How and Where to Translate? The Impact of Translation Strategies in Cross-lingual LLM Prompting

Aman Gupta
Amazon
San Francisco, USA
amangta@amazon.com

Yingying Zhuang
Amazon
San Francisco, USA
yyzhuang@amazon.com

Zhou Yu
Amazon
Seattle, USA
amznzya@amazon.com

Ziji Zhang
Amazon
Seattle, USA
czhangzi@amazon.com

Anurag Beniwal
Amazon
San Francisco, USA
beanurag@amazon.com

Abstract

Despite advances in the multilingual capabilities of Large Language Models (LLMs), their performance varies substantially across different languages and tasks. In multilingual retrieval-augmented generation (RAG)-based systems, knowledge bases (KB) are often shared from high-resource languages (such as English) to low-resource ones, resulting in retrieved information from the KB being in a different language than the rest of the context. In such scenarios, two common practices are pre-translation to create a mono-lingual prompt and cross-lingual prompting for direct inference. However, the impact of these choices remains unclear. In this paper, we systematically evaluate the impact of different prompt translation strategies for classification tasks with RAG-enhanced LLMs in multilingual systems. Experimental results show that an optimized prompting strategy can significantly improve knowledge sharing across languages, therefore improve the performance on the downstream classification task. The findings advocate for a broader utilization of multilingual resource sharing and cross-lingual prompt optimization for non-English languages, especially the low-resource ones.

CCS Concepts

• **Information systems** → **Language models**.

Keywords

Prompt Optimization, Cross-lingual Prompting, Multilingual Information Retrieval

ACM Reference Format:

Aman Gupta, Yingying Zhuang, Zhou Yu, Ziji Zhang, and Anurag Beniwal. 2025. How and Where to Translate? The Impact of Translation Strategies in Cross-lingual LLM Prompting. In *Proceedings of First International KDD Workshop on Prompt Optimization, 2025 (Prompt Optimization KDD '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Prompt Optimization KDD '25, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Text classification has always been an active Natural Language Processing (NLP) task with applicability to a wide range of real-world problems, including sentiment analysis, topic labeling, conversation dialogue state tracking, etc. It aims to choose one or multiple labels from a list of pre-defined categories given some informative texts. As Large Language Models (LLMs) have emerged as a powerful tool for language understanding and generation[30, 31], LLM-based classification has become an active area of research and has achieved remarkable advancements [7, 25, 26]. In dialogue systems and conversation AIs, a Retrieval-Augmented Generation (RAG) component is often deployed to generate a list of probable candidates for the task and a LLM is then used as a re-ranker to choose the final label from the list [5]. A multilingual retriever has become a common practice where the retrieved information and candidate lists are in a different language from the rest contexts and the label. Figure 1 demonstrates such an example for user intent classification where given a user utterance in French, a multilingual retriever will identify texts in English in the Knowledge Base (KB) with similar semantics and use their corresponding user intents as the candidate list. Therefore the candidate list is in English while the user utterance is in French, resulting in a cross-lingual prompt.

One common strategy is to pass this prompt directly to a LLM with multilingual capability to predict the final user intent. Many of today's frontier LLMs, such as Anthropic's Claude 3 [1] and Cohere's Command A [9] are focused to excel in the multilingual settings to support such business needs. However, the majority of the state-of-the-art LLMs cover only a small percentage of the world's spoken languages and heavily favor those with abundant resources such as English [14, 19, 20, 28]. This creates a gap in the availability of models that can be applied for direct inference with cross-lingual prompts, specially when the prompt contains minority languages, as well as a bias in performance towards English and other high-resource languages compared to the lower resourced ones.

The second option is pre-translation, where parts of the prompt are translated to ensure monolingualism (often English) before applying a LLM as the re-ranker [3, 21]. This removes the multilingual dependency of the LLM therefore ensures the optimal English performance is utilized. However, as translation may contain errors or terms native speakers don't actually use, or fail to account for the

contexts of the local language, this strategy introduces complexities and risks of information loss [20].

In this study, we answer the specific question of *which components of the input prompt should be pre-translated and into which language to achieve optimal performance*. We systematically evaluate six translation strategies across prompt components—from keeping all components in English to selectively translating candidates, identity statements, rules, or all components to the source language. We perform experiments on French and Hindi as the source languages as they respectively represent the high-resource and low-resource scenarios, with five diverse language models on the task of dialogue intent detection. Our performance results suggest that while an optimal pre-translated prompt can lead to improved performance, the impact of pre-translation varies significantly based on the model choice and the nature of the source language.

2 Related Work

Various studies have proposed to use selective pre-translation in multi-lingual settings, where only specific parts of the prompt is pre-translated to optimize performance. For example, [15] demonstrated that translating only the context to English outperforms direct-inference in summarization and NLI tasks, while [3] showed that translating only the few-shot examples while keeping the rest of the prompt in the source language yielded optimal results for low-resource languages while does not make significant impact for high-resource languages compared to monolingual prompting. Furthermore, [11] proposed a cross-lingual prompting method that translates only the question and answer parts for the QA task to optimize in-context learning. One common theme from all these studies is **how and where to translate** in cross-lingual prompting is highly dependent on 1) the task, 2) the language setting (high vs low resource), 3) translation quality, 4) the tokenizer, 5) and finally the LLM deployed for inference. This calls for the need to systematically study the effectiveness of various prompting strategies across different tasks and LLMs.

[17] is aimed to address this need where they evaluate the impact of different pre-translation strategies across a range of tasks, including NLI, QA, NER, and Summarization, in order to provide general guidelines for choosing optimal strategies in various multilingual settings. However, they did not study LLM-based classification tasks where the prompt contains a list of probable candidates and the LLM is prompted to choose the final answer from the list based on the contexts provided in the prompt. In our work we specifically fill this gap for multilingual classification tasks, motivated by real-world applications where information is often retrieved from a multilingual KB. The ability to share knowledge in multiple languages is a fundamental component to ensure the development of Conversation AIs systems to reach a broader range of communities, and we hope our work spurs research to foster inclusivity, accessibility and collaboration for both businesses and individuals.

3 Methodology

In this section, we outline the experimental framework we used to evaluate the performance impact of translating prompts into different languages. We begin by presenting the overall task description in Section 3.1, followed by a brief overview of the dataset in Section 3.2. Section 3.3 describes the different prompts we used and, Section 3.4 details our choice of models for evaluation.

3.1 Task Description

We assess the results on the task of intent classification in dialogue systems, where the goal is to detect the user's intent based on their utterances. We follow the two-step RAG setup for intent detection[6, 16, 27], where a retrieval model first retrieves the list of candidate intents and then a LLM picks the final intent from the candidates. For the purpose of this study we focus on prompts for the LLM to pick the final intent from the candidates. We enforce full recall from the retrieval model that is, the correct intent is always present in the candidates. Overall, the prompt contains i) an *identity statement (**I**)* which describes the high level idea of the task, ii) *task rules (**R**)* which present instructions such as how to work on the task and what output format to follow, iii) a *candidate list (**C**)* of the candidate intents and iv) the *speaker utterance (**U**)* for which we need to detect their intent. Figure 1 shows the prompt that we use for the task. We use accuracy as the primary performance evaluation metric. For each instance, we check if only the correct intent (verbatim) is present and no other labels are present in the model response. If these conditions are met, then it's considered accurate; otherwise, it's considered inaccurate.

Strategy Code	Identity	Rules	Candidates
B (Baseline)	English	English	English
C (Candidates)	English	English	Source
I (Identity)	Source	English	English
R (Rules)	English	Source	English
I&R (Identity & Rules)	Source	Source	English
A (All)	Source	Source	Source

Table 1: Translation strategies for prompt components across experimental configurations, showing which components remain in English versus being translated to the source language (Hindi or French).

3.2 Dataset

Data Anonymization Due to business considerations, we are not permitted to share the results using the original user utterances. As a result, we manually anonymized both the labels and user utterances to ensure no personal information is included. Additionally, specific product and service names were denonymized to prevent the identification of the company from the user utterances or label descriptions. Despite these modifications, the conclusions drawn from our experiments remain valid.

The dataset comprises 6,000 utterances in French and in Hindi, respectively. We use 80 intent classes to categorize each user utterance. For each utterance, we select the candidate intents using embedding-based similarity, taking the top-k intents up to a certain threshold. To maintain full recall, we manually add the correct intent when it is not present among these top-k candidates.

3.3 Cross-lingual Prompt Experiments

We specifically selected *Hindi (Hi)* and *French (Fr)* as the source languages for experimentation because they provide excellent diversity and represent different positions on the spectrum of language

Example Prompt

You are an intelligent AI agent and are assigned the task of intent detection based on speaker utterances. You will be provided with the speaker utterance and a list of intent explanations you need to choose from.

You must follow these instructions:
1) Carefully read the utterance and review all candidate intent explanations.
2) Select the ONE intent explanation that most accurately represents the speaker utterance.
3) Respond ONLY with the exact text of the chosen intent explanation. Do not modify the wording of the intent explanation.
4) Do not add any explanations, greetings, or additional text.
5) If multiple intents seem applicable, choose most specific one

Candidate Intent Descriptions:
1. User wants to know the weather forecast for a specific location.
2. User wants to set a reminder or alarm for a future event.
3. User is requesting information about their account balance.
4. User wants to play a specific song or music.
5. User is asking for directions to a location.

Utterance:
"Peux-tu me rappeler de faire les courses à 17 heures ?"

Figure 1: The above figure presents the prompt we use for intent detection and its components: identity statement (blue), task rules (red), candidate intent descriptions (green), and user utterance (purple). The groundtruth response in this case would the utterance "User wants to set a reminder or alarm for a future event."

resource availability and linguistic similarity to English. Hindi is an under-represented language with lower resource availability and greater linguistic dissimilarity to English, while French is a high-resource language with relatively close linguistic similarities to English[2, 8].

Throughout our experiments we keep the utterance (U) in the source language and selectively translate a few other components of the prompt from English into the source language. We use a LLM for translation and manually inspect a subset to ensure good translation quality. We prioritize component combinations that reflect authentic real-world usage scenarios. We name and abbreviate the experiment configurations based on what sections we translate into the source language. Table 1 represents the set of translation components that we use. We treat the configuration with all components in English as **B (Baseline)** and compare the rest of the strategies against it. For multilingual RAG systems, knowledge bases are often maintained in English due to resource availability. Two common approaches emerge: (1) using cross-lingual retrieval[29] to find English documents for source language queries, resulting in English candidates, or (2) pre-translating the knowledge base into the source language and retrieving from the translated KB, resulting in source language candidates. Our Baseline configuration represents the first approach, while configuration **C (Candidates)** simulates the second approach where candidates are in the source language.

Additionally, configuration **I (Identity)** tests the impact of translating only the identity statement, providing insight into whether task framing in the source language affects comprehension. Configuration **R (Rules)** examines the effect of having instructions in the source language while keeping all other elements in English. The **I&R (Identity & Rules)** configuration combines both identity and rules translations to evaluate their synergistic effect, while still maintaining English candidates. Finally, configuration **A (All)** represents full localization, with all prompt components translated to match the utterance language. In configurations **C** and **A**, since the candidates are in the source language, the model response is also in the source language as the model just has to copy one of the candidates. We do not evaluate translating the user utterance into English, and instead primarily focus on methods where static prompt components can be pre-translated offline without impacting runtime performance.

3.4 Models

For our experiments, we selected a diverse set of state-of-the-art multilingual LLMs to ensure comprehensive evaluation across different architectural designs and pre-training approaches. Specifically, we evaluated Llama-3.1-8B[10], Qwen2.5-7B-Instruct[22], BLOOMZ-7b1[18], Mistral-Nemo-Instruct-2407[4], and BLOOMZ-7b1-mt[18]. This selection provides a balanced representation of both general-purpose models (Llama-3.1-8B) and those with enhanced instruction-following capabilities (Qwen2.5-7B-Instruct, Mistral-Nemo-Instruct-2407). Moreover, we deliberately included models with varying degrees of multilingual expertise, from those primarily trained on English (Llama-3.1-8B) to those specifically designed for multilingual tasks (BLOOMZ-7b1, BLOOMZ-7b1-mt). All models have comparable parameter sizes (approximately 7B-13B parameters range), allowing for fair comparison.

3.5 Implementational Details

As mentioned earlier, we evaluate these models with various prompt configurations without explicit tuning on the dataset. We use standard inference parameters of temperature as 0.1, top-p as 0.7 and max_tokens as 512 to have deterministic generation properties. We perform distributed inference using vllm[12] on a machine with 4 NVIDIA L4 GPUs.

4 Results

We compute the average accuracy across all the prompt configurations and present the results in Table 2. Given the baseline performance we report the percentage improvement or decrement in accuracy for each model and prompt configuration.

We observe relatively low absolute baseline performance across all models for the intent detection task. We attribute this to two main factors: i) The intent detection instances are out-of-domain (OOD) for the LLMs, as they have not encountered them during training, and ii) LLMs struggle with verbatim copying portions of the user prompt [23]. Furthermore, we observe varying trends across different LLMs for each prompt configuration. Specifically,

	Hindi (Hi)						French (Fr)					
Model	**B**	**C**	**I**	**R**	**I&R**	**A**	**B**	**C**	**I**	**R**	**I&R**	**A**
Llama-3.1-8B	0.386	-4.9%	-0.8%	-6.7%	-7.3%	-10.9%	0.286	-7.3%	-12.6%	-13.6%	-24.5%	-32.2%
Qwen2.5-7B-Instruct	0.405	-12.8%	-3.2%	+5.4%	+6.7%	-2.2%	0.439	-12.5%	+3.4%	+6.6%	+5.7%	-13.4%
BLOOMZ-7b1	0.308	-1.9%	+0.3%	+3.2%	-3.2%	-7.1%	0.186	+6.5%	+2.2%	+1.1%	+7.0%	+13.4%
Mistral-Nemo-Instruct-2407	0.454	-13.0%	-2.9%	-15.0%	-16.7%	-33.3%	0.443	-2.3%	+0.5%	-2.3%	-2.3%	-9.0%
BLOOMZ-7b1-mt	0.308	-0.3%	+1.9%	+2.3%	+1.0%	-1.3%	0.185	+7.0%	+1.6%	-1.6%	+4.3%	+9.7%

Table 2: Performance comparison of different prompt translation strategies across languages (Hindi and French). Values for baseline (B) are absolute accuracy scores, while other columns (C, I, R, I&R, A) show percentage changes relative to the baseline. Dark green cells indicate strong improvements (≥10%), medium green for moderate improvements (5-10%), and light green for small improvements (0-5%). Similarly, dark red shows strong declines (≥10%), medium red for moderate declines (5-10%), and light red for small declines (0-5%).

Llama-3.1-8B and Mistral-Nemo-Instruct-2407 experience performance decreases when any component is translated into the source language. In contrast, the BLOOMZ-based models show decent improvements in performance across most prompt configurations. Qwen2.5-7B-Instruct demonstrate mixed results, with improvements in the **I**, **R** and **I&R** strategies while decrements in **C** and **A**. From a language perspective, we notice that all models struggled with generating Hindi text (in configurations **C** & **A**), experiencing a decline in performance, while for French, generation in the native language proves beneficial, but only for the BLOOMZ models. In general, having Identity, Rules, or both in the respective language leads to some performance improvements.

5 Discussion and Recommendations

We hypothesize and discuss the following insights about choosing the optimal prompt language for a task:

LLM generation abilities are better in English, especially for low resource languages. Most models perform poorly when the candidates are presented in the source language, thereby forcing the model to generate the response in source language. The only exception is that for the BLOOMZ models in the case of the French language. Such observation is expected and can be attributed to the fact that the amount of training data in English is disproportionately large as compared to other languages. Similar observations have been made in several other works[13, 24].

Keeping Identity and Instructions in the source language can lead to improvements. Our results indicate that translating the identity statement (**I**) and rules (**R**) into the source language often yields performance improvements, particularly for models like Qwen2.5-7B-Instruct and the BLOOMZ variants. We hypothesize that framing the task in the user's native language helps the model better understand the context and intent of the query, while still leveraging the model's stronger English generation capabilities. This suggests that a hybrid approach—where task framing occurs in the source language but generation remains in English—may be optimal for many multilingual applications.

Model architecture and pre-training strategy significantly impact cross-lingual performance. The varying performance across different models suggests that architectural design and pre-training objectives play crucial roles in determining cross-lingual capabilities. For instance, the BLOOMZ models, which were specifically trained with multilingual objectives, demonstrate more consistent improvements across different prompt configurations compared to the others. This indicates that the choice of LLM should be carefully considered based on the specific language requirements of the application.

Language resource availability and task requirements drive optimal translation strategy. Our results reveal contrasting optimal strategies between Hindi (low-resource) and French (high-resource). Hindi generation showed performance degradation across most models, while French generation benefited certain models, suggesting that language resource availability during pre-training significantly impacts prompt translation effectiveness. The inconsistent performance across configurations also indicates that optimal strategies depend on specific task requirements—for intent classification, translating identity and rules proved beneficial but may not generalize to other tasks. This highlights the need for optimization based on both language resource levels and task-specific demands.

6 Conclusion and Future Work

This work addresses a fundamental challenge in multilingual dialogue systems: optimizing cross-lingual prompt strategies without sacrificing runtime performance. Throughout this study, we systematically evaluate the impact of translating parts of the system prompt into the source language of dialogue queries for the dialogue intent detection task. Our results show that the choice of optimal prompt strategy can vary greatly based on the type of language and the inference model being used. In general, we observed improvements when the identity and instructions of the prompt were translated to the source language. For low-resource languages that are lexically dissimilar to English, like Hindi, it's always better to have the generation in English unless the task requires Hindi text. For high-resource languages like French, it can be better to have generation in the source language, depending on the choice of the inference model. Different LLMs behave differently when presented with cross-lingual prompts, which can be attributed to the different data and recipes used to train these models. While models like BLOOMZ-7b1 and BLOOMZ-7b1-mt show consistent improvements in performance with translated prompts, others like Llama-3.1-8B perform better with English-only prompts.

For future work, we plan to extend our analysis to a broader range of language tasks beyond intent classification, including question answering, summarization, and entity extraction, to determine whether our hypotheses about translation strategies generalize across diverse task types. Additionally, we aim to expand our language coverage to include more typologically diverse languages representing various resource levels to better understand how linguistic distance from English impacts optimal prompt translation strategies. We also intend to investigate the role of translation quality more systematically by comparing professional human translations against various machine translation approaches, as well as exploring hybrid prompting strategies that dynamically select which components to translate based on model confidence scores.

References

[1] [n. d.]. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family. Accessed: 2025-04-20.

[2] Ashish Sunil Agrawal, Barah Fazili, and Preethi Jyothi. 2024. Translation Errors Significantly Impact Low-Resource Languages in Cross-Lingual Learning. arXiv:2402.02080 [cs.CL] https://arxiv.org/abs/2402.02080

[3] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528* (2023).

[4] Mistral AI and NVIDIA. 2024. Mistral-Nemo-Instruct-2407: A 12B Parameter Multilingual LLM. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

[5] Nicholas Ampazis. 2024. Improving rag quality for large language models with topic-enhanced reranking. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 74–87.

[6] Gaurav Arora, Shreya Jain, and Srujana Merugu. 2024. Intent Detection in the Age of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 1559–1570. doi:10.18653/v1/2024.emnlp-industry.114

[7] Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned'Small'LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660* (2024).

[8] Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 405–433. doi:10.18653/v1/2024.naacl-long.24

[9] Team Cohere. 2025. Command A: An Enterprise-Ready Large Language Model. arXiv:2504.00698 [cs.CL] https://arxiv.org/abs/2504.00698

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[11] Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. 2024. Cross-lingual QA: A Key to Unlocking In-context Cross-lingual Performance. arXiv:2305.15233 [cs.CL] https://arxiv.org/abs/2305.15233

[12] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[13] Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13171–13189. doi:10.18653/v1/2023.findings-emnlp.878

[14] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *arXiv e-prints* (2024), arXiv–2404.

[15] Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models. arXiv:2403.10258 [cs.CL] https://arxiv.org/abs/2403.10258

[16] Junhua Liu, Tan Yong Keat, Bin Fu, and Kwan Hui Lim. 2024. LARA: Linguistic-Adaptive Retrieval-Augmentation for Multi-Turn Intent Classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 1096–1106. doi:10.18653/v1/2024.emnlp-industry.82

[17] Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 1331–1354. https://aclanthology.org/2025.findings-naacl.73/

[18] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786* (2022).

[19] Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. arXiv:2305.14456 [cs.CL] https://arxiv.org/abs/2305.14456

[20] Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: large language models in non-English content analysis. *arXiv preprint arXiv:2306.07377* (2023).

[21] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057* (2022).

[22] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

[23] Noah Wang, Feiyu Duan, Yibo Zhang, Wangchunshu Zhou, Ke Xu, Wenhao Huang, and Jie Fu. 2024. PositionID: LLMs can Control Lengths, Copy and Paste with Explicit Positional Awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 16877–16915. doi:10.18653/v1/2024.findings-emnlp.983

[24] Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities. arXiv:2411.04986 [cs.CL] https://arxiv.org/abs/2411.04986

[25] Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477* (2024).

[26] Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. Pushing the limit of LLM capacity for text classification. *arXiv preprint arXiv:2402.07470* (2024).

[27] Ziji Zhang, Michael Yang, Zhiyu Chen, Yingying Zhuang, Shu-Ting Pi, Qun Liu, Rajashekar Maragoud, Vy Nguyen, and Anurag Beniwal. 2025. REIC: RAG-Enhanced Intent Classification at Scale. arXiv:2506.00210 [cs.CL] https://arxiv.org/abs/2506.00210

[28] Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055* (2024).

[29] Yingying Zhuang, Aman Gupta, and Anurag Beniwal. 2025. Multilingual Information Retrieval with a Monolingual Knowledge Base. arXiv:2506.02527 [cs.CL] https://arxiv.org/abs/2506.02527

[30] Yingying Zhuang, Yichao Lu, and Simi Wang. 2021. Weakly Supervised Extractive Summarization with Attention. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li (Eds.). Association for Computational Linguistics, Singapore and Online, 520–529. doi:10.18653/v1/2021.sigdial-1.54

[31] Yingying Zhuang, Jiecheng Song, Narayanan Sadagopan, and Anurag Beniwal. 2023. Self-supervised Pre-training and Semi-supervised Learning for Extractive Dialog Summarization. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 1069–1076. doi:10.1145/3543873.3587680