022). "DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 4207–4218. DOI: 10.18653/v1/2022.naacl-main.311. URL: https://aclanthology.org/2022.naacl-main.311.

Clark, Jonathan H., Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki (2020). "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages." In: *Transactions of the Association for Computational Linguistics* 8, pp. 454–470. DOI: 10.1162/tacl_a_00317. URL: https://aclanthology.org/2020.tacl-1.30.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes (Sept. 2017). "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 670–680. DOI: 10.18653/v1/D17-1070. URL: https://aclanthology.org/D17-1070.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov (2018). "XNLI: Evaluating Cross-lingual Sentence Representations." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485. DOI: 10.18653/v1/D18-1269. URL: https://aclanthology.org/D18-1269.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational

Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dodge, Jesse, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith (2020). *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping*. arXiv: 2002.06305 [cs.CL]. URL: https://arxiv.org/abs/2002.06305.

Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui (2022). "A survey for in-context learning." In: *arXiv*. URL: https://arxiv.org/abs/2301.00234.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (2025). *Ethnologue: Languages of the World*. Ed. by David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 28th. Dallas, Texas: SIL International. URL: http://www.ethnologue.com.

Enevoldsen, Kenneth et al. (2025). "MMTEB: Massive Multilingual Text Embedding Benchmark." In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=zl3pfz4VCV.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang (May 2022). "Language-agnostic BERT Sentence Embedding." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 878–891. DOI: 10.18653/v1/2022.acl-long.62. URL: https://aclanthology.org/2022.acl-long.62.

Feng, Steven Y., Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy (Aug. 2021). "A Survey of Data Augmentation Approaches for NLP." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 968–988. DOI: 10.18653/v1/2021.findings-acl.84. URL: https://aclanthology.org/2021.findings-acl.84.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin (2002). "Placing Search in Context: The Concept Revisited." In: *ACM Transactions on Information Systems* 20.1, pp. 116–131.

Fu, Jinlan, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu (2023a). "GPTScore: Evaluate as You Desire." In: *arXiv*. arXiv: 2302.04166 [cs.CL]. URL: https://arxiv.org/abs/2302.04166.

Fu, Yao, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot (2023b). "Complexity-Based Prompting for Multi-step Reasoning." In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=yf1icZHC-l9.

Gao, Luyu, Zhuyun Dai, and Jamie Callan (2021a). "Rethink Training of BERT Rerankers in Multi-stage Retrieval Pipeline." In: *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR*

2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II. Berlin, Heidelberg: Springer-Verlag, 280–286. ISBN: 978-3-030-72239-5. DOI: 10.1007/978-3-030-72240-1_26. URL: https://doi.org/10.1007/978-3-030-72240-1_26.

Gao, Tianyu, Xingcheng Yao, and Danqi Chen (Nov. 2021b). "SimCSE: Simple Contrastive Learning of Sentence Embeddings." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552. URL: https://aclanthology.org/2021.emnlp-main.552.

Gautam, Vagrant, Miaoran Zhang, and Dietrich Klakow (Dec. 2023). "A Lightweight Method to Generate Unanswerable Questions in English." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 7349–7360. DOI: 10.18653/v1/2023.findings-emnlp.491. URL: https://aclanthology.org/2023.findings-emnlp.491/.

Giorgi, John, Osvald Nitski, Bo Wang, and Gary Bader (Aug. 2021). "DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 879–895. DOI: 10.18653/v1/2021.acl-long.72. URL: https://aclanthology.org/2021.acl-long.72.

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff (May 2012). "Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 759–765. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

Gonen, Hila, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer (Dec. 2023). "Demystifying Prompts in Language Models via Perplexity Estimation." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 10136–10148. DOI: 10.18653/v1/2023.findings-emnlp.679. URL: https://aclanthology.org/2023.findings-emnlp.679.

Gordon, Andrew, Zornitsa Kozareva, and Melissa Roemmele (2012). "SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning." In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*

(SemEval 2012). Montréal, Canada: Association for Computational Linguistics, pp. 394–398. URL: https://aclanthology.org/S12-1052.

Gupta, Prakhar and Martin Jaggi (Aug. 2021). "Obtaining Better Static Word Embeddings Using Contextual Embedding Models." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 5241–5253. DOI: 10.18653/v1/2021.acl-long.408. URL: https://aclanthology.org/2021.acl-long.408.

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith (July 2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8342–8360. DOI: 10.18653/v1/2020.acl-main.740. URL: https://aclanthology.org/2020.acl-main.740.

Haddow, Barry, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch (Sept. 2022). "Survey of Low-Resource Machine Translation." In: *Computational Linguistics* 48.3, pp. 673–732. DOI: 10.1162/coli_a_00446. URL: https://aclanthology.org/2022.cl-3.6.

Harris, Zellig S (1954). "Distributional structure." In: *Word* 10.2-3, pp. 146–162. URL: https://doi.org/10.1080/00437956.1954.11659520.

He, Junxian, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig (2022). "Towards a Unified View of Parameter-Efficient Transfer Learning." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=0RDcd5Axok.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

Hedderich, Michael A., Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow (June 2021). "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 2545–2568. DOI: 10.18653/v1/2021.naacl-main.201. URL: https://aclanthology.org/2021.naacl-main.201.

Hellrich, Johannes, Bernd Kampe, and Udo Hahn (June 2019). "The Influence of Down-Sampling Strategies on SVD Word Embedding Stability." In: *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Minneapolis, USA: Association for Computational Linguistics, pp. 18–26. DOI: 10.18653/v1/W19-2003. URL: https://aclanthology.org/W19-2003.

Hendel, Roee, Mor Geva, and Amir Globerson (2023). "In-Context Learning Creates Task Vectors." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 9318–9333. DOI: 10.18653/v1/2023.findings-emnlp.624. URL: https://aclanthology.org/2023.findings-emnlp.624.

Hill, Felix, Kyunghyun Cho, and Anna Korhonen (June 2016). "Learning Distributed Representations of Sentences from Unlabelled Data." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 1367–1377. DOI: 10.18653/v1/N16-1162. URL: https://aclanthology.org/N16-1162.

Hill, Felix, Roi Reichart, and Anna Korhonen (Dec. 2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation." In: *Computational Linguistics* 41.4, pp. 665–695. DOI: 10.1162/COLI_a_00237. URL: https://aclanthology.org/J15-4004.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019). "Parameter-Efficient Transfer Learning for NLP." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2790–2799. URL: http://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf.

Hu, Edward J, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2022). "LoRA: Low-Rank Adaptation of Large Language Models." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson (2020). "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4411–4421. URL: https://proceedings.mlr.press/v119/hu20b.html.

Huang, Haoyang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei (Dec. 2023). "Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-

Thought Prompting." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 12365–12394. DOI: 10.18653/v1/2023.findings-emnlp.826. URL: https://aclanthology.org/2023.findings-emnlp.826.

Humeau, Samuel, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston (2020). "Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkxgnnNFvH.

Jiang, Chao, Hsiang-Fu Yu, Cho-Jui Hsieh, and Kai-Wei Chang (June 2018). "Learning Word Embeddings for Low-Resource Languages by PU Learning." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1024–1034. DOI: 10.18653/v1/N18-1093. URL: https://aclanthology.org/N18-1093.

Jiang, Ting, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang (Dec. 2022). "PromptBERT: Improving BERT Sentence Embeddings with Prompts." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 8826–8837. DOI: 10.18653/v1/2022.emnlp-main.603. URL: https://aclanthology.org/2022.emnlp-main.603.

Johnson, Melvin et al. (2017). "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." In: *Transactions of the Association for Computational Linguistics* 5, pp. 339–351. DOI: 10.1162/tacl_a_00065. URL: https://aclanthology.org/Q17-1024.

Jørgensen, Tollef (June 2024). "PEAR at SemEval-2024 Task 1: Pair Encoding with Augmented Re-sampling for Semantic Textual Relatedness." In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, pp. 1405–1411. DOI: 10.18653/v1/2024.semeval-1.202. URL: https://aclanthology.org/2024.semeval-1.202/.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (July 2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. URL: https://aclanthology.org/2020.acl-main.560.

Jungmaier, Jakob, Nora Kassner, and Benjamin Roth (May 2020). "Dirichlet-Smoothed Word Embeddings for Low-Resource Settings." English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3560–3565. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.437.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). *Scaling Laws for Neural Language Models*. arXiv: 2001.08361 [cs.LG]. URL: https://arxiv.org/abs/2001.08361.

Karpinska, Marzena, Bofang Li, Anna Rogers, and Aleksandr Drozd (July 2018). "Subcharacter Information in Japanese Embeddings: When Is It Worth It?" In: *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*. Melbourne, Australia: Association for Computational Linguistics, pp. 28–37. DOI: 10.18653/v1/W18-2905. URL: https://aclanthology.org/W18-2905.

Khattab, Omar et al. (2023). "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines." In: *arXiv*. URL: https://arxiv.org/abs/2310.03714.

Kiela, Douwe, Alexis Conneau, Allan Jabri, and Maximilian Nickel (June 2018). "Learning Visually Grounded Sentence Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 408–418. DOI: 10.18653/v1/N18-1038. URL: https://aclanthology.org/N18-1038.

Kiet, Nguyen Tuan and Dang Van Thin (June 2024). "NRK at SemEval-2024 Task 1: Semantic Textual Relatedness through Domain Adaptation and Ensemble Learning on BERT-based models." In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, pp. 76–81. DOI: 10.18653/v1/2024.semeval-1.13. URL: https://aclanthology.org/2024.semeval-1.13/.

Kim, Hyuhng Joon, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang goo Lee (2022). *Self-Generated In-Context Learning: Leveraging Auto-regressive Language Models as a Demonstration Generator*. arXiv: 2206.08082 [cs.CL]. URL: https://arxiv.org/abs/2206.08082.

Kim, Taeuk, Kang Min Yoo, and Sang-goo Lee (Aug. 2021). "Self-Guided Contrastive Learning for BERT Sentence Representations." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, pp. 2528–2540. DOI: 10.18653/v1/2021.acl-long.197. URL: https://aclanthology.org/2021.acl-long.197.

Kiros, Ryan, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Skip-Thought Vectors." In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/f442d33fa06832082290ad8544a8da27-Paper.pdf.

Kojima, Takeshi, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). "Large Language Models are Zero-Shot Reasoners." In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Kudo, Taku and John Richardson (Nov. 2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: https://aclanthology.org/D18-2012.

Lai, Siwei, Kang Liu, Shizhu He, and Jun Zhao (2016). "How to generate a good word embedding." In: *IEEE Intelligent Systems* 31.6, pp. 5–14.

Lai, Viet, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen (2023). "ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 13171–13189. DOI: 10.18653/v1/2023.findings-emnlp.878. URL: https://aclanthology.org/2023.findings-emnlp.878.

Lange, Lukas, Jannik Strötgen, Heike Adel, and Dietrich Klakow (Nov. 2021). "To Share or not to Share: Predicting Sets of Sources for Model Transfer Learning." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 8744–8753. DOI: 10.18653/v1/2021.emnlp-main.689. URL: https://aclanthology.org/2021.emnlp-main.689.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš (Nov. 2020). "From Zero to Hero: On the Limitations of Zero-Shot

Language Transfer with Multilingual Transformers." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4483–4499. DOI: 10.18653/v1/2020.emnlp-main.363. URL: https://aclanthology.org/2020.emnlp-main.363.

Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni (2015). "Combining Language and Vision with a Multimodal Skip-gram Model." In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, pp. 153–163. DOI: 10.3115/v1/N15-1016. URL: https://aclanthology.org/N15-1016.

Lee, Chankyu, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping (2025). "NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models." In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=lgsyLSsDRe.

Lei, Yibin, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates (Aug. 2024). "Meta-Task Prompting Elicits Embeddings from Large Language Models." In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 10141–10157. DOI: 10.18653/v1/2024.acl-long.546. URL: https://aclanthology.org/2024.acl-long.546/.

Lester, Brian, Rami Al-Rfou, and Noah Constant (Nov. 2021). "The Power of Scale for Parameter-Efficient Prompt Tuning." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3045–3059. DOI: 10.18653/v1/2021.emnlp-main.243. URL: https://aclanthology.org/2021.emnlp-main.243.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). "Improving Distributional Similarity with Lessons Learned from Word Embeddings." In: *Transactions of the Association for Computational Linguistics* 3, pp. 211–225. DOI: 10.1162/tacl_a_00134. URL: https://aclanthology.org/Q15-1016.

Li, Bohan, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li (Nov. 2020). "On the Sentence Embeddings from Pre-trained Language Models." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9119–9130. DOI: 10.18653/v1/2020.emnlp-main.733. URL: https://aclanthology.org/2020.emnlp-main.733.

Li, Jianjian, Shengwei Liang, Yong Liao, Hongping Deng, and Haiyang Yu (June 2024). "USTCCTSU at SemEval-2024 Task 1: Reducing

Anisotropy for Cross-lingual Semantic Textual Relatedness Task." In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, pp. 881–887. DOI: 10.18653/v1/2024.semeval-1.126. URL: https://aclanthology.org/2024.semeval-1.126/.

Li, Songtao and Hao Tang (2025). *Multimodal Alignment and Fusion: A Survey*. arXiv: 2411.17040 [cs.CV]. URL: https://arxiv.org/abs/2411.17040.

Li, Xiang Lisa and Percy Liang (Aug. 2021). "Prefix-Tuning: Optimizing Continuous Prompts for Generation." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353. URL: https://aclanthology.org/2021.acl-long.353.

Li, Xiaonan, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu (July 2023a). "Unified Demonstration Retriever for In-Context Learning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 4644–4668. DOI: 10.18653/v1/2023.acl-long.256. URL: https://aclanthology.org/2023.acl-long.256.

Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang (2023b). *Towards General Text Embeddings with Multi-stage Contrastive Learning*. arXiv: 2308.03281 [cs.CL]. URL: https://arxiv.org/abs/2308.03281.

Liang, Wang, Yang Nan, Huang Xiaolong, Yang Linjun, Majumder Rangan, and Wei Furu (2024). *Multilingual E5 Text Embeddings: A Technical Report*. arXiv: 2402.05672 [cs.CL]. URL: https://arxiv.org/abs/2402.05672.

Libovický, Jindřich, Rudolf Rosa, and Alexander Fraser (Nov. 2020). "On the Language Neutrality of Pre-trained Multilingual Representations." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 1663–1674. DOI: 10.18653/v1/2020.findings-emnlp.150. URL: https://aclanthology.org/2020.findings-emnlp.150.

Lin, Pin-Jie, Miaoran Zhang, Marius Mosbach, and Dietrich Klakow (Aug. 2024). "Exploring the Effectiveness and Consistency of Task Selection in Intermediate-Task Transfer Learning." In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Ed. by Xiyan Fu and Eve Fleisig. Bangkok, Thailand: Association for Computational Linguistics, pp. 170–185. ISBN: 979-8-89176-097-4. DOI: 10.18653/v1/2024.acl-srw.24. URL: https://aclanthology.org/2024.acl-srw.24/.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). "Microsoft COCO: Common Objects in Context." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, pp. 740–755. URL: https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Lin, Xi Victoria et al. (Dec. 2022). "Few-shot Learning with Multilingual Generative Language Models." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9019–9052. DOI: 10.18653/v1/2022.emnlp-main.616. URL: https://aclanthology.org/2022.emnlp-main.616.

Lin, Yu-Hsiang et al. (July 2019). "Choosing Transfer Languages for Cross-Lingual Learning." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3125–3135. DOI: 10.18653/v1/P19-1301. URL: https://aclanthology.org/P19-1301.

Lison, Pierre and Andrey Kutuzov (May 2017). "Redefining Context Windows for Word Embedding Models: An Experimental Study." In: *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 284–288. URL: https://aclanthology.org/W17-0239.

Littell, Patrick, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin (Apr. 2017). "URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 8–14. URL: https://aclanthology.org/E17-2002.

Liu, Fangyu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov (2022a). "Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=AmUhwTOHgm.

Liu, Fangyu, Ivan Vulić, Anna Korhonen, and Nigel Collier (Nov. 2021). "Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1442–1459. DOI: 10.18653/v1/2021.emnlp-main.109. URL: https://aclanthology.org/2021.emnlp-main.109.

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee (2023). "Visual Instruction Tuning." In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko,

M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., pp. 34892–34916. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Liu, Jiachang, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen (May 2022b). "What Makes Good In-Context Examples for GPT-3?" In: *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Dublin, Ireland and Online: Association for Computational Linguistics, pp. 100–114. DOI: 10.18653/v1/2022.deelio-1.10. URL: https://aclanthology.org/2022.deelio-1.10.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A robustly optimized BERT pretraining approach." In: *arXiv preprint arXiv:1907.11692*. URL: https://arxiv.org/abs/1907.11692.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp (May 2022). "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556. URL: https://aclanthology.org/2022.acl-long.556.

Luong, Thang, Richard Socher, and Christopher Manning (Aug. 2013). "Better Word Representations with Recursive Neural Networks for Morphology." In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 104–113. URL: https://aclanthology.org/W13-3512.

Lyu, Xinxi, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi (July 2023). "Z-ICL: Zero-Shot In-Context Learning with Pseudo-Demonstrations." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 2304–2317. DOI: 10.18653/v1/2023.acl-long.129. URL: https://aclanthology.org/2023.acl-long.129.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli (May 2014). "A SICK cure for the evaluation of compositional distributional semantic models." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 216–223. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin (May 2018). "Advances in Pre-Training Distributed Word Representations." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1008.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013b). "Distributed Representations of Words and Phrases and their Compositionality." In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Miller, George A and Walter G Charles (1991). "Contextual correlates of semantic similarity." In: *Language and cognitive processes* 6.1, pp. 1–28.

Min, Sewon, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer (May 2022a). "Noisy Channel Language Model Prompting for Few-Shot Text Classification." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5316–5330. DOI: 10.18653/v1/2022.acl-long.365. URL: https://aclanthology.org/2022.acl-long.365.

Min, Sewon, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi (July 2022b). "MetaICL: Learning to Learn In Context." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 2791–2809. DOI: 10.18653/v1/2022.naacl-main.201. URL: https://aclanthology.org/2022.naacl-main.201.

Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer (Dec. 2022c). "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?" In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 11048–11064. DOI: 10.18653/v1/2022.emnlp-main.759. URL: https://aclanthology.org/2022.emnlp-main.759.

Mishra, Swaroop, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi (May 2022a). "Reframing Instructional Prompts to GPTk's Language." In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computa-

tional Linguistics, pp. 589–612. DOI: 10.18653/v1/2022.findings-acl.50. URL: https://aclanthology.org/2022.findings-acl.50.

Mishra, Swaroop, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi (May 2022b). "Cross-Task Generalization via Natural Language Crowdsourcing Instructions." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3470–3487. DOI: 10.18653/v1/2022.acl-long.244. URL: https://aclanthology.org/2022.acl-long.244.

Mizrahi, Moran, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky (2024). "State of What Art? A Call for Multi-Prompt LLM Evaluation." In: *Transactions of the Association for Computational Linguistics* 12, pp. 933–949. DOI: 10.1162/tacl_a_00681. URL: https://aclanthology.org/2024.tacl-1.52/.

Mohammad, Saif M and Graeme Hirst (2012). *Distributional Measures as Proxies for Semantic Relatedness*. arXiv: 1203.1889 [cs.CL]. URL: https://arxiv.org/abs/1203.1889.

Mosbach, Marius, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar (July 2023). "Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 12284–12314. DOI: 10.18653/v1/2023.findings-acl.779. URL: https://aclanthology.org/2023.findings-acl.779.

Mostafazadeh, Nasrin, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen (Aug. 2016). "Story Cloze Evaluator: Vector Space Representation Evaluation by Predicting What Happens Next." In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 24–29. DOI: 10.18653/v1/W16-2505. URL: https://aclanthology.org/W16-2505.

Muennighoff, Niklas, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela (2025). "Generative Representational Instruction Tuning." In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BC4lIvfSzv.

Muennighoff, Niklas, Nouamane Tazi, Loic Magne, and Nils Reimers (May 2023a). "MTEB: Massive Text Embedding Benchmark." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2014–2037. DOI: 10.18653/v1/2023.eacl-main.148. URL: https://aclanthology.org/2023.eacl-main.148.

Muennighoff, Niklas et al. (July 2023b). "Crosslingual Generalization through Multitask Finetuning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers). Toronto, Canada: Association for Computational Linguistics, pp. 15991–16111. DOI: 10.18653/v1/2023.acl-long.891. URL: https://aclanthology.org/2023.acl-long.891.

Muhammad, Shamsuddeen et al. (Dec. 2023). "AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 13968–13981. DOI: 10.18653/v1/2023.emnlp-main.862. URL: https://aclanthology.org/2023.emnlp-main.862.

Ogueji, Kelechi, Yuxin Zhu, and Jimmy Lin (Nov. 2021). "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages." In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 116–126. DOI: 10.18653/v1/2021.mrl-1.11. URL: https://aclanthology.org/2021.mrl-1.11.

Ojo, Jessica, Kelechi Ogueji, Pontus Stenetorp, and David I. Adelani (2023). "How good are Large Language Models on African Languages?" In: *arXiv*. arXiv: 2311.07978 [cs.CL]. URL: https://arxiv.org/abs/2311.07978.

Ojo, Jessica, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani (July 2025). "AfroBench: How Good are Large Language Models on African Languages?" In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 19048–19095. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.976. URL: https://aclanthology.org/2025.findings-acl.976/.

OpenAI et al. (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

Opitz, Juri and Anette Frank (Nov. 2022). "SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features." In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, pp. 625–638. URL: https://aclanthology.org/2022.aacl-main.48.

Ousidhoum, Nedjma et al. (2024a). "SemEval Task 1: Semantic Textual Relatedness for African and Asian Languages." In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, pp. 1963–

1978. DOI: 10.18653/v1/2024.semeval-1.272. URL: https://aclanthology.org/2024.semeval-1.272.

Ousidhoum, Nedjma et al. (2024b). "SemRel2024: A Collection of Semantic Textual Relatedness Datasets for 13 Languages." In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 2512–2530. DOI: 10.18653/v1/2024.findings-acl.147. URL: https://aclanthology.org/2024.findings-acl.147.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). "Training language models to follow instructions with human feedback." In: *Advances in Neural Information Processing Systems* 35, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi (June 2018). "Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 528–540. DOI: 10.18653/v1/N18-1049. URL: https://aclanthology.org/N18-1049.

Pan, Jane, Tianyu Gao, Howard Chen, and Danqi Chen (July 2023). "What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 8298–8319. DOI: 10.18653/v1/2023.findings-acl.527. URL: https://aclanthology.org/2023.findings-acl.527.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation." In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

Parović, Marinela, Goran Glavaš, Ivan Vulić, and Anna Korhonen (July 2022). "BAD-X: Bilingual Adapters Improve Zero-Shot Cross-Lingual Transfer." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 1791–1799. DOI: 10.18653/v1/2022.naacl-main.130. URL: https://aclanthology.org/2022.naacl-main.130.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

Pfeiffer, Jonas, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe (July 2022). "Lifting the Curse of Multilinguality by Pre-training Modular Transformers." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 3479–3495. DOI: 10.18653/v1/2022.naacl-main.255. URL: https://aclanthology.org/2022.naacl-main.255.

Pfeiffer, Jonas, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder (Nov. 2020). "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7654–7673. DOI: 10.18653/v1/2020.emnlp-main.617. URL: https://aclanthology.org/2020.emnlp-main.617.

Philippy, Fred, Siwen Guo, and Shohreh Haddadan (July 2023). "Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 5877–5891. DOI: 10.18653/v1/2023.acl-long.323. URL: https://aclanthology.org/2023.acl-long.323.

Pires, Telmo, Eva Schlinger, and Dan Garrette (July 2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

Ponti, Edoardo Maria, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen (Nov. 2020). "XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Pro-*

cessing (EMNLP). Online: Association for Computational Linguistics, pp. 2362–2376. DOI: 10.18653/v1/2020.emnlp-main.185. URL: https://aclanthology.org/2020.emnlp-main.185.

Popović, Maja (Sept. 2015). "chrF: character n-gram F-score for automatic MT evaluation." In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049. URL: https://aclanthology.org/W15-3049.

– (Sept. 2017). "chrF++: words helping character n-grams." In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 612–618. DOI: 10.18653/v1/W17-4770. URL: https://aclanthology.org/W17-4770.

Pruksachatkun, Yada, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman (July 2020). "Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5231–5247. DOI: 10.18653/v1/2020.acl-main.467. URL: https://aclanthology.org/2020.acl-main.467.

Qin, Libo, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu (2024). *Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers*. arXiv: 2404.04925 [cs.CL]. URL: https://arxiv.org/abs/2404.04925.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners." In: *OpenAI blog* 1.8, p. 9.

Radford, Alec et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision." In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Ed. by Marina Meila and Tong Zhang, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *Journal of machine learning research* 21.140, pp. 1–67.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (Nov. 2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://aclanthology.org/D16-1264.

Ramesh, Gowtham et al. (Feb. 2022). "Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages."

In: *Transactions of the Association for Computational Linguistics* 10, pp. 145–162. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00452. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00452/1987010/tacl_a_00452.pdf. URL: https://doi.org/10.1162/tacl_a_00452.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

– (Nov. 2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. URL: https://aclanthology.org/2020.emnlp-main.365.

Reynolds, Laria and Kyle McDonell (2021). "Prompt programming for large language models: Beyond the few-shot paradigm." In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7. URL: https://arxiv.org/abs/2102.07350.

Ri, Ryokan and Yoshimasa Tsuruoka (July 2020). "Revisiting the Context Window for Cross-lingual Word Embeddings." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 995–1005. DOI: 10.18653/v1/2020.acl-main.94. URL: https://aclanthology.org/2020.acl-main.94.

Roy Dipta, Shubhashis and Sai Vallurupalli (June 2024). "UMBCLU at SemEval-2024 Task 1: Semantic Textual Relatedness with and without machine translation." In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, pp. 1351–1357. DOI: 10.18653/v1/2024.semeval-1.195. URL: https://aclanthology.org/2024.semeval-1.195/.

Rubenstein, Herbert and John B Goodenough (1965). "Contextual correlates of synonymy." In: *Communications of the ACM* 8.10, pp. 627–633.

Rubin, Ohad, Jonathan Herzig, and Jonathan Berant (July 2022). "Learning To Retrieve Prompts for In-Context Learning." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 2655–

2671. DOI: 10.18653/v1/2022.naacl-main.191. URL: https://aclanthology.org/2022.naacl-main.191.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (May 2022). "Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold." In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 2340–2354. DOI: 10.18653/v1/2022.findings-acl.184. URL: https://aclanthology.org/2022.findings-acl.184.

SU, Hongjin et al. (2023). "Selective Annotation Makes Language Models Better Few-Shot Learners." In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=qY1hlv7gwg.

Sajjad, Hassan, Firoj Alam, Fahim Dalvi, and Nadir Durrani (Oct. 2022). "Effect of Post-processing on Contextualized Word Representations." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 3127–3142. URL: https://aclanthology.org/2022.coling-1.277.

Sakaizawa, Yuya and Mamoru Komachi (May 2018). "Construction of a Japanese Word Similarity Dataset." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1152.

Sanh, Victor et al. (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=9Vrb9D0WI4.

Schick, Timo and Hinrich Schütze (Nov. 2021a). "Generating Datasets with Pretrained Language Models." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6943–6951. DOI: 10.18653/v1/2021.emnlp-main.555. URL: https://aclanthology.org/2021.emnlp-main.555.

– (June 2021b). "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 2339–2352. DOI: 10.18653/v1/2021.naacl-main.185. URL: https://aclanthology.org/2021.naacl-main.185.

Schnabel, Tobias, Igor Labutov, David Mimno, and Thorsten Joachims (Sept. 2015). "Evaluation methods for unsupervised word embeddings." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Com-

putational Linguistics, pp. 298–307. DOI: 10.18653/v1/D15-1036. URL: https://aclanthology.org/D15-1036.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr (2024). "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting." In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=RIu5lyNXjT.

Seonwoo, Yeon, Guoyin Wang, Changmin Seo, Sajal Choudhary, Jiwei Li, Xiang Li, Puyang Xu, Sunghyun Park, and Alice Oh (July 2023). "Ranking-Enhanced Unsupervised Sentence Representation Learning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 15783–15798. DOI: 10.18653/v1/2023.acl-long.879. URL: https://aclanthology.org/2023.acl-long.879.

Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou (2023a). "Large language models can be easily distracted by irrelevant context." In: *International Conference on Machine Learning*. PMLR, pp. 31210–31227. URL: https://proceedings.mlr.press/v202/shi23a.html.

Shi, Freda et al. (2023b). "Language models are multilingual chain-of-thought reasoners." In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=fR3wGCk-IXp.

Shi, Ning et al. (June 2024). "UAlberta at SemEval-2024 Task 1: A Potpourri of Methods for Quantifying Multilingual Semantic Textual Relatedness and Similarity." In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, pp. 1798–1805. DOI: 10.18653/v1/2024.semeval-1.254. URL: https://aclanthology.org/2024.semeval-1.254/.

Shi, Peng, Rui Zhang, He Bai, and Jimmy Lin (Dec. 2022). "XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing." In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5248–5259. DOI: 10.18653/v1/2022.findings-emnlp.384. URL: https://aclanthology.org/2022.findings-emnlp.384.

Si, Chenglei, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang (2023). "Prompting GPT-3 To Be Reliable." In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=98p5x51L5af.

Springer, Jacob Mitchell, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan (2025). "Repetition Improves Language Model Embeddings." In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Ahlrf2HGJR.

Su, Jianlin, Jiarun Cao, Weijie Liu, and Yangyiwen Ou (2021). "Whitening Sentence Representations for Better Semantics and Faster Retrieval." In: *arXiv preprint arXiv:2103.15316*. URL: https://arxiv.org/pdf/2103.15316.pdf.

Tan, Hao and Mohit Bansal (Nov. 2019). "LXMERT: Learning Cross-Modality Encoder Representations from Transformers." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5100–5111. DOI: 10.18653/v1/D19-1514. URL: https://aclanthology.org/D19-1514.

– (Nov. 2020). "Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2066–2080. DOI: 10.18653/v1/2020.emnlp-main.162. URL: https://aclanthology.org/2020.emnlp-main.162.

Tang, Zineng, Jaemin Cho, Hao Tan, and Mohit Bansal (2021). "VidLanKD: Improving Language Understanding via Video-Distilled Knowledge Transfer." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 24468–24481. URL: https://papers.nips.cc/paper/2021/file/ccdf3864e2fa9089f9eca4fc7a48ea0a-Paper.pdf.

Tanwar, Eshaan, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty (July 2023). "Multilingual LLMs are Better Cross-lingual In-context Learners with Alignment." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 6292–6307. DOI: 10.18653/v1/2023.acl-long.346. URL: https://aclanthology.org/2023.acl-long.346.

Team, Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. (2023). *Gemini: a family of highly capable multimodal models*. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

Team, NLLB et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation*. arXiv: 2207.04672 [cs.CL]. URL: https://arxiv.org/abs/2207.04672.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. (2023). "Llama 2: Open foundation

and fine-tuned chat models." In: *arXiv*. URL: https://arxiv.org/abs/2307.09288.

Uemura, Kosei, Miaoran Zhang, and David Ifeoluwa Adelani (2025). *AfriMTEB and AfriE5: Benchmarking and Adapting Text Embedding Models for African Languages*. arXiv: 2510.23896 [cs.CL]. URL: https://arxiv.org/abs/2510.23896.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Venekoski, Viljami and Jouko Vankka (May 2017). "Finnish resources for evaluating language model semantics." In: *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 231–236. URL: https://aclanthology.org/W17-0228.

Von Oswald, Johannes, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov (2023). "Transformers Learn In-Context by Gradient Descent." In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 35151–35174. URL: https://proceedings.mlr.press/v202/von-oswald23a.html.

Wan, Xingchen, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Eisenschlos, Sercan Arik, and Tomas Pfister (Dec. 2023). "Universal Self-Adaptive Prompting." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 7437–7462. DOI: 10.18653/v1/2023.emnlp-main.461. URL: https://aclanthology.org/2023.emnlp-main.461.

Wang, Boshi, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun (July 2023a). "Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 2717–2739. DOI: 10.18653/v1/2023.acl-long.153. URL: https://aclanthology.org/2023.acl-long.153.

Wang, Dong, Ning Ding, Piji Li, and Haitao Zheng (Aug. 2021). "CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*

the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, pp. 2332–2342. DOI: 10.18653/v1/2021.acl-long.181. URL: https://aclanthology.org/2021.acl-long.181.

Wang, Lean, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun (2023b). "Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 9840–9855. URL: https://aclanthology.org/2023.emnlp-main.609.

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei (2024a). *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. arXiv: 2212.03533 [cs.CL]. URL: https://arxiv.org/abs/2212.03533.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei (Aug. 2024b). "Improving Text Embeddings with Large Language Models." In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 11897–11916. DOI: 10.18653/v1/2024.acl-long.642. URL: https://aclanthology.org/2024.acl-long.642/.

Wang, Mingyang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze (July 2023c). "NLNDE at SemEval-2023 Task 12: Adaptive Pretraining and Source Language Selection for Low-Resource Multilingual Sentiment Analysis." In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, pp. 488–497. DOI: 10.18653/v1/2023.semeval-1.68. URL: https://aclanthology.org/2023.semeval-1.68.

Wang, Tongzhou and Phillip Isola (2020). "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 9929–9939. URL: http://proceedings.mlr.press/v119/wang20k/wang20k.pdf.

Wang, Xinyi, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang (2023d). "Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning." In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=BGvkwZEGt7.

Wang, Yizhong et al. (Dec. 2022). "Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association

for Computational Linguistics, pp. 5085–5109. DOI: 10.18653/v1/2022.emnlp-main.340. URL: https://aclanthology.org/2022.emnlp-main.340.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Wei, Jerry, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. (2023). "Larger language models do in-context learning differently." In: *arXiv*. URL: https://arxiv.org/abs/2303.03846.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. arXiv: 2302.11382 [cs.SE]. URL: https://arxiv.org/abs/2302.11382.

Whitehouse, Chenxi, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata (June 2024). "Low-Rank Adaptation for Multilingual Summarization: An Empirical Study." In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 1202–1228. DOI: 10.18653/v1/2024.findings-naacl.77. URL: https://aclanthology.org/2024.findings-naacl.77/.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu (2016). "Towards Universal Paraphrastic Sentence Embeddings." In: *International Conference on Learning Representations (ICLR)*. URL: https://arxiv.org/abs/1511.08198.

Wieting, John and Kevin Gimpel (July 2018). "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 451–462. DOI: 10.18653/v1/P18-1042. URL: https://aclanthology.org/P18-1042.

Wieting, John, Graham Neubig, and Taylor Berg-Kirkpatrick (Nov. 2020). "A Bilingual Generative Transformer for Semantic Sentence Embedding." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 1581–1594. DOI: 10.18653/v1/2020.emnlp-main.122. URL: https://aclanthology.org/2020.emnlp-main.122.

Williams, Adina, Nikita Nangia, and Samuel Bowman (June 2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: https://aclanthology.org/N18-1101.

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

Workshop, BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. (2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv: 2211.05100 [cs.CL]. URL: https://arxiv.org/abs/2211.05100.

Wu, Shijie and Mark Dredze (Nov. 2019). "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077.

Wu, Zhengxuan, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts (2024). "ReFT: Representation Finetuning for Language Models." In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. Curran Associates, Inc., pp. 63908–63962. DOI: 10.52202/079017-2041. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/75008a0fba53bf13b0bb3b7bff986e0e-Paper-Conference.pdf.

Wu, Zhenyu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao (July 2023). "OpenICL: An Open-Source Framework for In-context Learning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Toronto, Canada: Association for Computational Linguistics, pp. 489–498. DOI: 10.18653/v1/2023.acl-demo.47. URL: https://aclanthology.org/2023.acl-demo.47.

Wu, Zhuofeng, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma (2020). *CLEAR: Contrastive Learning for Sentence Representation*. arXiv: 2012.15466 [cs.CL]. URL: https://arxiv.org/abs/2012.15466.

Xie, Sang Michael, Aditi Raghunathan, Percy Liang, and Tengyu Ma (2022a). "An Explanation of In-context Learning as Implicit Bayesian Inference." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=RdJVFCHjUMI.

Xie, Zhihui, Handong Zhao, Tong Yu, and Shuai Li (Dec. 2022b). "Discovering Low-rank Subspaces for Language-agnostic Multilingual Representations." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5617–5633. DOI: 10.18653/v1/2022.emnlp-main.379. URL: https://aclanthology.org/2022.emnlp-main.379.

Xu, Heng-Da, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao (Aug. 2021). "Read, Listen, and See: Leveraging Multimodal Information Helps Chinese Spell Checking." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 716–728. DOI: 10.18653/v1/2021.findings-acl.64. URL: https://aclanthology.org/2021.findings-acl.64.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41.

Yan, Yuanmeng, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu (Aug. 2021). "ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 5065–5075. DOI: 10.18653/v1/2021.acl-long.393. URL: https://aclanthology.org/2021.acl-long.393.

Yang, Yinfei, Yuan Zhang, Chris Tar, and Jason Baldridge (Nov. 2019). "PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3687–3692. DOI: 10.18653/v1/D19-1382. URL: https://aclanthology.org/D19-1382.

Yao, Shaowei and Xiaojun Wan (July 2020). "Multimodal Transformer for Multimodal Machine Translation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4346–4350. DOI: 10.1

8653/v1/2020.acl-main.400. URL: https://aclanthology.org/2020.acl-main.400.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan (2023). "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., pp. 11809–11822. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.

Yao, Yao, Zuchao Li, and Hai Zhao (June 2024). "GoT: Effective Graph-of-Thought Reasoning in Language Models." In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 2901–2921. DOI: 10.18653/v1/2024.findings-naacl.183. URL: https://aclanthology.org/2024.findings-naacl.183/.

Yin, Shukang, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen (Nov. 2024). "A survey on multimodal large language models." In: *National Science Review* 11.12. ISSN: 2053-714X. DOI: 10.1093/nsr/nwae403. URL: http://dx.doi.org/10.1093/nsr/nwae403.

Yoo, Kang Min, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim (Dec. 2022). "Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2422–2437. DOI: 10.18653/v1/2022.emnlp-main.155. URL: https://aclanthology.org/2022.emnlp-main.155.

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier (2014). "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions." In: *Transactions of the Association for Computational Linguistics* 2, pp. 67–78. DOI: 10.1162/tacl_a_00166. URL: https://aclanthology.org/Q14-1006.

Zablocki, Eloi, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari (2018). "Learning multi-modal word representation grounded in visual context." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11939.

Zhang, Junlei, Zhenzhong Lan, and Junxian He (Dec. 2023a). "Contrastive Learning of Sentence Embeddings from Scratch." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 3916–3932. DOI: 10.18653/v1/2023.emnlp-main.238. URL: https://aclanthology.org/2023.emnlp-main.238/.

Zhang, Miaoran, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach (Aug. 2024a). "The Impact of Demonstrations on Multilingual In-Context Learning: A Multidimensional Analysis." In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 7342–7371. DOI: 10.18653/v1/2024.findings-acl.438. URL: https://aclanthology.org/2024.findings-acl.438.

Zhang, Miaoran, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow (July 2022). "MCSE: Multimodal Contrastive Learning of Sentence Embeddings." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 5959–5969. DOI: 10.18653/v1/2022.naacl-main.436. URL: https://aclanthology.org/2022.naacl-main.436.

Zhang, Miaoran, Mingyang Wang, Jesujoba Alabi, and Dietrich Klakow (June 2024b). "AAdaM at SemEval-2024 Task 1: Augmentation and Adaptation for Multilingual Semantic Textual Relatedness." In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá. Mexico City, Mexico: Association for Computational Linguistics, pp. 800–810. DOI: 10.18653/v1/2024.semeval-1.114. URL: https://aclanthology.org/2024.semeval-1.114.

Zhang, Ruochen, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji (Dec. 2023b). "Multilingual Large Language Models Are Not (Yet) Code-Switchers." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 12567–12582. DOI: 10.18653/v1/2023.emnlp-main.774. URL: https://aclanthology.org/2023.emnlp-main.774.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020a). "BERTScore: Evaluating Text Generation with BERT." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, Yan, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing (Nov. 2020b). "An Unsupervised Sentence Embedding Method by Mutual Information Maximization." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 1601–1610. DOI: 10.18653/v1/2020.emnlp-main.124. URL: https://aclanthology.org/2020.emnlp-main.124.

Zhang, Yuan, Jason Baldridge, and Luheng He (June 2019a). "PAWS: Paraphrase Adversaries from Word Scrambling." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1298–1308. DOI: 10.18653/v1/N19-1131. URL: https://aclanthology.org/N19-1131.

Zhang, Zhuosheng, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao (2019b). "Neural machine translation with universal visual representation." In: *International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=Byl8hhNYPS.

Zhao, Yanpeng and Ivan Titov (Nov. 2020). "Visually Grounded Compound PCFGs." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4369–4379. DOI: 10.18653/v1/2020.emnlp-main.354. URL: https://aclanthology.org/2020.emnlp-main.354.

Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh (2021). "Calibrate before use: Improving few-shot performance of language models." In: *International Conference on Machine Learning*. PMLR, pp. 12697–12706. URL: https://proceedings.mlr.press/v139/zhao21c.html.

Zhou, Kun, Beichen Zhang, Xin Zhao, and Ji-Rong Wen (May 2022). "Debiased Contrastive Learning of Unsupervised Sentence Representations." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6120–6130. DOI: 10.18653/v1/2022.acl-long.423. URL: https://aclanthology.org/2022.acl-long.423.

Zhu, Dawei, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow (Nov. 2024a). "Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice?" In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 388–409. DOI: 10.18653/v1/2024.emnlp-main.24. URL: https://aclanthology.org/2024.emnlp-main.24/.

Zhu, Dawei, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow (July 2023). "Weaker Than You Think: A Critical Look at Weakly Supervised Learning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 14229–14253. DOI: 10.18653/v1/2023.acl-long.796. URL: https://aclanthology.org/2023.acl-long.796.

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li (June 2024b). "Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis." In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 2765–2781. DOI: 10.18653/v1/2024.findings-naacl.176. URL: https://aclanthology.org/2024.findings-naacl.176.

Zouhar, Vilém, Marius Mosbach, Miaoran Zhang, and Dietrich Klakow (May 2022). "Knowledge Base Index Compression via Dimensionality and Precision Reduction." In: *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*. Dublin, Ireland and Online: Association for Computational Linguistics, pp. 41–53. DOI: 10.18653/v1/2022.spanlp-1.5. URL: https://aclanthology.org/2022.spanlp-1.5.

APPENDIX

A.1 NUMBER OF DEMONSTRATIONS

The language-specific results for each task are shown in Figures A.1–A.10. The order of languages follows their data ratio in the CommonCrawl corpus[1] from high-resource to low-resource. We observe large variations in model performance across different languages. For instance, there exists a large performance disparity between English and Urdu in XNLI. In XCOPA, the performance of Quechua is significantly worse compared to other languages.

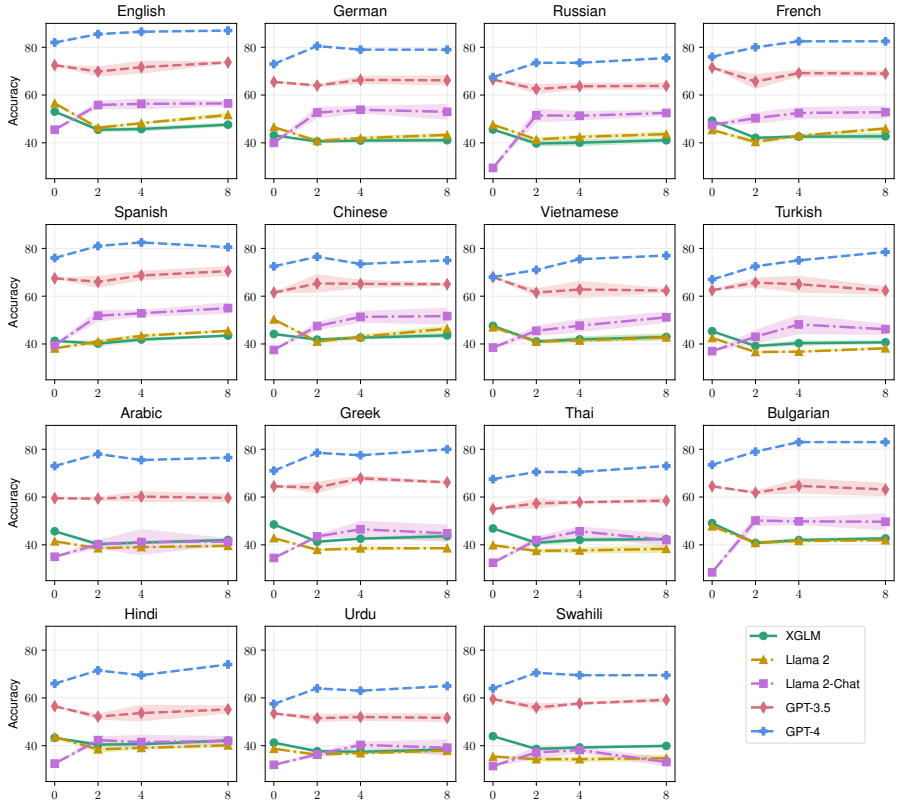


Figure A.1: Language-specific performance on XNLI with a varying number of demonstrations.

A.2 DEMONSTRATION QUALITY

In Figures A.11–A.20, we show the language-specific results for each task, in which we can see language discrepancies with different types of demonstrations.

1 http://commoncrawl.org

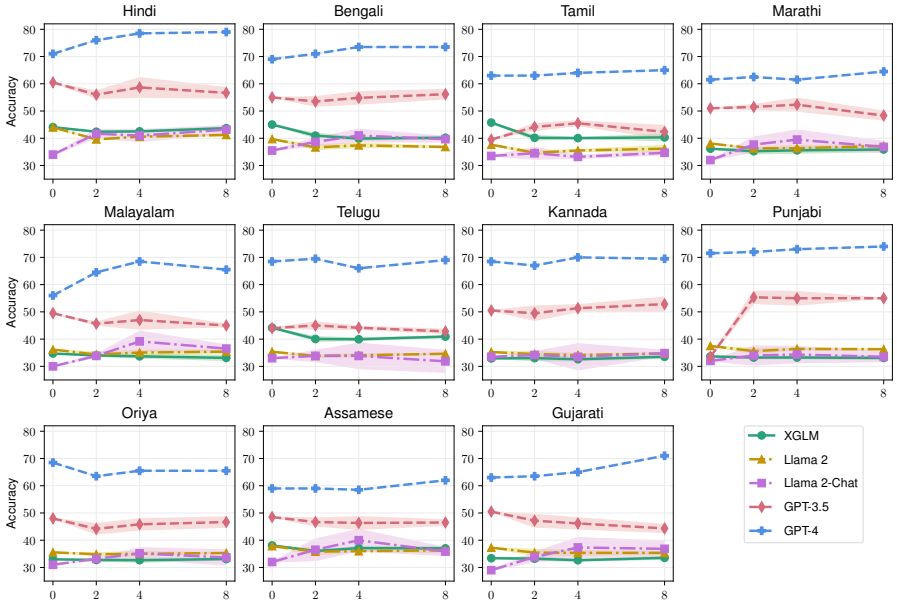


Figure A.2: Language-specific performance on IndicXNLI with a varying number of demonstrations.

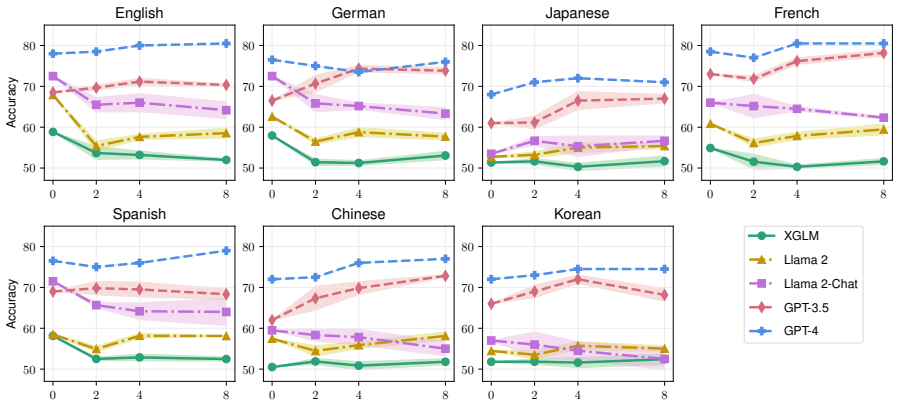


Figure A.3: Language-specific performance on PAWS-X with a varying number of demonstrations.

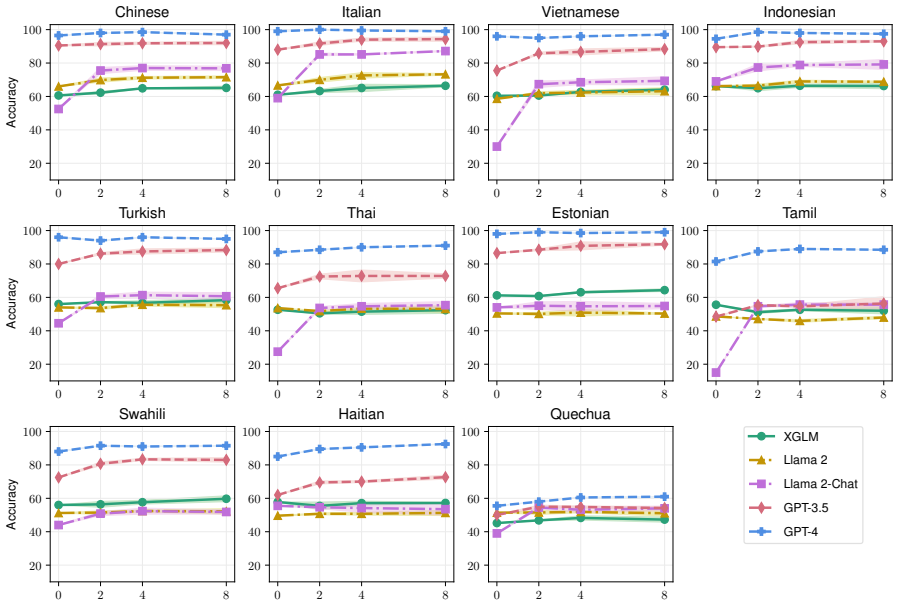


Figure A.4: Language-specific performance on XCOPA with a varying number of demonstrations.

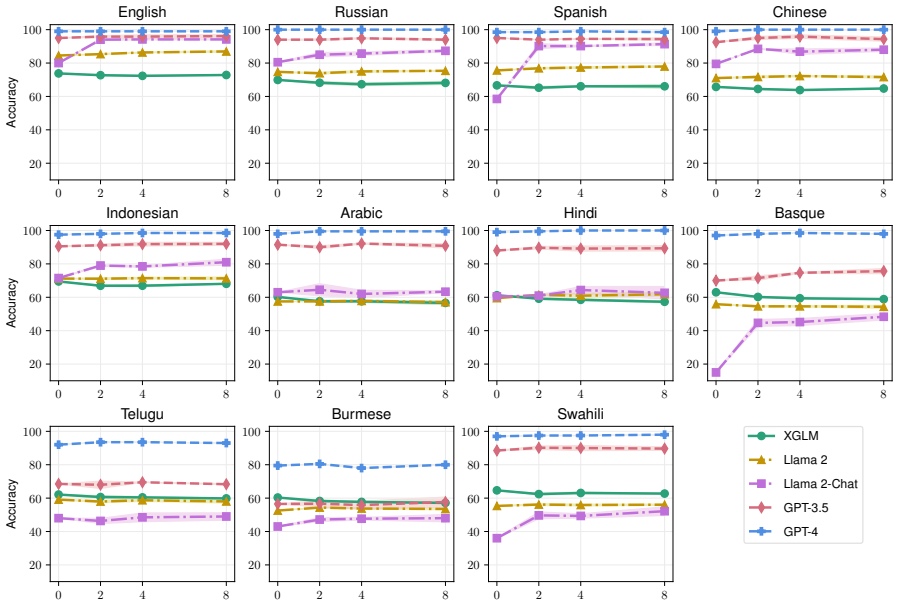


Figure A.5: Language-specific performance on XStoryCloze with a varying number of demonstrations.

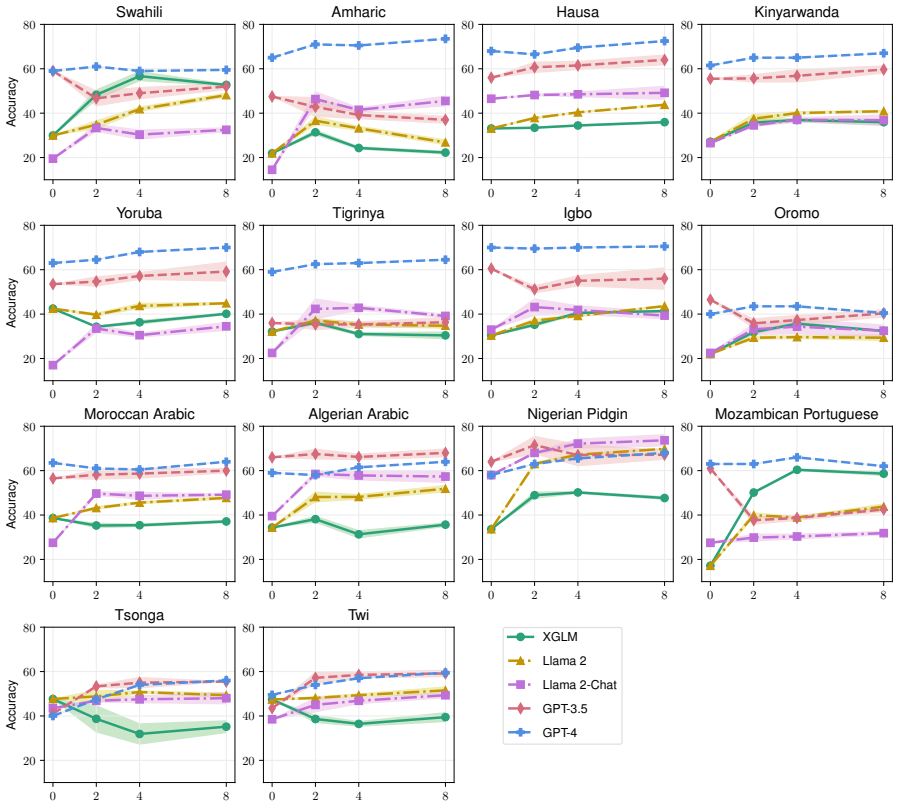


Figure A.6: Language-specific performance on AfriSenti with a varying number of demonstrations.

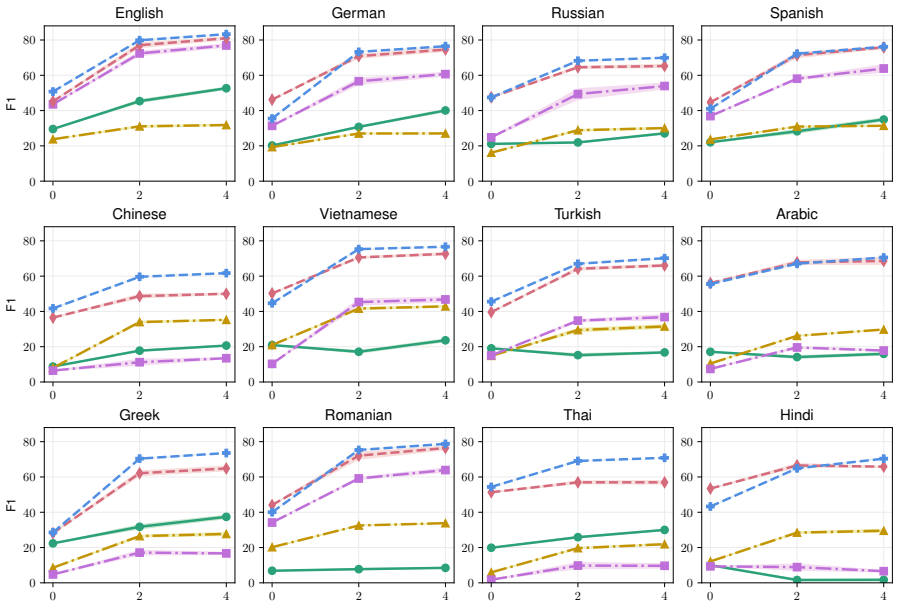


Figure A.7: Language-specific performance on XQuAD with a varying number of demonstrations.

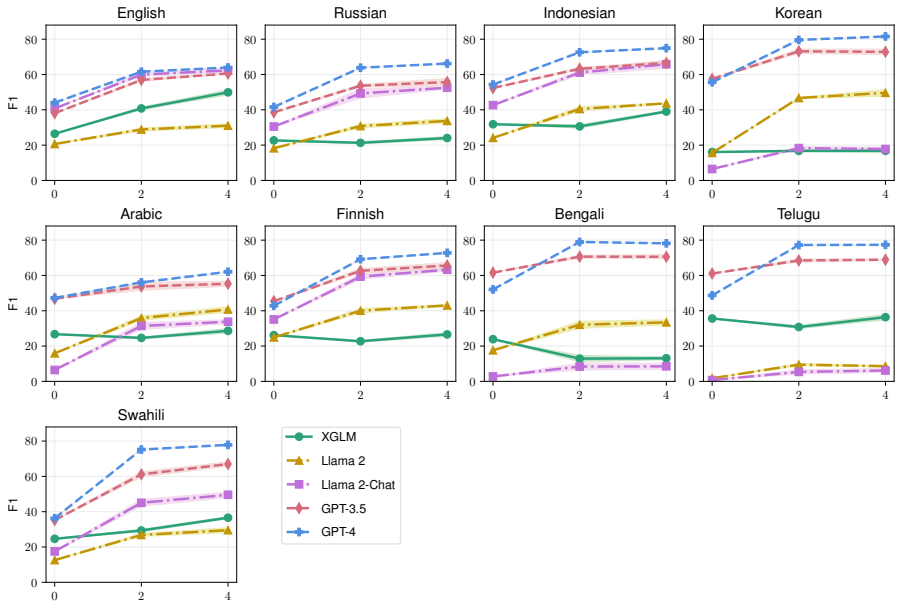


Figure A.8: Language-specific performance on TyDiQA with a varying number of demonstrations.

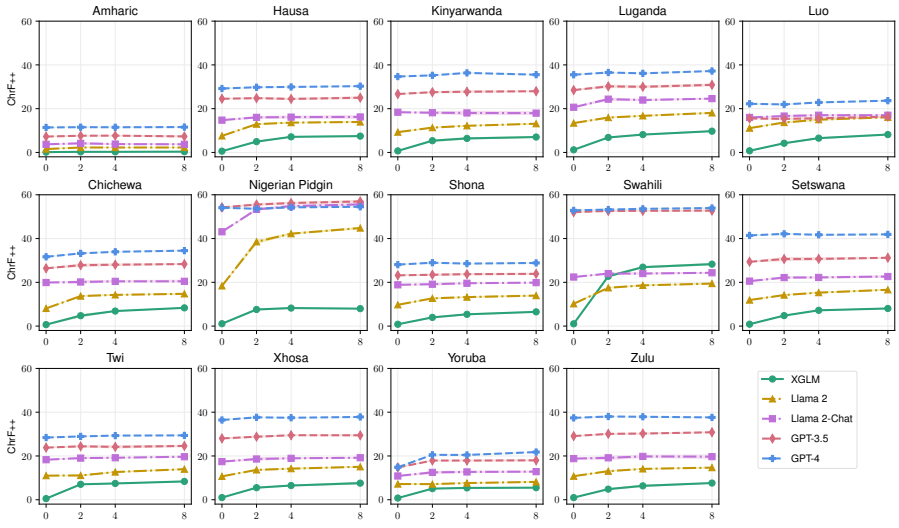


Figure A.9: Language-specific performance on MAFAND (en-xx) with a varying number of demonstrations.

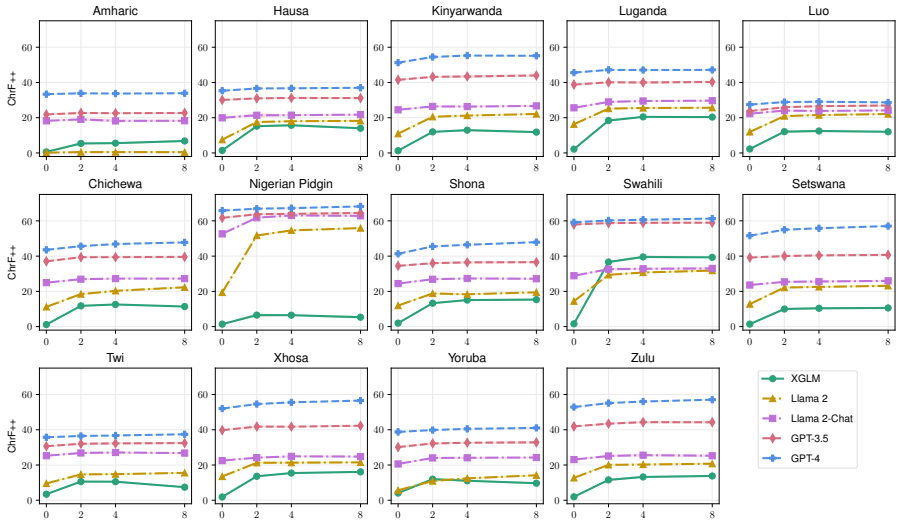


Figure A.10: Language-specific performance on MAFAND (xx-en) with a varying number of demonstrations.

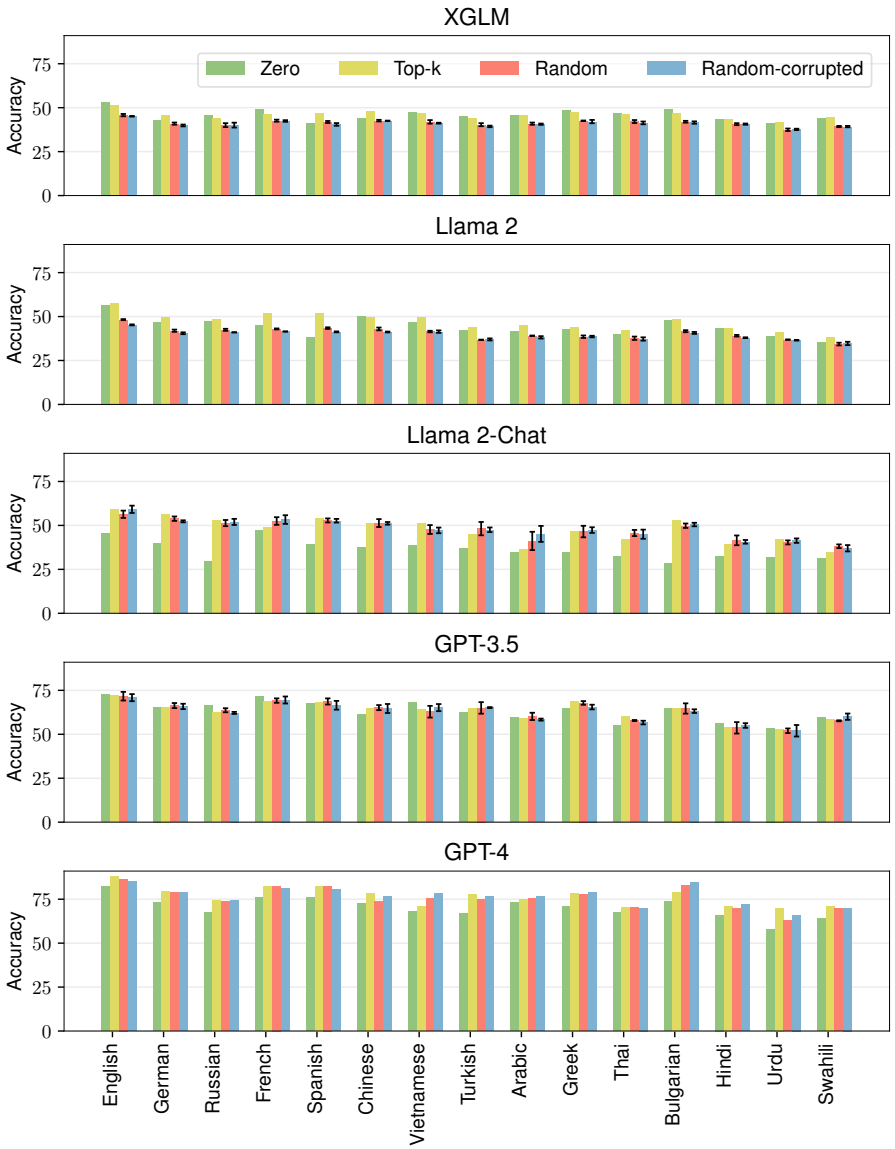


Figure A.11: Language-specific performance on XNLI with different types of demonstrations.

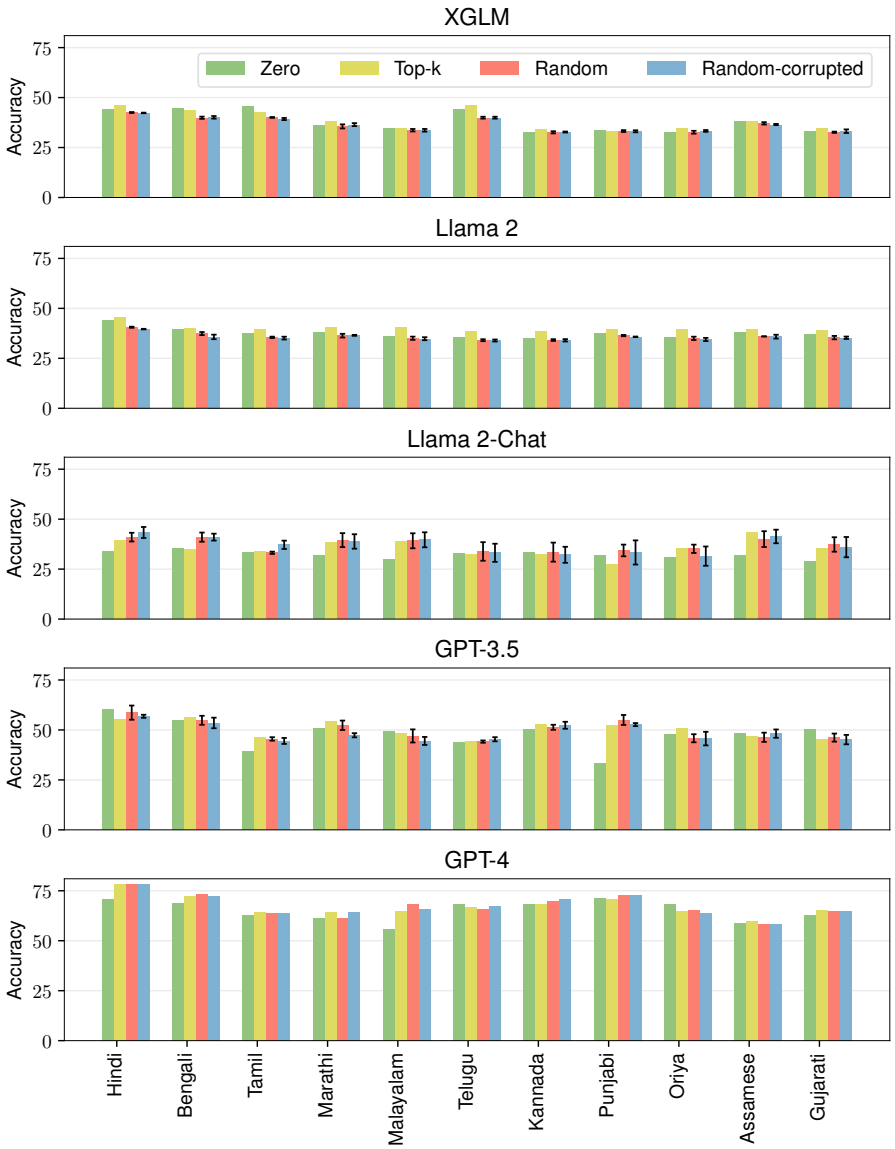


Figure A.12: Language-specific performance on IndicXNLI with different types of demonstrations.

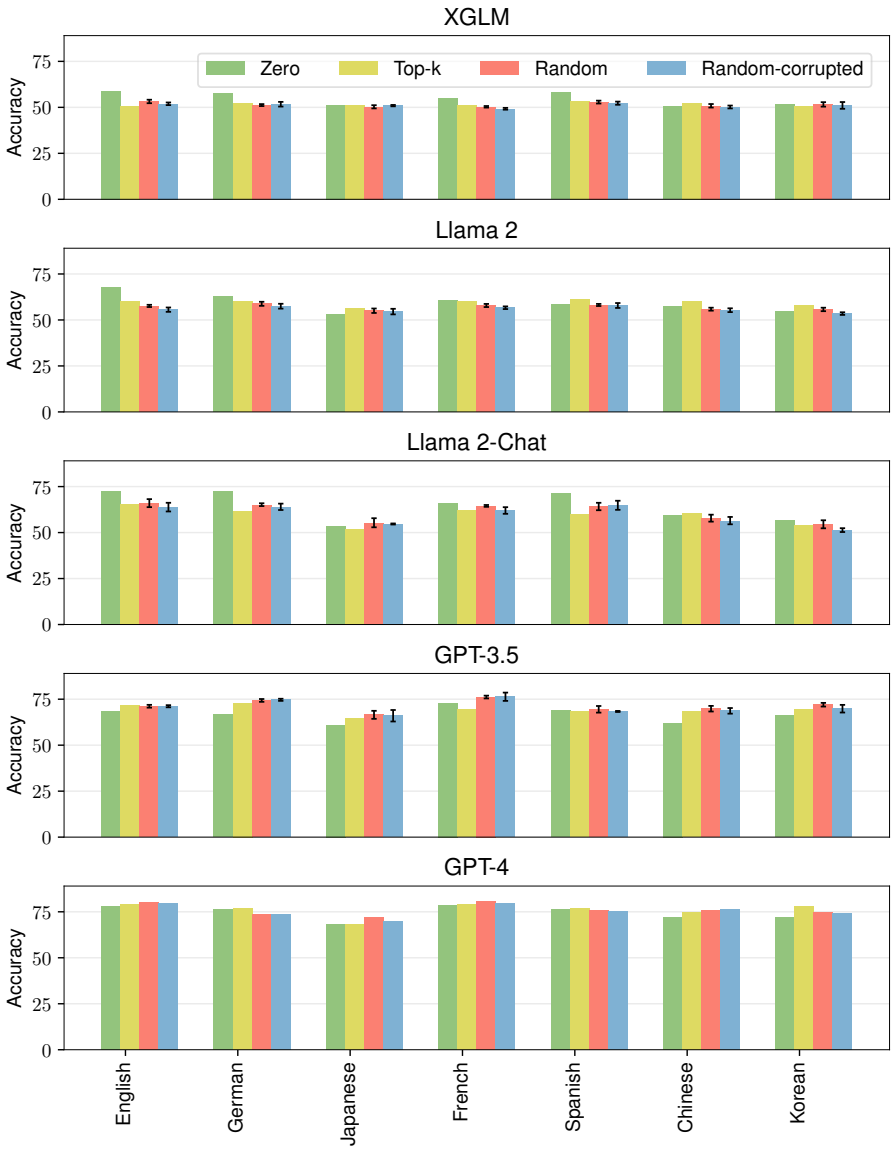


Figure A.13: Language-specific performance on PAWS-X with different types of demonstrations.

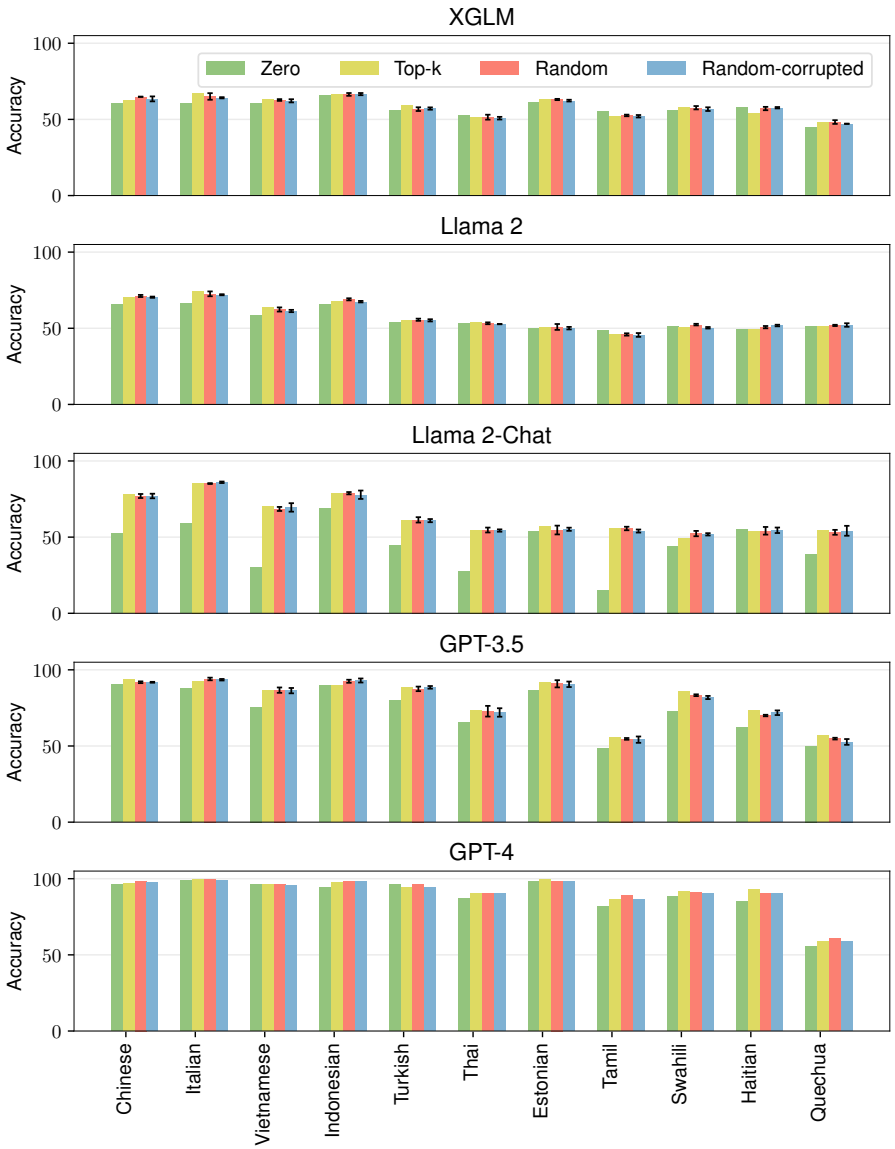


Figure A.14: Language-specific performance on XCOPA with different types of demonstrations.

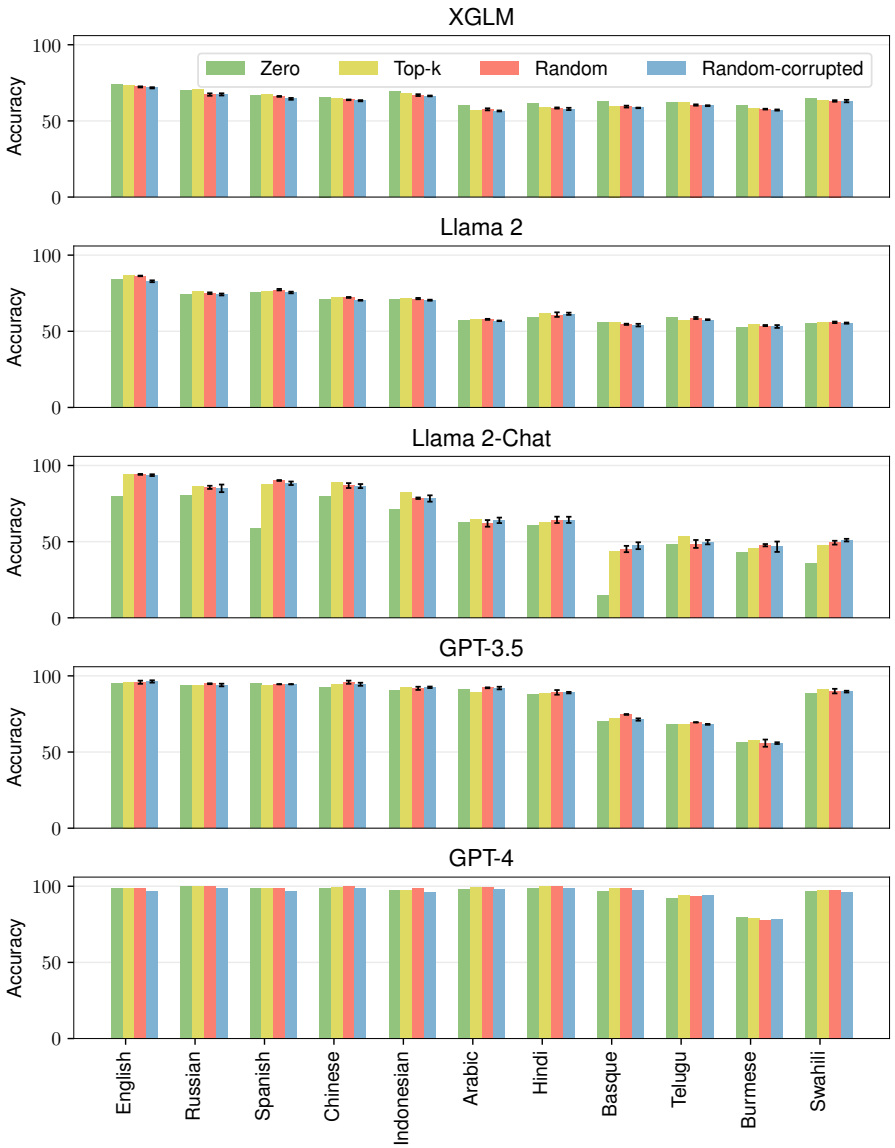


Figure A.15: Language-specific performance on XStoryCloze with different types of demonstrations.

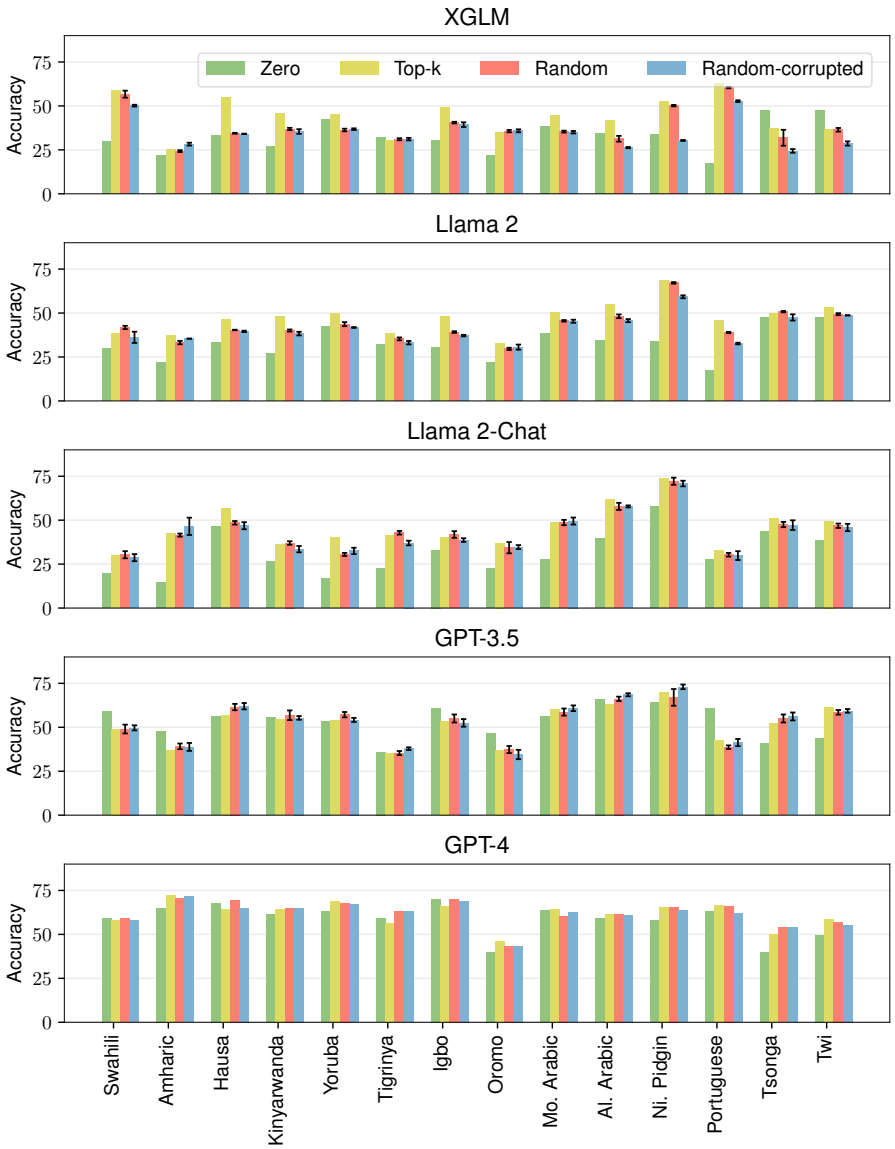


Figure A.16: Language-specific performance on AfriSenti with different types of demonstrations.

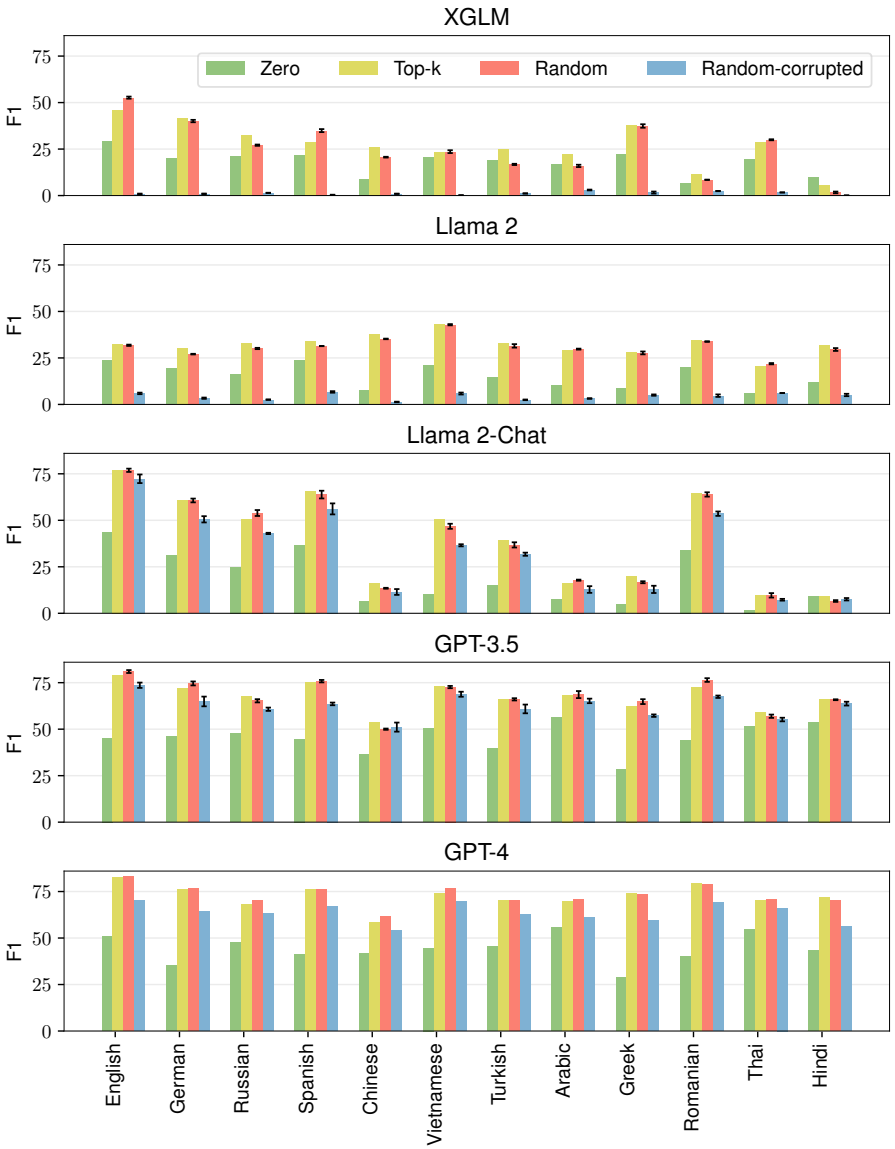


Figure A.17: Language-specific performance on XQuAD with different types of demonstrations.

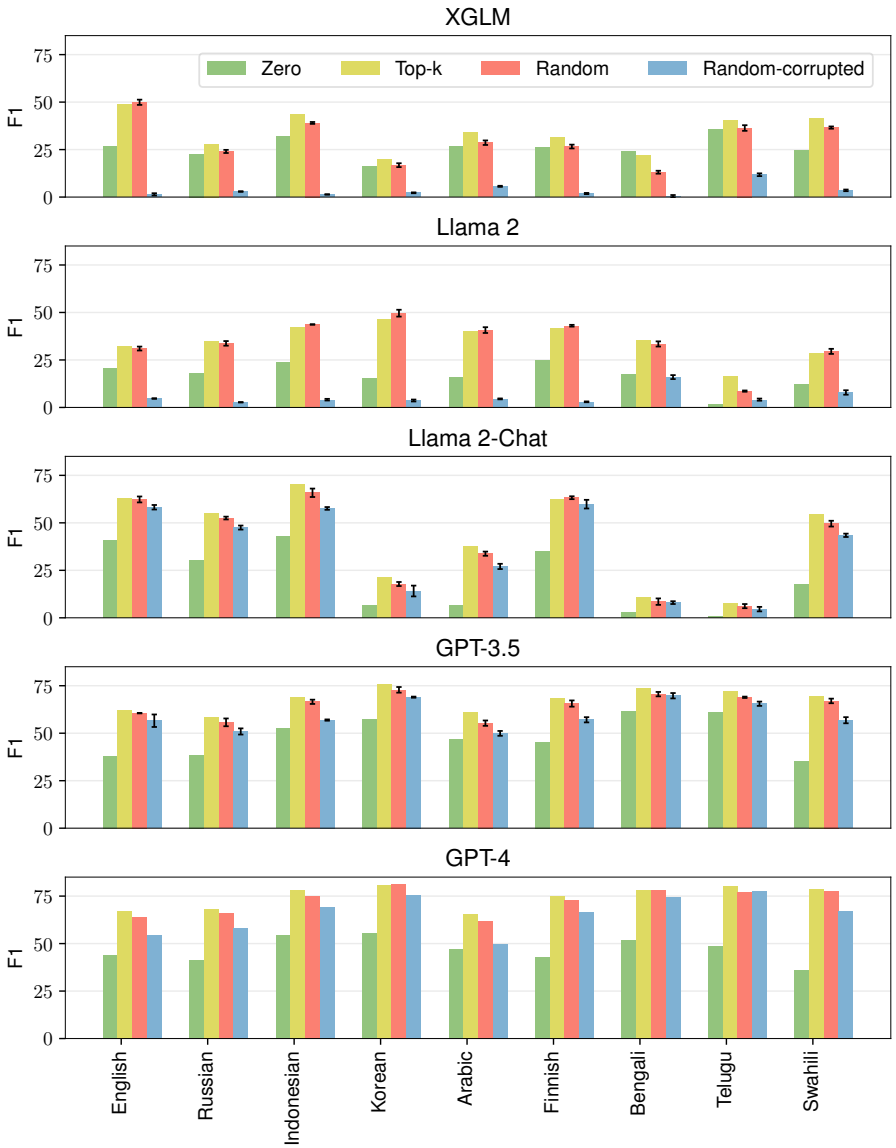


Figure A.18: Language-specific performance on TyDiQA with different types of demonstrations.

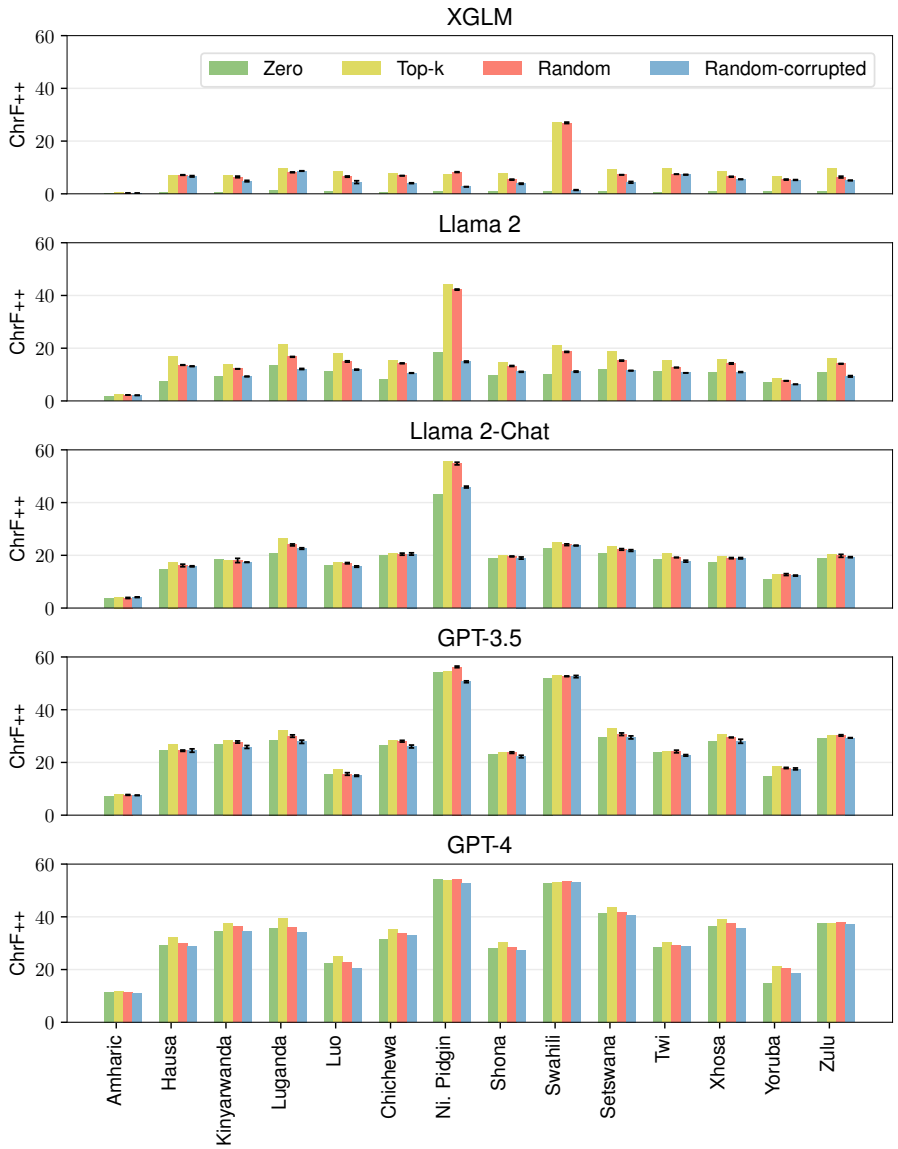


Figure A.19: Language-specific performance on MAFAND (en-xx) with different types of demonstrations.

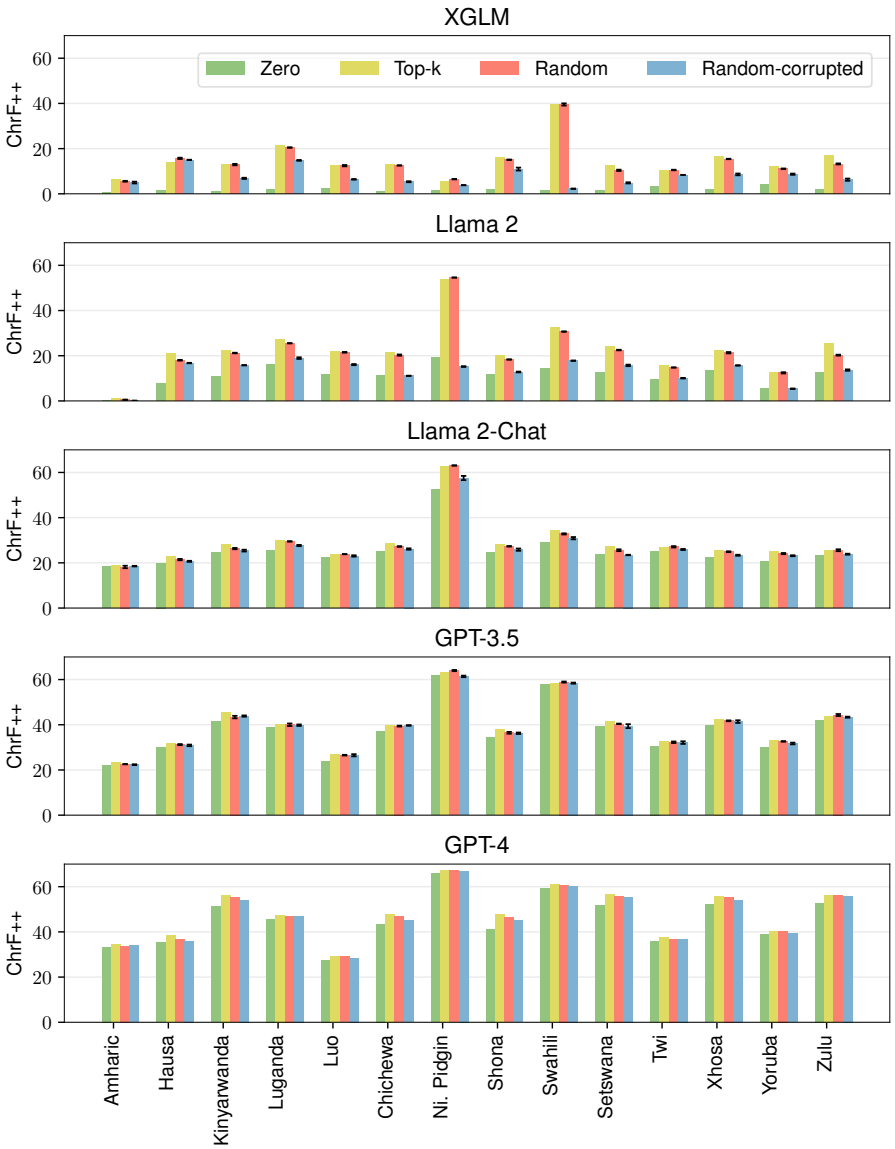


Figure A.20: Language-specific performance on MAFAND (xx-en) with different types of demonstrations.

DECLARATION

I hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

Saarbrücken, 2025

Miaoran Zhang