Investigating Language and Retrieval Bias in Multilingual Previously Fact-Checked Claim Detection

Ivan Vykopal[†,1,2], **Antonia Karamolegkou**[†,3], **Jaroslav Kopčan**[2], **Qiwei Peng**[3], **Tomáš Javůrek**[2], **Michal Gregor**[2] and **Marián Šimko**[2]
[1]Brno University of Technology,
[2]Kempelen Institute of Intelligent Technologies,
[3]University of Copenhagen,
[†]Contributed equally

Abstract

Multilingual Large Language Models (LLMs) offer powerful capabilities for cross-lingual fact-checking. However, these models often exhibit *language bias*, performing disproportionately better on high-resource languages such as English than on low-resource counterparts. We also present and inspect a novel concept - *retrieval bias*, when information retrieval systems tend to favor certain information over others, leaving the retrieval process skewed. In this paper, we study language and retrieval bias in the context of Previously Fact-Checked Claim Detection (PFCD). We evaluate six open-source multilingual LLMs across 20 languages using a fully multilingual prompting strategy, leveraging the AMC-16K dataset. By translating task prompts into each language, we uncover disparities in monolingual and cross-lingual performance and identify key trends based on model family, size, and prompting strategy. Our findings highlight persistent bias in LLM behavior and offer recommendations for improving equity in multilingual fact-checking. To investigate retrieval bias, we employed multilingual embedding models and look into the frequency of retrieved claims. Our analysis reveals that certain claims are retrieved disproportionately across different posts, leading to inflated retrieval performance for popular claims while under-representing less common ones.

1 Introduction

Recent advances in multilingual Large Language Models (LLMs) have enabled powerful capabilities in multilingual and cross-lingual natural language understanding. Recent models support reasoning and generation across dozens or even hundreds of languages, powering applications ranging from machine translation to multilingual retrieval. Among these, fact-checking applications have benefited significantly: LLMs are increasingly used to identify claims, verify them, or even identify whether a given claim has already been verified (Chen and Shu, 2024; Vykopal et al., 2024).

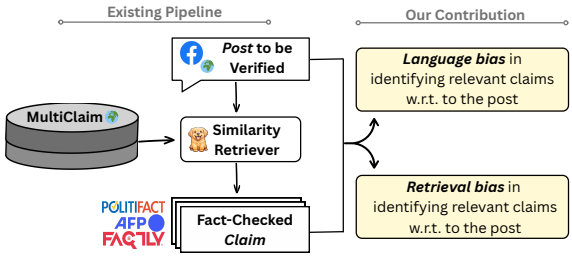

Figure 1: Our contribution and suggested analysis in the existing PFCD pipeline.

Despite these advances, multilingual LLMs exhibit persistent and often severe disparities in performance across different languages. This phenomenon, referred to as *language bias*, occurs when models systematically favor high-resource languages, such as English, Chinese, or Spanish-over underrepresented ones, leading to uneven model behavior and outcomes (Xu et al., 2025). Language bias is particularly problematic in fact-checking contexts, where accurate performance in low-resource languages is crucial for equitable access to trustworthy information. These disparities are rooted in the pretraining data, where high-resource languages dominate, and are amplified during model alignment and instruction tuning (Yang et al., 2024).

In addition to language bias, we focus on what we call *retrieval bias*. Whereas prior work on factuality in LLMs typically refers to the correctness of generated content, we define retrieval bias as the tendency of a model to retrieve certain fact-checked claims more frequently than others, independent of their actual relevance. In PFCD, this means the model might systematically favor or recall a subset of popular or easily phrased debunks, regardless of whether they truly match the input post. Retrieval bias can arise from differences in phrasing, claim popularity, or language-induced artifacts: for exam-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5195–5221
March 24-29, 2026

ple, claims that appear more often in training data or have generic wording may be retrieved more often. This phenomenon is distinct from generation accuracy and has not been explicitly characterized in previous work.

In this paper, we systematically investigate both language bias and retrieval bias in multilingual PFCD. We build upon the ***AMC-16K*** dataset (Vykopal et al., 2025), a subset of ***MultiClaim*** (Pikuliak et al., 2023b), which contains 16,000 annotated pairs of social media posts and corresponding fact-checked claims in 20 languages, covering both monolingual and cross-lingual scenarios. A key novelty of our study is the use of *fully multilingual prompting*: all prompts and task instructions are translated into the respective input languages, enabling direct comparison with English-only prompting strategies. Our contribution and the analysis are shown in Figure 1. We assess six state-of-the-art open-source LLMs (`Qwen3` $8B$ and $14B$, `Llama3.1` $8B$ and $70B$, `Gemma3` $12B$ and $27B$) on ***AMC-16K***. Our experiments use various prompting strategies (zero-shot, zero-shot with task description, few-shot, and chain-of-thought) to measure their robustness.[1]

The goal of this study is primarily diagnostic rather than algorithmic. We aim to systematically characterize and quantify language and retrieval bias in multilingual PFCD, rather than propose a new method. By providing a rigorous empirical evaluation across languages, models, and prompting strategies, and by formally defining retrieval bias as a distinct failure mode, we establish a foundation for measurement and analysis that can inform the development of future mitigation techniques. While we include an exploratory analysis of LLM-based re-ranking to illustrate how retrieval bias may be reduced, mitigation is not the primary focus of this paper.

We find that (1) most models exhibit significant language bias: performance often drops on non-English inputs and in languages less represented during training, consistent with Yang et al. (2024) and Huang et al. (2023), and (2) models also exhibit retrieval bias: certain claims are retrieved far more often, typically those with simpler phrasing or higher prevalence, indicating that the models are influenced by claim distribution as well as relevance. These findings highlight the need to account for both language and retrieval biases when developing and evaluating multilingual PFCD systems.

[1]Code and data are available at: https://github.com/kinit-sk/llms-biases.

2 Background

2.1 Fact-Checked Claim Detection

The task of retrieving previously fact-checked claims (PFCD) was introduced by Shaar et al. (2020), who developed the first benchmark and demonstrated its utility for reducing redundant verification efforts. Since then, research has expanded PFCD to multilingual settings with researchers providing new data sources (Kazemi et al., 2021) or organizing shared tasks with multilingual tracks (Nakov et al., 2022; Peng et al., 2025b). Recently, Pikuliak et al. (2023a) introduced MultiClaim, a large multilingual dataset covering 39 languages. Building upon the MultiClaim, Vykopal et al. (2024) released *AMC-16K*, a curated subset with 16K claim–post pairs across 20 languages. PFCD systems have evolved from embedding-based ranking to entailment-style models (Shaar et al., 2022) and, more recently, LLM-based generation or reranking (Zheng et al., 2024; Pisarevskaya and Zubiaga, 2025). However, most studies still focus on English or a few high-resource languages, leaving performance disparities largely unexplored.

2.2 Language Bias

Language bias in multilingual NLP refers to systematic disparities in model performance across languages. Xu et al. (2025) provide a comprehensive review of such disparities, noting that LLMs consistently favor high-resource languages due to imbalanced training corpora and optimization strategies. Empirical work by Yang et al. (2024) shows that retrieval success, classification, and summarization tasks all suffer from reduced performance in low-resource languages. Several mitigation strategies have been proposed, including balanced pretraining (Liu et al., 2020), multilingual alignment (Conneau et al., 2020), and synthetic data augmentation (Hedderich et al., 2021). Despite this, most multilingual models continue to perform best in English.

2.3 Retrieval Bias

We introduce the term retrieval bias to describe a tendency of claim-matching models to favor certain fact-checked claims over others, leading to a retrieval bias in PFCD. In our context, retrieval bias manifests when a model systematically retrieves

some claims (e.g., those with high semantic similarity, profile topics, or generic phrasing) more often. For example, if a model frequently matches unrelated posts to a well-known debunk about vaccine safety simply because that debunk appears often in data or is phrased generically, this indicates retrieval bias. Such biases can arise from the uneven distribution of claims in the training or retrieval corpus: popular or frequently paraphrased claims are more "accessible" to the model. They may also result from language-driven effects if certain claims translate more naturally into the prompt language.

To our knowledge, this specific bias has not been explicitly defined in PFCD research, so we formalize it here. We will analyze retrieval bias by measuring the frequency with which each claim is retrieved (or scored as relevant) by the model, controlling for actual ground-truth relevance. If some claims are chosen disproportionately often across different posts, this signals retrieval bias. Understanding this bias is important because it can inflate apparent performance if models "over-retrieve" popular claims, and it can also highlight blind spots for less common claims.

Retrieval bias is conceptually related to popularity bias in recommender systems (RS), where algorithms tend to favor already popular items (e.g., widely streamed songs or frequently purchased products) at the expense of long-tail items (Klimashevskaia et al., 2024). In both cases, system outputs become skewed toward certain content, inflating apparent performance while masking blind spots for rarer or more specific instances. Whereas RS research has developed formal metrics and mitigation strategies for popularity bias (Abdollahpouri et al., 2021), retrieval bias in PFCD has not been explicitly defined or systematically studied.

3 Language Bias

To systematically examine the language bias, we evaluate the behavior of multilingual LLMs on the task of detecting previously fact-checked claims. Our analysis considers both monolingual and cross-lingual scenarios across 20 languages, enabling us to explore how LLMs address linguistic diversity.

3.1 Methodology

Our experimental design builds upon the methodology introduced by Vykopal et al. (2025), who evaluated seven multilingual LLMs across five prompting strategies in both monolingual and cross-lingual settings. However, their experiments relied only on English-language instructions, regardless of the language used in posts or fact-checked claims. To address this limitation, we extend their work by translating prompt instructions into the 20 languages represented in the dataset (see Figure 2) and Chinese. This allows us to assess how multilingual prompting impacts model performance and to analyze language biases in LLM behavior.

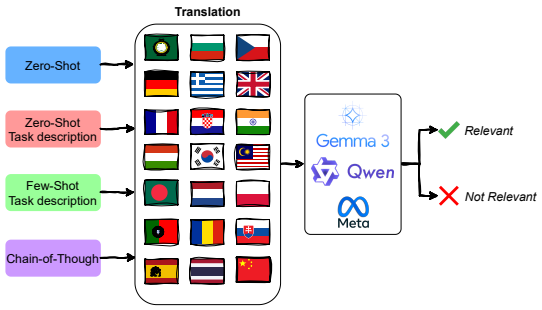

Figure 2: A pipeline for language bias experiments. First, we translate instructions into 21 languages and then prompt LLMs with translated instructions on the entire dataset using all four prompting strategies.

Dataset. We employed the ***AMC-16K*** dataset (Vykopal et al., 2025) as the basis for our experiments. The dataset is a subset of the ***MultiClaim*** dataset (Pikuliak et al., 2023a), consisting of 16,000 pairs of social media posts and corresponding fact-checked claims. These pairs are divided into two settings, monolingual and cross-lingual. In the *monolingual setting*, the post and the fact-checked claim are in the same language, with 8K samples spanning 20 languages. In the *cross-lingual setting*, the post and the fact-checked claim are in different languages, comprising another 8K pairs that cover 20 diverse language combinations.

Large Language Models. We selected six multilingual LLMs from three model families: `Qwen3`, `Llama3.1` and `Gemma3`. Each model family is represented by two variants with a different model size. All selected models are open-source, similarly to Vykopal et al. (2025), to ensure the reproducibility of our experiments. Table 1 provides details on the number of parameters, supported languages, and whether we used "*thinking*" mode for a specific model. The "*thinking*" mode represents a reasoning-enhanced inference that is offered by `Qwen3` models. However, for our experiments, we leveraged the "*thinking*" mode, only with the small-

Model	# Params	# Langs	Thinking	Citation
Qwen3	8B	100	✓	Yang et al. (2025)
	14B		✗	
Llama3.1 Instruct	8B	8	✗	Grattafiori et al. (2024)
	70B		✗	
Gemma3 IT	12B	130	✗	Team et al. (2025)
	27B		✗	

Table 1: Overview of multilingual LLMs used in our experiments.

est Qwen3 model, due to computational reasons. By experimenting with "*thinking*" mode, we aim to assess whether enabling internal reasoning has an effect on the performance and the bias in LLMs.

Prompting Strategies. To examine how different prompting strategies affect LLMs' performance across languages, we adopted techniques introduced by Vykopal et al. (2025). We explore four main approaches: *(1) zero-shot*; *(2) zero-shot with task description*; *(3) few-shot with task description*; and *(4) chain-of-thought*. In *zero-shot prompting*, the LLM receives only a post and a fact-checked claim without any task explanation, with a simple question to determine the relevance. *Zero-Shot with task description* adds a brief instruction to clarify and describe the task. *Few-shot with task descriptions* supplements this with 10 labeled demonstrations to guide the LLM. Finally, *chain-of-though (CoT)* prompting encourages step-by-step reasoning before producing a decision.

We excluded the *crosslingual-thought prompting (XLT)* (Huang et al., 2023) strategy used in the original study, as it instructs LLMs to translate both the post and the fact-checked claim into English. Our aim is to assess the model's multilingual abilities without translation as an intermediate step, which could mask language-specific biases.

To support fair multilingual evaluation, we translated all prompt templates and task descriptions into 19 non-English languages present in the ***AMC-16K*** dataset and Chinese, resulting in instructions across 21 languages, while the data cover 20. The translations were generated using the GPT-4.1 API[2] and then manually reviewed by native or proficient speakers to ensure syntactic accuracy and semantic fidelity.

Evaluation. To evaluate the capabilities of LLMs for detecting previously fact-checked claims, each LLM was instructed to produce a binary label, determining the relevance between social media posts and fact-checked claims. Given the class imbalance in the dataset, where only 16% of the pairs are relevant, we use the *Macro F1* as the main evaluation metric, as it balances performance across both relevant and irrelevant classes.

To further analyze LLM's behavior, similarly to Vykopal et al. (2025), we also report the *True Negative Rate (TNR)*, which captures how well the LLM identifies irrelevant pairs, and the *False Negative Rate (FNR)*, which reflects the frequency of missed relevant pairs.

3.2 Experiments and Results

In this section, we present findings on language bias in LLMs for the task of detecting previously fact-checked claims. We evaluated model performance in two experimental setups: *monolingual* and *cross-lingual*. For each setting, we report and discuss key observations based on the results.

Monolingual Settings. In Table 2, *Monolingual* column shows the average performance differences when using target-language instruction and English instructions, averaged across 20 languages for each model and prompting strategies.

Gemma3 27B consistently preferred target-language instructions across all prompting strategies. This LLM showed positive performance differences in all settings when the instruction was provided in the target languages, suggesting its strong multilingual capabilities.

Prompting strategies that include task description or few-shot examples reduce language bias. Across most LLMs, using target-language instructions in zero-shot settings resulted in strongly negative performance differences, indicating a bias towards English. However, adding task-specific context narrowed these differences, moving performance closer to that of the English baseline, but still achieving a positive increase towards target-language instructions. This reveals that enhanced prompting helps mitigate language-specific performance gaps and reduces bias.

Qwen showed strong language bias in zero-shot and CoT prompting. Both model sizes had large negative differences when using the target language for instructions in zero-shot and CoT settings, indicating a bias toward English instructions. These large deviations suggest that the Qwen is less effective at processing multilingual prompts unless additional context or demonstrations are provided.

[2]https://openai.com/api/

	Monolingual					Cross-lingual - post-language					Cross-lingual - claim-language				
Model	**ZS**	**ZS + Task**	**FS + Task**	**CoT**	**Avg.**	**ZS**	**ZS + Task**	**FS + Task**	**CoT**	**Avg.**	**ZS**	**ZS + Task**	**FS + Task**	**CoT**	**Avg.**
Qwen3 8B	69.70	77.81	70.70	83.07	75.32	60.25	70.40	63.84	76.65	67.79	60.25	70.40	63.84	76.65	67.79
	-6.01	-1.50	0.76	-6.31	-3.26	-4.97	-1.75	0.51	-8.52	-3.68	-1.42	-2.40	0.98	-4.28	-1.78
Qwen3 14B	73.40	77.08	74.72	**85.15**	77.59	64.12	70.99	69.29	**78.24**	70.66	64.12	70.99	69.29	**78.24**	70.66
	-9.23	3.19	0.39	-4.82	-2.62	-7.67	-0.30	0.58	-5.69	-3.27	-4.94	0.81	-0.38	-3.11	-1.91
Gemma3 12B	56.44	58.72	70.07	58.90	61.03	52.86	52.22	65.92	50.78	55.45	52.86	52.22	65.92	50.78	55.45
	-3.67	6.36	0.26	1.99	1.23	-4.91	3.18	0.90	0.42	-0.10	-2.08	2.60	-1.31	0.85	0.02
Gemma3 27B	59.35	55.54	72.72	52.41	60.00	53.69	49.54	64.94	46.37	53.64	53.69	49.54	64.94	46.37	53.64
	1.64	10.85	0.83	5.60	4.73	-3.41	3.61	-1.12	3.24	0.58	0.50	4.16	0.83	2.09	1.90
Llama3.1 8B	50.46	71.13	51.12	67.18	59.97	45.53	66.48	50.56	58.82	55.35	45.53	66.48	50.56	58.82	55.35
	-1.55	-7.20	4.41	-2.34	-1.67	-2.65	-5.72	3.97	-3.98	-2.10	0.12	-1.31	1.92	1.05	0.45
Llama3.1 70B	73.97	74.38	54.04	75.48	69.47	65.57	65.77	54.34	67.66	63.33	65.57	65.77	54.34	67.66	63.33
	-5.18	0.48	0.67	-1.98	-1.50	0.45	3.00	-2.82	-1.88	-0.31	-1.42	1.46	-0.74	-2.37	-0.77
Average	63.89	69.11	65.56	70.67	67.23	57.00	62.57	61.48	63.09	61.04	57.00	62.57	61.48	63.09	61.04
	-4.00	2.03	1.22	-1.31	-0.51	-3.86	0.34	0.34	-2.74	-1.48	-1.54	0.89	0.22	-0.96	-0.35
Qwen3 8B (thinking)	**77.19**	**82.74**	**81.27**	-	**80.40**	**66.29**	**74.43**	**73.22**	-	**71.31**	**66.29**	**74.43**	**73.22**	-	**71.31**
	-2.50	-1.25	-0.96	-	-0.93	-2.28	-1.45	-0.15	-	-1.29	-1.09	-2.08	0.21	-	-0.99

Table 2: Macro F1 performance across 20 languages and language combinations for multiple models in three settings: *monolingual*, *cross-lingual with post-language instructions*, and *cross-lingual with claim-language instructions*. The first row per model shows absolute performance with English instructions; the second row shows the difference when using the target language for instruction. ZS = zero-shot; FS = few-shot; Task = using Task description. Green cells indicate improvements; red cells indicate declines. The highest absolute score for each column is in **bold**.

Cross-Lingual Settings. In the cross-lingual settings (see Table 2, cross-lingual columns), we compared the post-language and claim-language instruction against the English-language instruction. For post-language instruction, we investigate the difference between using the language of the post for the instruction compared to English instruction, while in claim-language instruction, we are using the language of the fact-checked claim.

For the post-language setup, we found that few-shot prompting combined with a task description reduced language bias and led to positive performance gains across most LLMs. This suggests that instruction clarity, especially when combined with the same language as the input post, can improve model performance and steer the language bias. In addition, Gemma3 models showed strong robustness to instruction language, with consistently improved performance when using the post language in all strategies except zero-shot.

In the claim-language instruction setting, we observed that zero-shot prompting generally led to a negative impact on performance, indicating a bias toward English when no additional context was provided. While adding a task description or a few-shot examples helped mitigate this effect, the resulting improvements were maller than those observed in the post-language setup. These findings suggest that, compared to post-language prompts, using the claim language for instructions offers less benefit overall, particularly in low-context settings.

Thinking. In our experiments, we also considered "*thinking*" mode for the Qwen3 8B model and evaluated whether thinking affects language bias. We found that enabling the thinking mode in Qwen3 8B did not improve performance when using target-language instructions in a monolingual setting. In general, the thinking mode benefited the English-language instructions, suggesting a potential bias in how reasoning capabilities are tuned across languages. However, in a zero-shot setting, the average performance decrease was less severe when thinking was enabled (-6 without thinking vs. -2.5 with thinking), indicating a partial mitigation of the negative effect.

For the cross-lingual, we observed that the thinking mode decreased the language bias in LLMs (average performance was decreased by around 1.8 and 2.9 times). When including the task description along with the demonstrations, the mitigation of the language bias was the highest, resulting only in marginal differences in Macro F1. Overall, these results demonstrate that activating thinking reduces the performance gap between English and target-language instructions, highlighting its potential for mitigating language bias.

4 Retrieval Bias

We define retrieval bias as the model's tendency to retrieve certain fact-checked claims more frequently than others, regardless of their actual relevance to the post. To examine retrieval bias, we

use both quantitative and qualitative methods. First measure retrieval success using standard metrics. Then, we analyze how frequently different claims are retrieved in the top-K results to identify retrieval biases. Finally, we perform topic modeling to explore the main themes of these frequently retrieved claims.

4.1 Methodology

We follow (Pikuliak et al., 2023b) and frame the retrieval task as a semantic matching and ranking problem, where the utilized embedding model must rank all available claims for each given post. We base our experiments on the test split of the ***MultiClaim*** dataset, and employ `T5` (Ni et al., 2022) and `Multilingual E5` (Wang et al., 2024) as embedding models. These models serve as our retrieval system by encoding both posts (queries) and claims (candidates) into dense vector representations, which are then ranked using cosine similarity scores. The split consists of 1,239 post-claim pairs. For each post, we compute claim embeddings, calculate cosine similarities with the post embedding, and rank by similarity to retrieve the top-K candidates, as depicted in Figure 3. To ensure computational efficiency while maintaining comprehensive coverage, we have limited the overall retrieval only to the top 20 claims for each post text. In particular, we analyze two key aspects: first, the retrieval success (both quantitatively and qualitatively) and second, the retrieval topic distribution.

Retrieval Success. We begin by evaluating retrieval performance using *Success@K*, *Mean Average Precision (MAP)*, and *Mean Reciprocal Rank (MRR)*. *Success@K* measures the proportion of posts for which at least one gold-relevant claim appears within the top-K retrieved candidates. *MAP* captures the average precision across all posts, emphasizing the rank positions of relevant claims and providing a holistic view of ranking quality. *MRR* focuses on the position of the first correct (gold-relevant) claim by averaging the reciprocal of its rank. These metrics are computed using the gold claim-post mapping, following the evaluation protocol of Pikuliak et al. (2023b). To further analyze retrieval quality and biases, we then examine the distribution of retrieved claims across posts to identify whether certain fact-checked claims are retrieved disproportionately often, regardless of their relevance. Finally, to better understand thematic patterns in these frequently retrieved claims, we apply topic modeling techniques, which reveal dominant topics and potential biases influencing retrieval behavior.

Topic distribution. To understand the thematic distribution of the retrieved claims and identify potential retrieval biases, we apply topic modeling on the claim contents. Specifically, we use *Semantic Signal Separation (S^3)* via the TurfTopic modeling framework (Kardos et al., 2025), which leverages semantic embeddings rather than raw text. This allows for more nuanced topic discovery in multilingual and noisy claim data. We first encode the retrieved claim texts using the `paraphrase-multilingual-MiniLM-L12-v2` model from the SentenceTransformers library (Reimers and Gurevych, 2019). This model generates dense sentence embeddings capturing semantic similarity across languages and contexts. The embeddings are then input to the Semantic Signal Separation model, which decomposes the embedding space into a predefined number of topics (in our case, 30). The model is fit using a fit-transform procedure that returns topic assignment probabilities for each claim. We assign each claim to the topic with the highest probability, producing a topic distribution that reflects the major semantic themes in the retrieved data. This topic distribution enables qualitative analysis of frequent claim types and highlights potential thematic retrieval biases in the system.

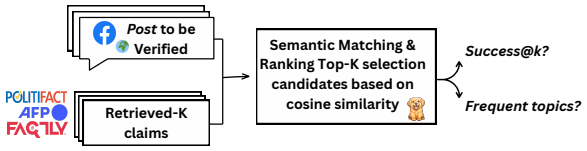

Figure 3: The retriever takes a post as a query, searches through a database of previously fact-checked claims to identify and rank the most relevant claims based on cosine similarity, returning top-K candidates. We then calculate the success of the retrieval and potentially frequently retrieved claims.

4.2 Experiments and Results

In this section, we present findings from our retrieval bias analysis. Building on the methodology described above, we evaluate retrieval behavior across two dimensions.

Retrieval Success. The results of retrieval success across the different metrics are presented in Table 3. The `Multilingual E5` model slightly outperforms `T5` at Success@1 (38.8% vs. 36.7%) and

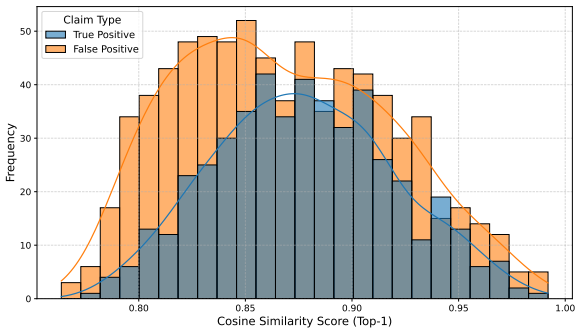

Figure 4: Distribution of Top-1 cosine similarity scores for the `Multilingual E5` retriever, comparing true positives (TP), i.e., claims correctly ranked at the top position, and false positives (FP), i.e., claims incorrectly ranked first.

Success@20 (88.5% vs. 87.5%), with marginal gains in MAP and MRR as well. The scores at Success@10 are consistent with previous works that use a subset of the ***MultiClaim*** dataset (Pikuliak et al., 2023b; Peng et al., 2025b). This indicates that both models achieve comparable retrieval quality, with `Multilingual E5` providing a modest advantage in early retrieval effectiveness.

Model	@1	@3	@5	@10	@20	MAP / MRR
`T5`	36.72%	64.41%	74.01%	83.21%	87.49%	51.23 / 52.43
`Multilingual E5`	38.80%	63.40%	74.40%	82.90%	88.50%	52.69 / 53.98

Table 3: Success@K and Mean Average Precision (MAP)/Mean Reciprocal Rank (MRR) for dense retrieval models on ***MultiClaim***.

To investigate the low success@1 scores, we analyze the cosine similarity distributions of top-1 true positive (TP) and false positive (FP) retrieved claims. Figure 4 shows a substantial overlap between the two distributions, indicating that false positives often achieve similarity scores comparable to correct matches. This highlights a fundamental limitation of similarity-based retrieval for PFCD.

This analysis provides direct quantitative evidence that dense retrieval models can be systematically misled by surface-level semantic or topical similarity, assigning high confidence to claims that are not factually aligned with the input post. As a result, retrieval errors persist even when similarity-based metrics suggest a strong match, highlighting a limitation of embedding-only retrieval approaches that is not captured by standard success or ranking metrics.

Apart from quantitative analysis, we also conduct a qualitative inspection of the most frequently retrieved top-1 claims by the `Multilingual E5` model. Table 4 presents the ten claims that most often appeared in the top-1 retrieval position across posts. While the majority are well-formed factual statements, typically addressing political events or health-related misinformation, we also observe two noisy or low-information entries: a truncated JavaScript snippet and a Google Drive link. These cases highlight that, despite generally favoring semantically rich and interpretable claims at rank 1, the model can still surface non-claim-like artifacts. The presence of such items suggests potential vulnerabilities in embedding-based retrieval when encountering malformed inputs. For comparison, we report the top-20 most frequent retrieved claims (regardless of rank) in the Appendix in Table 10, where such noisy entries are even more prominent.

This behavior reflects a known limitation (i.e. vocabulary and semantic mismatch) of dense retrieval models, which rely on distributional semantic similarity and are not designed to distinguish well-formed factual claims from structurally malformed inputs (Karpukhin et al., 2020). As a result, common retrieval practice typically involves discarding or filtering such noisy entries prior to retrieval. Our findings highlight the importance of such preprocessing steps when deploying dense retrievers in large, heterogeneous claim corpora.

Claim (truncated)	Freq	CosSim
(function(d, s, id) { var js, fjs...	12	0.811
New Australian law passed in Parliament...	6	0.904
Trump tweets on South Africa's...	5	0.901
https://drive.google.com/file/d/1TF...	5	0.799
South Korean protesters bombed with...	5	0.909
A message is being widely shared on...	5	0.898
Report quotes Duterte saying COVID-19...	5	0.842
Accurate representation of vaccinations...	5	0.876
Alexandre Trudeau, the brother of...	4	0.946
Pictures show Azerbaijan has destroyed...	4	0.910

Table 4: Top-1 retrieved claims: truncated content with frequency (freq) and average cosine similarity score (CosSim).

Lastly, as an attempt to improve retrieval success and the success@1 scores, we explore using LLMs to classify each claim–post pair as *relevant* or *not* in the zero-shot setting. As LLMs do not return a ranked list but only binary predictions, we report **Top-1 relevance classification accuracy**-i.e., the fraction of posts for which the top-scoring LLM-predicted relevant claim aligns with the gold label. Results are shown in Table 5.

Incorporating LLM-based relevance filtering im-

Model	Top-1 Accuracy
T5	36.72%
Multilingual E5	38.80%
Qwen3 8B	63.12%
Gemma3 12B	42.80%
Llama3.1 8B	38.86%

Table 5: Top-1 relevance classification accuracy (or success@1) for LLMs and retrieval models, showing that LLMs can help improve the success of the retrieval.

proves these scores, demonstrating the model's ability to mitigate retrieval bias by deprioritizing frequent but irrelevant claims.

Topic Modeling. Table 6 summarizes the normalized distribution of major thematic groups across posts, retrieved claims, and top-1 retrieval outcomes. By directly comparing topic frequencies in the input posts and retrieved claims, the results indicate a topic skew in the retrieval process, with certain high-profile themes retrieved far more frequently than their prevalence in the posts would suggest. For example, COVID-19–related content constitutes only 23.0% of posts but accounts for nearly half of all retrieved claims (49.5%) and an even larger share of false positive top-1 claims (44.8%). This discrepancy may suggest that dense retrieval disproportionately favors dominant, high-frequency topics beyond what would be expected from the post distribution alone.

Conversely, themes that are prominent in posts, such as domestic politics (22.4%), are almost entirely absent from both true and false positive top-1 retrievals. Together, these patterns demonstrate that retrieval errors could amplify corpus-level topic skew. We further illustrate the semantic structure underlying this behavior in the Appendix using a t-SNE projection of claim embeddings (Figure 7), alongside full top-10 frequency lists for all analyzed subsets.

5 Discussion

Our analysis highlights two central potential sources of bias in PFCD: *language bias* and *retrieval bias*. These biases affect model performance in distinct but interconnected ways and have compounding effects on the fairness and effectiveness of multilingual fact-checking systems. Importantly, these biases do not arise in isolation. PFCD benchmarks may reflect biases, overrepresenting globally salient topics and high-resource languages, thereby amplifying both retrieval and language bias. Based

Thematic Group	Posts %	Claims %	TP%	FP%
COVID/Vaccines	23.0%	49.5%	23.2%	**44.8%**
Visual/Photos	0.0%*	22.5%	14.7%	12.8%
Nigeria/Africa	0.0%*	10.3%	**22.5%**	7.7%
Ukraine/War	9.8%	9.0%	0.0%*	**10.7%**
Politics	**22.4%**	8.7%	0.0%*	0.0%*

Table 6: Distribution of major thematic groups across posts used in our analysis ($n = 1{,}239$), retrieved claims ($n = 13{,}861$), and top-1 retrieved claims split into true positives (TP, $n = 481$) and false positives (FP, $n = 758$). Percentages indicate the share of each theme within the corresponding subset. Asterisks (*) denote themes that do not appear among the top-10 most frequent topics for that subset; full top-10 topic lists are provided in the Appendix.

on our experiments and analysis, we argue that in addition to aggregate performance scores, future evaluations could benefit from reporting performance stratified by language, topic coverage, and frequency-normalized success.

Language Bias and Its Impact. Language bias emerges as systematic performance disparities across languages, with models consistently favoring high-resource languages such as English. Despite using multilingual prompting as suggested by previous works, we observe that models like Qwen3 and Llama3.1 perform significantly worse on low-resource languages, particularly under zero-shot settings. We find that language bias can be partially mitigated through enhanced prompting strategies. Few-shot prompting and task descriptions significantly reduce performance gaps between English and target-language instructions. Moreover, enabling reasoning-enhanced inference ("thinking mode") further narrows this gap, particularly in monolingual settings. This suggests that both contextual clarity and internal reasoning mechanisms help models better generalize across languages.

The implications of language bias are profound: users in underrepresented linguistic communities are less likely to benefit from accurate fact-checking, exacerbating global information inequities. Moreover, the performance gap persists even when semantic equivalence is maintained across translations, suggesting that model alignment and pretraining corpora remain skewed toward dominant languages.

Retrieval Bias and Its Impact. Retrieval bias refers to the model's tendency to favor certain claims during retrieval, independent of their actual relevance. Our analysis shows that some claims-

especially those with generic phrasing and high topical prevalence are disproportionately retrieved. This is evident in the topic modeling results, where high-frequency topics (e.g., COVID-19, war) dominate the top retrieved claims, indicating a systematic topic skew toward globally salient narratives. We quantify retrieval bias using *Success@K*, measuring the likelihood that a relevant claim appears among the top-K retrieved results. Without LLM reranking, the success@1 is only 38.8%, improving to 88.5% at $K = 20$. However, when we incorporate LLM-based relevance filtering, success@1 improves, indicating that LLMs can help mitigate retrieval bias by reranking based on semantic relevance rather than frequency or phrasing alone. This bias can lead to reduced topical coverage and overlooked relevant claims, especially for underrepresented narratives. As a result, the effectiveness of fact-check retrieval systems may be compromised in diverse real-world settings.

This retrieval bias is further amplified by noise in the retrieval corpus, where structurally malformed or low-information entries (e.g., URLs or truncated snippets) are repeatedly retrieved due to superficial similarity rather than factual content. Together, these findings suggest that retrieval bias in multilingual PFCD cannot be fully understood without accounting for corpus noise and topic skew, the latter of which may partially originate from the underlying fact-checking corpus itself (e.g., MultiClaim). We believe that mitigation likely requires complementary strategies beyond embedding similarity, such as lightweight claim-level filtering or structural validation of retrieved candidates.

Interplay and Broader Implications. Language and retrieval biases are not isolated phenomena. Language bias can amplify retrieval bias when certain claims are more accessible or better represented in specific languages. Conversely, retrieval bias can mask language bias by over-representing high-frequency claims that are easier to match across languages. Together, these biases challenge the equity and robustness of multilingual PFCD systems. Addressing them requires not only architectural and training improvements but also evaluation frameworks that explicitly account for both semantic relevance and linguistic diversity. Addressing them requires not only architectural and training improvements, but also evaluation frameworks that explicitly account for topic concentration and language resource imbalance (see Appendix D for a detailed analysis about the interaction between language bias and retrieval bias focused on topic concentration).

6 Conclusion

This study investigates language and retrieval biases in multilingual LLMs when applied to the task of PCFD. Through a systematic evaluation across 20 languages and six models, we show that most models perform best in English and suffer measurable drops in accuracy when operating in low-resource languages. Our results also suggest that the prompt language and prompting strategy influence performance, with few-shot prompting and translated task descriptions helping mitigate some of the bias. We also identify a distinct form of retrieval bias, where certain claims-often generic, high-profile, or frequently occurring-are retrieved disproportionately. This can distort evaluation metrics and limit the diversity of retrieved evidence. However, we find that LLM-based reranking improves semantic alignment and helps mitigate this effect.

We hope this analysis provides a foundation for more equitable multilingual fact-checking systems. Future directions could explore deeper linguistic analysis of failures, adaptive prompting strategies, and fairness-aware retrieval objectives. Beyond technical solutions, interdisciplinary approaches-such as sociolinguistic audits of retrieved claims-may offer new lenses on model behavior, revealing how cultural framing and linguistic nuance shape retrieval outcomes.

Limitations

Language Coverage. Despite covering 20 languages, our study remains limited by the scope of the ***AMC-16K*** dataset and the selection of language pairs. Some regions and scripts (e.g., indigenous languages, South Asian scripts beyond Devanagari) are underrepresented. Additionally, our evaluation assumes task equivalence across languages, which may not fully capture cultural or linguistic nuances.

Model Selection and Prompting Strategies. We also only test six open-source models, which may not reflect the capabilities of proprietary LLMs. Finally, while we apply multilingual prompting, we do not consider code-switching or mixed-language input, which may be more common in real-world scenarios.

Lack of Mitigation Methods. A key limitation of our study is that we do not propose a new method to mitigate language or retrieval bias. Our focus lies on analyzing, defining, and measuring these biases for the multilingual PFCD task. While we explore LLM-based re-ranking as an illustrative mitigation strategy, this analysis is not intended as a comprehensive solution. Developing robust and scalable mitigation techniques, particularly for low-resource languages, remains an important direction for future research, and we view our findings as a diagnostic foundation upon which such methods can be built.

Acknowledgments

This research was partially supported by *DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies*, a project funded by Horizon Europe under GA No.101079164, by the *European Union NextGenerationEU* through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00007, and by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

Antonia Karamolegkou was supported by the Onassis Foundation - Scholarship ID: F ZP 017-2/2022-2023'.

Ethical Consideration

Intended Use. All experiments in this work were conducted using publicly available, research-focused datasets: ***MultiClaim*** (Pikuliak et al., 2023b) and its subset ***AMC-16K*** (Vykopal et al., 2025). These datasets consist of social media posts and fact-checking information collected from public sources. Since both datasets are focused on the fact-checking domain, social media posts and fact-checked claims include harmful content. Our goal is to promote fairness and transparency in multilingual model evaluation, particularly for low-resource languages that are historically underrepresented in NLP systems.

Usage of AI Assistants. We have used the AI assistant for grammar checks and sentence structure improvements. We have not used AI assistants in the research process beyond the experiments detailed in the Methodology in Section 3 and 4.

References

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 119–129, New York, NY, USA. Association for Computing Machinery.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 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*, pages 2545–2568, Online. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 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*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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*, pages 6282–6293, Online. Association for Computational Linguistics.

Márton Kardos, Kenneth C. Enevoldsen, Jan Kostkan, Ross Deans Kristensen-McLachlan, and Roberta Rocca. 2025. Turftopic: Topic modelling with contextual representations from sentence transformers. *Journal of Open Source Software*, 10(111):8183.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A. Hale. 2021. Claim matching beyond English to scale global fact-checking. 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)*, pages 4504–4517, Online. Association for Computational Linguistics.

Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34(5):1777–1834.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52, Marseille, France. European Language Resources association.

Preslav Nakov and 1 others. 2022. Overview of the CLEF-2022 checkthat! lab. In *Working Notes of CLEF 2022-Conference and Labs of the Evaluation Forum*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podrouzek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025a. SemEval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2498–2511, Vienna, Austria. Association for Computational Linguistics.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podroužek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025b. SemEval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023a. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, Singapore. Association for Computational Linguistics.

Matúš Pikuliak and 1 others. 2023b. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Dina Pisarevskaya and Arkaitz Zubiaga. 2025. Zero-shot and few-shot learning with instruction-following LLMs for claim matching in automated fact-checking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9721–9736, Abu Dhabi, UAE. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Shaden Shaar, Nikola Georgiev, Firoj Alam, Giovanni Da San Martino, Aisha Mohamed, and Preslav Nakov. 2022. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, Tatiana Anikina, Michal Gregor, and Marian Simko. 2025. Large language models for multilingual previously fact-checked claim detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15741–15765, Suzhou, China. Association for Computational Linguistics.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative Large Language Models in Automated Fact-Checking: A Survey. *Preprint*, arXiv:2407.02351.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: corpora, alignment, and bias. *Front. Comput. Sci.*, 19(11).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2024. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 280–292, Miami, Florida, USA. Association for Computational Linguistics.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9274–9281, Bangkok, Thailand. Association for Computational Linguistics.

A Computational Resources

For our experiments, we leveraged a computational infrastructure consisting of A40 PCIe 40GB and H100 NVL 94GB NVIDIA GPUs, while our experiments ran in parallel on multiple GPUs. In total, our experiments required approximately 3250 GPU hours.

B Language Bias

In this section, we provide additional findings and results on the identification of language biases in LLMs. We extend the prompting setup by (Vykopal et al., 2025) and translate the prompts into different languages as described in the main paper. We provide the system prompt and zeroshot prompt template used in our experiments in Figures 5 below.

B.1 Experimental Setup

For all experiments for the language biases, we used the same hyperparameter settings during inference. In particular, we set the temperature to 0 to ensure mostly deterministic outputs and to minimize variability due to sampling. However, for the experiments with thinking mode with the `Qwen3 8B`, we adopted the recommended configuration provided by the model's authors. Specifically, we set the temperature to 0.6, TopP to 0.95, TopK to 20, and MinP to 0.

B.2 Monolingual Results

The comparison of absolute performance differences across the three combinations: monolingual, cross-lingual with the instruction in the post language and cross-lingual with the instruction in the claim language is shown in Table 11. The full results comparing the use of target language and English instructions across all models and prompting strategies in monolingual settings are presented in Table 13.

Overall Trends. In the monolingual setting, we observed several patterns that emerged regarding the effect of using target language instructions versus English. While the overall differences were modest, a slightly greater number of model-prompting combinations performed better with target language instructions, suggesting **a subtle positive bias toward using the target language**. However, zero-shot prompting consistently led to poorer results when instructions were not in English (except `Gemma3 27B`), highlighting the dominance of English in the zero-shot setting without additional context.

Language Families. *Germanic* languages, such as Dutch and German, tended to benefit from the target language instructions, with average improvements of around 4% in macro F1. Similarly, *Slavic* languages, such as Slovak, Czech, Polish, Serbo-Croatian, and Bulgarian, showed gains of approximately 2.5-3.5% on average. In contrast, *Sino-Tibetan* and other language families showed uniformly negative results with target language instructions, with some drops reaching up to 14% for Burmese, pointing to a pronounced bias toward English for really low-resource languages.

Writing Script. Script also played a significant role. Languages written in Latin script were more likely to benefit from target language instructions, whereas those using non-Latin scripts tended to perform worse, particularly under zero-shot and CoT prompting. There were exceptions, such as Arabic, Bulgarian, and Serbo-Croatian, despite using non-Latin scripts, that still demonstrated performance improvements across multiple settings and models. In addition, larger models generally

```
You are a fact-checker responsible for determining the relevance of previously
debunked claims to a given social media post. Your task is to assess if the claim
is relevant to the post and whether it is possible to infer main statements from
the post from the debunked claim. Based on your analysis, provide one of the
following answers:
- 'Yes' if the claim is relevant to the social media post.
- 'No' if the debunked claim is not relevant to the social media post.
```

```
Claim: {document}
Post: {query}

Is the claim relevant to the social media post? Respond with a single word,
either ""Yes"" or ""No"", in English only.
```

Figure 5: System prompt used for the zero-shot English setting. This prompt was translated into different languages using the GPT-4.1 API as detailed in the main paper.

demonstrated greater gains from target language instructions, indicating that scale may help mitigate English-centric bias. These trends likely reflect the pertaining data distribution, which is heavily skewed toward Latin-script and English-language content.

B.3 Cross-lingual Results

Post-Language Instructions. Detailed results for the cross-lingual setting, comparing post-language and English-language instructions, are provided in Table 15.

We identified several findings when the instructions were given in the language of the post. Zero-shot prompting generally led to weaker performance compared to English-language instruction, reinforcing the strong default bias toward English in multilingual LLMs. However, the `Qwen3` models, for instance, demonstrated improvements with post-language instructions, but only when few-shot demonstrations were provided, suggesting that these models benefit more from additional context when processing non-English instructions. On the other hand, `Gemma3 12B` achieved higher performance with the post-language instruction in most prompting strategies (except zero-shot), reducing the bias between post-language and English-language instructions.

Additionally, language similarity also played a role. Pairs of linguistically related languages, such as Slovak and Czech and Polish and Serbo-Croatian, often achieved better performance with post-language instructions. This indicates that shared linguistic structures and vocabulary may help models better align the instruction with the input content in such cases.

Claim-Language Instructions. Results for the cross-lingual setting using claim-language instructions, where the instruction is given in the language of the claim (i.e., the second language in each pair), are presented in Table 16.

Chinese vs English. To investigate language bias between *Chinese* and *English*, we compared model performance when instructions were provided in *Chinese* versus *English*. As shown in Table 7, most models exhibited lower performance when prompted in Chinese, particularly under zero-shot

Model	ZS	ZS + Task	FS + Task	CoT	Avg.
Qwen3 8B	-18.40	-9.41	2.44	-3.63	-7.25
Qwen3 14B	-20.46	-3.44	2.93	-4.37	-6.34
Gemma3 12B	-11.07	-5.44	2.99	-2.25	-3.94
Gemma3 27B	-2.11	5.18	3.91	4.00	2.75
Llama3.1 8B	-12.02	-25.06	-2.31	-8.60	-12.00
Llama3.1 70B	-11.22	0.26	-3.86	-8.02	-1.7

Table 7: Macro F1 score differences when using *Chinese* vs. *English* language for instructions across prompting strategies, evaluated on all annotated pairs (16K). Negative values (red) indicate lower performance with Chinese prompts; positive values (green) indicate improvements.

Lang	Rep. (%)	Unf. (%)	Lang	Rep. (%)	Unf. (%)
Arabic	0.15	0.55	Korean	0.30	2.26
Bulgarian	0.31	1.37	Malay	0.20	2.94
Czech	0.27	0.35	Burmese	**1.23**	**9.31**
German	0.08	1.13	Dutch	0.08	1.22
Greek	0.18	0.57	Polish	0.15	0.88
English	0.07	2.90	Portuguese	0.11	1.22
French	0.07	0.70	Romanian	0.15	0.84
Serbo-Croatian	0.34	0.61	Slovak	0.38	0.99
Hindi	0.31	1.45	Spanish	0.18	0.46
Hungarian	0.31	1.04	Thai	0.22	1.49

Table 8: Percentage of generation failures per instruction language, aggregated across all 27 model–prompting configurations. Rep. denotes repetitive outputs; Unf. denotes unfinished predictions.

(ZS) and zero-shot with task description (ZS +Task) settings. This suggests that English prompts remain more effective for guiding model behavior, mostly due to the model's pre-training data and instruction tuning being predominantly in English. Only the Gemma3 27B demonstrated improvements with Chinese prompts across 3 prompting strategies, resulting in an average improvement of 2.75. This highlights that expanding and diversifying the pertaining data to include more languages may lead to a trade-off: while it can improve performance in underrepresented languages, it may slightly degrade performance in English due to a more distributed linguistic focus.

B.4 Thinking

To further investigate the role of reasoning processes in instruction understanding, we conducted an ablation study on the impact of the thinking mode across different prompting strategies. In the monolingual setting, the results are shown in Table 18, where we compare performance with and without "thinking mode" across 20 languages. To extend this analysis to multilingual scenarios, we also evaluated the effect of the thinking mode in cross-lingual settings. Table 20 presents the results when using post-language instructions (the first language in each pair), while Table 21 shows the same analysis using claim-language instructions (the second language in each pair).

B.5 Analysis of Failures

Failure Cases and Generation Consistency. To better understand the source of performance degradation in low-resource language, we analyze generation failures that prevent reliable label extraction, focusing on two recurrent failure modes: *repetitive sequences* and *unfinished predictions*. Both analyses are aggregated across all evaluated prompting strategies and model-technique combinations. Similar errors have also been identified in prior work (Vykopal et al., 2025).

Failure Types. *Repetitive sequences* correspond to degenerate generation in which the model repeatedly produces identical character sequences until the end of decoding. We automatically detect these cases using a sliding-window heuristic that flags outputs containing at least three consecutive repetitions, similarly to Vykopal et al. (2024).

Unfinished predictions denote cases where no binary label could be extracted using language-specific regular expressions, typically due to excessively long generations or deviations from the expected output format.

Language-Level Patterns. Table 8 reports failure rates by instruction languages across all model-prompting combinations (including thinking mode) for the entire AMC-16k dataset (Vykopal et al., 2025). The most sever issues occur in low-resource and underrepresented languages, with **Burmese** standing out markedly: over 1% of outputs exhibit repetitive degeneration and more than 9% fail to produce a parsable label. Elevated failure rates are also observed in several non-Latin-script languages (e.g., Slovak, Serbo-Croatian, Bulgarian, Hindi or Hungarian), whereas high-resource Western European languages show mostly low failure percentages. These results suggest that part of the performance drop in low-resource setting stems from instruction-following and decoding failures, not only semantic misclassification, likely exacerbated by script and data underrepresentation during pre-training.

Model	Rep. (%)	Unf. (%)
Qwen3 8B	1.77	0.43
Qwen3 14B	0.96	0.34
Gemma3 12B	0.12	3.76
Gemma3 27B	0.16	0.06
LLaMA3.1 8B	**2.13**	**24.82**
LLaMA3.1 70B	1.85	12.76
Qwen3 8B (thinking)	0.00	0.00

Table 9: Percentage of generation failures aggregated across all languages and prompting strategies. Rep. denotes degenerate repetitive outputs; Unf. denotes cases where no binary label could be extracted.

Claim (truncated)	Freq	CosSim
(function(d, s, id) { var js, fjs...	159	0.778
https://scontent.ftxl2-1.fna.fbcdn...	64	0.780
https://drive.google.com/file/d/1T...	63	0.780
https://www.instagram.com/tv/CTjM561		0.784
1. All calls will be recorded...	44	0.779
https://www.facebook.com/photo.php.	37	0.779
The corona virus is large in size...	33	0.816
https://www.facebook.com/10007599.	30	0.786
Urgent information on Covid...	26	0.802
https://archive.ph/p0fQB	25	0.785

Table 10: Top-20 retrieved claims: truncated content with frequency (freq) and average cosine similarity score (CosSim).

Model-Level Patterns. To disentangle linguistic effects from model behavior, Table 9 aggregates the same failure types by model across all languages and prompting techniques. We observe substantial variation across models. Llama3.1 models, especially the 8B variant, exhibit the highest unfinished rates (up to 25%), indicating difficulty in adhering to output constraints in multilingual settings. Gemma3 models are the most robust overall, with very low failure rates, particularly for the 27B variant. Qwen3 models show moderate repetition but low unfinished rates. Notable, enabling "*thinking*" mode for Qwen3 eliminates both failure types in our experiments, suggesting that *thinking* mode improves output stability.

C Retrieval Bias

To measure retrieval bias we use two retriever models, Multilingual E5 and T5, that were used in the MultiClaim dataset paper (Pikuliak et al., 2023b) and they also serve as baselines in the corresponding SemEval-2025 task in multilingual and crosslingual fact-checked claim retrieval (Peng et al., 2025a).

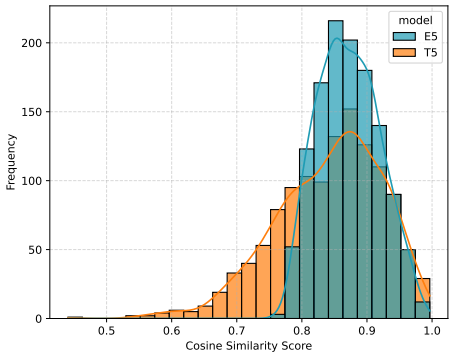

Figure 6: Distribution of Top-1 Cosine Similarity Scores between the E5 and T5 models.

To better understand the behavior of dense retrievers beyond aggregate performance metrics, we analyze the similarity scores used to rank candidate claims. Specifically, we examine the cosine similarity distributions of top-1 retrieved claims for both T5 and Multilingual E5. This allows us to investigate whether the similarity scores are high for the top-1 retrieved claim as shown in Figure 6. We can see a distribution between 0.8 and 1.0 for Multilingual E5 and 0.6 and 1.0 for T5. This relatively narrow band of high scores can lead to confirmation-style matches, where retrieved fact-checks are "close enough" semantically but potentially off-topic or unable to verify the post. High similarity does not always imply relevance, indicating that embedding-based retrieval alone is insufficient.

Table 10 shows the most frequently retrieved claims across the top-20 results for the Multilingual E5 model. Unlike the top-1 results (Table 4), this broader view reveals a heavier presence of noisy or structurally non-claim-like content. For example, the most frequent entry is a JavaScript snippet retrieved in 159 different cases. Similarly, raw URLs—including Facebook, Instagram, and Google Drive links—appear multiple times with high retrieval frequency. These items generally lack informative or verifiable content and suggest that the model may favor certain boilerplate or template-like artifacts during dense retrieval. This pattern highlights a retrieval bias toward recurring surface forms, particularly in the absence of strong semantic grounding. The average similarity scores for these claims remain relatively high, indicating that the model considers them relevant despite their lim-

	Monolingual					Cross-lingual - post-language					Cross-lingual - claim-language				
Model	**ZS**	**ZS + Task**	**FS + Task**	**CoT**	**Avg.**	**ZS**	**ZS + Task**	**FS + Task**	**CoT**	**Avg.**	**ZS**	**ZS + Task**	**FS + Task**	**CoT**	**Avg.**
Qwen3 8B	69.70 7.06	77.81 3.82	70.70 6.15	83.07 6.31	75.32 5.84	60.25 6.17	70.40 3.97	63.84 6.84	76.65 8.85	67.79 6.46	60.25 3.06	70.40 3.99	63.84 3.22	76.65 5.08	67.79 3.84
Qwen3 14B	73.40 9.60	77.08 4.94	74.72 3.49	**85.15** 5.01	77.59 5.78	64.12 8.33	70.99 3.46	69.29 1.56	**78.24** 6.72	70.66 5.02	64.12 5.85	70.99 2.57	69.29 2.21	**78.24** 3.22	70.66 3.46
Gemma3 12B	56.44 9.47	58.72 7.21	70.07 3.67	58.90 6.32	61.03 6.67	52.86 8.30	52.22 4.46	65.92 3.50	50.78 5.04	55.45 5.33	52.86 4.39	52.22 2.95	65.92 1.67	50.78 2.79	55.45 2.95
Gemma3 27B	59.35 6.82	55.54 10.85	72.72 2.34	52.41 6.53	60.00 6.64	53.69 4.78	49.54 3.99	64.94 2.04	46.37 4.42	53.64 3.81	53.69 2.83	49.54 4.16	64.94 1.86	46.37 2.89	53.64 2.94
Llama3.1 8B	50.46 8.76	71.13 8.59	51.12 5.43	67.18 4.81	59.97 6.90	45.53 8.04	66.48 8.76	50.56 4.66	58.82 5.87	55.35 6.83	45.53 7.61	66.48 4.40	50.56 2.57	58.82 3.06	55.35 4.41
Llama3.1 70B	73.97 6.82	74.38 7.69	54.04 3.48	75.48 7.10	69.47 6.27	65.57 7.09	65.77 9.22	54.34 4.52	67.66 4.73	63.33 6.39	65.57 3.59	65.77 4.12	54.34 2.70	67.66 3.35	63.33 3.44
Average	63.89 8.09	69.11 7.18	65.56 4.09	70.67 6.01	67.23 6.34	57.00 7.12	62.57 5.64	61.48 3.85	63.09 5.94	61.04 5.64	57.00 4.56	62.57 3.70	61.48 2.37	63.09 3.40	61.04 3.51
Qwen3 8B (thinking)	**77.19** 3.71	**82.74** 2.23	**81.27** 2.56	- -	**80.40** 2.83	**66.29** 3.50	**74.43** 3.32	**73.22** 3.64	- -	**71.31** 3.49	**66.29** 1.73	**74.43** 2.68	**73.22** 2.70	- -	**71.31** 2.37

Table 11: Macro F1 performance across 20 languages and language combinations for multiple models in three settings: monolingual, cross-lingual with post-language instructions, and cross-lingual with claim-language instructions. The first row per model shows absolute performance with English instructions; the second row shows the absolute difference when using the target language for instruction. ZS = zero-shot; FS = few-shot; Task = using Task description. The highest absolute score for each column is in **bold**.

Model	Technique	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Qwen3 8B	ZS	77.18	72.26	63.19	64.14	77.41	75.52	73.27	59.14	74.88	69.9	74.51	69.12	65.65	67.8	57.96	70.75	69.73	64.97	71.05	75.48	69.7
	ZS + Task	88.31	88.95	68.97	78.9	82.31	78.26	80.39	63.24	79.79	78.22	76.54	76.02	82.47	75.56	67.48	78.13	75.13	76.22	80.63	80.61	77.81
	FS + Task	81.22	79.9	65.33	66.93	74.05	67.22	79.05	64.24	77.18	67.09	73.75	64.57	72.1	68.27	61.49	71.12	68.67	64.83	76.11	70.86	70.7
	CoT	89.44	91.26	81.94	80.42	82.51	82.45	88.51	73.6	87.3	80.04	86.07	86.06	78.01	81.03	74.16	80.63	84.16	80.48	84.81	88.57	83.07
Average		84.04	83.09	69.86	72.60	79.07	75.86	80.31	65.06	79.79	73.81	77.72	73.94	74.56	73.17	65.27	75.16	74.42	71.63	78.15	78.88	75.32
Qwen3 14B	ZS	77.43	80.9	61.97	73.9	80.22	75.34	77.01	60.32	77.31	68.47	79.34	77.35	73.32	72.49	62.45	69.07	73.32	70.79	73.96	83.01	73.4
	ZS + Task	86.1	83.79	66.34	79.65	84.62	79.66	82.35	65.99	80.12	70.32	74.79	80.5	79.88	75.57	67.69	77.32	73.73	70.7	79.7	82.68	77.08
	FS + Task	84.12	82.35	63.83	75.89	81.21	78.62	82.23	64.96	79.84	70.82	70.61	75.95	76.61	76.25	66.54	75.3	72.08	63.64	78.62	74.88	74.72
	CoT	89.69	90.69	79.49	85.24	85.32	84.52	88.83	76.71	88.89	81.51	85.01	90.01	86.15	83.57	78.07	86.36	83.57	84.31	86.89	88.15	85.15
Average		84.34	84.43	67.91	78.67	82.84	79.54	82.61	67.00	81.54	72.78	77.44	80.95	78.99	76.97	68.69	77.01	75.68	72.36	79.79	82.18	77.59
Gemma3 12B	ZS	68.92	56.06	39.47	48.88	68.25	63.74	67.68	39.99	64.89	51.37	61.35	69.5	57.61	49.61	42.93	55.05	46.45	48.93	57.67	70.37	56.44
	ZS + Task	70.04	62.21	49.27	52.8	67.33	60.96	65.17	44.73	67.2	53.49	59.92	65.43	65.53	51.46	48.63	59.48	52.52	51.06	59.71	67.55	58.72
	FS + Task	76.97	80.46	67.32	68.27	77.74	69.57	75.04	65.11	71.31	62.52	69.74	63.63	69.37	71.63	61.67	75.34	66.23	63.72	72.21	73.6	70.07
	CoT	71.85	61.02	49.78	54.34	69.17	65.45	62.25	45.42	69.43	55.11	60.94	64.93	63.62	52.57	48.58	59.48	51.57	49.13	60.9	62.36	58.9
Average		71.95	64.94	51.46	56.07	70.62	64.93	67.54	48.81	68.21	55.62	62.99	65.87	64.03	56.32	50.45	62.34	54.19	53.21	62.62	68.47	61.03
Gemma3 27B	ZS	68.36	60.17	48.51	54.86	71.4	65.78	65.25	40.22	69.65	55.11	64.05	69.75	62.22	50.12	44.63	59.74	48.54	53.49	60.61	74.46	59.35
	ZS + Task	66.05	59.81	48.3	48.32	66.19	57.92	62.79	38.0	64.85	50.92	56.93	60.98	64.15	48.08	44.17	53.12	47.59	51.2	59.08	62.36	55.54
	FS + Task	84.04	81.41	65.28	64.65	83.13	74.48	76.19	58.5	75.96	64.58	74.54	71.26	83.35	67.64	63.95	76.85	67.97	64.05	74.98	81.59	72.72
	CoT	67.69	51.94	45.07	42.74	63.29	57.63	57.63	41.46	63.3	47.53	54.16	58.27	59.17	42.93	42.46	59.19	39.83	43.25	55.64	55.01	52.41
Average		71.54	63.33	51.79	52.64	71.00	63.95	65.47	44.55	68.44	54.54	62.42	65.07	67.22	52.19	48.80	62.23	50.98	53.00	62.58	68.36	60.01
Llama3.1 8B	ZS	57.11	48.74	39.2	41.81	57.49	64.0	59.79	37.99	66.48	43.87	53.4	55.46	43.76	42.17	42.36	54.57	48.55	40.03	51.33	61.11	50.46
	ZS + Task	79.84	72.29	62.5	68.36	78.36	77.26	80.4	61.12	79.96	64.18	68.44	75.32	74.43	70.63	56.81	70.93	73.2	63.63	73.34	71.55	71.13
	FS + Task	56.97	52.69	49.3	49.74	48.31	47.3	50.71	48.19	54.87	51.47	53.66	41.85	52.0	45.13	52.28	51.3	54.93	58.04	50.0	53.66	51.12
	CoT	73.39	74.6	60.73	65.96	69.8	76.23	70.71	56.94	73.52	64.66	67.09	74.33	62.22	64.88	55.72	66.87	65.81	64.97	64.3	70.89	67.18
Average		66.83	62.08	52.93	56.47	63.49	66.20	65.40	51.06	68.71	56.05	60.65	61.74	58.10	55.70	51.79	60.92	60.62	56.67	59.74	64.30	59.97
Llama3.1 70B	ZS	79.21	81.31	69.19	74.95	81.04	78.21	79.34	64.38	85.59	69.41	73.46	81.89	56.59	70.78	61.25	73.26	72.4	73.26	77.11	76.81	73.97
	ZS + Task	80.45	88.23	66.76	71.4	80.72	77.37	77.31	62.76	85.52	67.11	74.76	78.28	70.41	68.76	62.64	74.58	73.45	68.23	77.84	81.11	74.38
	FS + Task	56.98	47.65	54.01	53.99	47.65	69.42	64.7	52.97	51.92	46.87	58.97	41.41	47.02	52.48	54.11	62.64	57.54	51.33	60.44	48.8	54.04
	CoT	82.62	46.51	77.75	78.92	77.68	81.34	80.54	68.15	76.67	72.64	83.19	81.21	56.48	78.56	70.6	77.83	81.97	76.79	75.43	84.78	75.48
Average		74.82	65.93	66.93	69.82	71.77	76.59	75.47	62.07	74.93	64.01	72.60	70.70	57.63	67.65	62.15	72.08	71.34	67.40	72.71	72.88	69.47

Table 12: Macro F1 performance using English instruction in monolingual settings across six LLMs and four prompting techniques. The average is calculated across all languages for each model and technique.

Model	Technique	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Qwen3 8B	ZS	-2.79	-10.89	0.43	8.16	-4.90	0.00	-2.93	-0.98	-0.40	-2.30	-17.20	-9.05	-21.36	-6.55	-5.36	-6.29	-6.70	1.89	-8.33	-24.67	-6.01
	ZS + Task	-1.13	-0.18	2.45	2.79	-1.68	0.00	3.79	2.50	-3.14	-3.19	-6.89	-3.44	-19.68	4.85	3.68	-2.71	-1.11	-3.58	3.13	-6.42	-1.50
	FS + Task	4.36	5.84	-3.90	6.86	-10.00	0.00	-4.76	2.42	7.17	3.11	-9.59	1.04	-22.66	9.88	-2.11	4.62	10.37	-0.83	2.83	10.65	0.76
	CoT	-7.42	-3.54	-6.76	-4.07	-5.27	0.00	-4.64	-4.17	-7.06	-3.59	-15.47	-10.17	-28.37	-0.04	-0.68	-4.11	-12.79	-1.60	-2.15	-4.38	-6.31
Average		-1.74	-2.20	-1.94	3.44	-5.46	0.00	-2.13	-0.06	-0.86	-1.49	-12.29	-5.40	-23.02	2.04	-1.12	-2.12	-2.56	-1.03	-1.13	-6.21	-3.26
Qwen3 14B	ZS	2.17	-0.33	1.54	-0.64	-2.96	0.00	-7.87	-0.22	-18.43	-4.02	-22.21	-4.21	-43.63	-1.91	-19.65	-10.05	-7.61	-2.17	-13.31	-29.16	-9.23
	ZS + Task	6.89	7.62	11.53	2.45	2.76	0.00	2.50	4.98	6.37	5.14	-3.94	1.20	-10.46	3.12	6.94	1.64	6.64	7.42	4.09	-3.15	3.19
	FS + Task	1.88	7.68	1.74	-0.97	1.46	0.00	-2.27	4.01	1.54	-2.80	2.82	1.03	-19.01	-0.95	3.92	-4.94	0.30	5.93	3.26	3.23	0.39
	CoT	0.99	-5.23	-1.75	0.94	-0.58	0.00	-3.40	-3.13	-2.92	-0.67	-9.55	-25.56	-26.19	-0.32	-1.25	-6.45	-4.18	-1.59	-0.99	-4.52	-4.82
Average		2.98	2.43	3.27	0.44	0.17	0.00	-2.76	1.41	-3.36	-0.59	-8.22	-6.88	-24.82	-0.01	-2.51	-4.95	-1.21	2.40	-1.74	-8.40	-2.62
Gemma3 12B	ZS	1.44	9.80	17.40	-0.40	-6.25	0.00	0.03	6.65	-18.48	-0.22	-23.97	-15.68	-32.20	-1.81	2.68	1.95	4.74	10.59	2.76	-32.36	-3.67
	ZS + Task	18.01	7.85	2.80	3.15	9.74	0.00	4.39	3.81	8.47	5.17	4.77	7.54	-8.58	17.67	14.70	6.43	0.38	12.93	3.31	4.58	6.36
	FS + Task	9.61	-0.80	-6.85	-4.71	0.21	0.00	-2.85	0.59	12.75	-1.54	2.38	3.35	-5.92	0.97	2.71	-6.53	-1.20	-3.79	2.90	3.82	0.26
	CoT	8.86	-1.25	3.94	14.21	-5.92	0.00	-0.75	8.72	4.23	7.68	-8.68	2.57	-22.79	15.44	5.45	-2.57	-1.33	5.19	1.60	5.20	1.99
Average		9.48	3.90	4.32	3.06	-0.56	0.00	0.20	4.94	1.74	2.78	-6.37	-0.56	-17.37	8.07	6.38	-0.18	0.65	6.23	2.64	-4.69	1.23
Gemma3 27B	ZS	2.75	10.45	11.71	4.38	2.33	0.00	1.58	19.24	-9.06	0.42	-5.79	0.11	-16.92	3.55	-0.09	-5.43	16.05	5.82	6.16	-14.46	1.64
	ZS + Task	12.74	14.50	6.90	19.16	7.35	0.00	7.34	11.11	14.26	6.77	3.81	16.11	14.68	13.62	13.89	10.86	13.70	3.99	13.10	13.19	10.85
	FS + Task	2.43	-3.08	0.30	6.53	0.40	0.00	-1.20	0.01	5.12	-0.53	-3.93	3.55	-2.75	4.44	4.80	0.42	1.41	-1.52	-2.09	2.35	0.83
	CoT	10.61	2.75	3.98	24.90	-3.07	0.00	-1.94	-3.21	10.23	9.18	-0.98	10.48	0.00	4.97	5.57	4.77	7.25	5.48	5.52	15.61	5.60
Average		7.13	6.16	5.72	13.74	1.75	0.00	1.45	6.79	5.14	3.96	-1.72	7.56	-1.25	6.64	6.04	2.65	9.60	3.44	5.67	4.17	4.73
Llama3.1 8B	ZS	13.13	-10.19	5.67	10.69	-0.76	0.00	-5.28	15.08	-17.93	-6.59	2.43	3.03	-8.50	-4.45	4.06	-8.73	7.36	10.67	-2.31	-38.37	-1.55
	ZS + Task	0.66	-5.34	-1.77	0.34	-15.35	0.00	-7.80	-2.08	-17.14	3.67	-54.58	-6.47	-26.67	-1.54	6.61	-4.94	-5.73	-1.70	2.58	-6.83	-7.20
	FS + Task	5.07	5.53	1.69	8.05	10.59	0.00	9.30	3.67	-3.20	2.19	5.20	8.39	-1.37	15.76	8.29	0.06	0.06	-5.62	11.69	2.85	4.41
	CoT	-6.69	-2.83	5.71	-0.21	-0.92	0.00	-1.47	4.26	-20.68	-5.69	-2.94	-8.12	-11.36	-0.17	7.04	-0.66	-9.71	1.36	0.39	5.92	-2.34
Average		3.04	-3.21	2.83	4.72	-1.61	0.00	-1.31	5.23	-14.74	-1.61	-12.47	-0.79	-11.98	2.40	6.50	-3.57	-2.01	1.18	3.09	-9.11	-1.67
Llama3.1 70B	ZS	-10.83	-6.35	3.42	1.22	-7.71	0.00	-2.81	-11.48	-20.65	-2.19	-6.28	-6.48	-10.16	2.70	9.05	-4.32	-3.65	-4.80	-4.92	-17.44	-5.18
	ZS + Task	-23.07	1.36	8.24	9.17	0.79	0.00	4.62	11.38	0.77	-2.44	0.70	2.31	-23.53	11.37	9.86	1.80	1.06	13.40	4.80	-23.07	0.48
	FS + Task	-6.86	5.74	-2.74	0.43	-8.18	0.00	-8.22	-0.26	4.17	3.28	1.34	6.28	2.15	6.28	3.52	-1.69	2.75	0.70	-0.11	4.82	0.67
	CoT	-11.84	33.18	-3.79	-1.29	4.50	0.00	-0.13	0.69	5.67	-3.75	-18.97	-6.11	-9.18	2.80	0.49	-9.80	-11.29	3.86	-5.08	-9.50	-1.98
Average		-13.15	8.48	1.28	2.38	-2.65	0.00	-1.64	0.08	-2.51	-1.27	-5.80	-1.00	-10.18	5.79	5.73	-3.50	-2.78	3.29	-1.33	-11.30	-1.50

Table 13: Performance difference in Macro F1 when using target language instruction versus English instruction in monolingual settings across six LLMs and four prompting techniques. Positive values (green) indicate improved performance with target language instructions; negative values (red) indicate better performance with English instructions. The average is calculated across all languages for each model and technique.

Model	Technique	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Qwen3 8B	ZS	68.93	70.63	63.07	61.95	62.53	62.79	56.81	63.81	55.33	65.66	57.51	63.18	50.41	59.14	50.12	62.21	45.24	65.4	66.88	53.31	60.25
	ZS + Task	80.93	84.45	78.01	71.11	72.93	69.69	64.19	73.3	57.43	75.06	60.87	69.56	64.84	68.57	65.03	71.24	60.67	74.09	74.31	71.73	70.4
	FS + Task	68.88	71.49	80.8	62.24	59.65	57.25	58.79	67.79	57.58	67.1	56.85	60.57	66.09	61.95	58.3	68.14	49.81	77.29	67.25	58.95	63.84
	CoT	86.1	91.4	84.03	80.35	80.95	77.63	64.96	67.85	72.82	80.03	74.1	78.98	74.16	74.65	69.97	75.25	69.62	78.52	75.8	75.85	76.65
Average		76.21	79.49	76.48	68.91	69.02	66.84	61.19	68.19	60.79	71.96	62.33	68.07	63.88	66.08	60.86	69.21	56.34	73.83	71.06	64.96	67.79
Qwen3 14B	80.05	74.69	69.86	70.83	68.21	64.95	57.89	73.88	52.89	62.72	54.05	68.44	56.26	64.8	51.47	64.7	49.14	70.7	69.28	57.59	64.12	
	ZS + Task	82.08	80.74	78.38	71.37	66.41	70.86	71.96	75.58	53.94	71.49	60.26	68.76	60.68	73.74	71.05	79.98	64.81	76.74	73.78	67.22	70.99
	FS + Task	75.17	82.93	83.29	66.64	69.54	66.2	68.21	73.89	55.0	67.79	60.0	67.0	62.02	69.35	67.33	74.49	58.97	77.27	74.49	66.18	69.29
	CoT	88.62	91.71	84.03	85.66	82.76	78.16	74.62	77.36	67.75	81.91	71.1	83.28	76.04	75.73	75.76	80.82	63.97	74.49	78.33	72.64	78.24
Average		81.48	82.52	78.89	73.63	71.73	70.04	68.17	75.18	57.40	70.98	61.35	71.87	63.75	70.91	66.40	75.00	59.22	74.80	73.97	65.91	70.66
Gemma3 12B	ZS	67.13	42.17	64.37	55.36	50.9	54.04	48.16	49.24	36.0	56.87	38.67	52.0	49.64	45.74	49.59	62.93	46.06	63.73	61.31	63.3	52.86
	ZS + Task	61.75	45.92	63.36	49.68	50.9	52.69	48.44	50.83	38.14	57.96	40.76	48.68	48.93	44.76	50.44	61.34	49.17	60.66	60.9	59.07	52.22
	FS + Task	71.86	77.38	87.05	66.97	62.04	57.54	57.28	63.36	58.45	61.47	62.77	69.87	67.08	66.97	60.22	68.39	49.87	70.92	70.52	68.4	65.92
	CoT	66.84	49.4	62.46	54.02	53.21	53.55	47.42	47.74	35.45	53.66	38.67	50.82	46.26	36.67	45.71	64.41	45.24	59.24	54.68	50.12	50.78
Average		66.90	53.72	69.31	56.51	54.26	54.46	50.33	52.79	42.01	57.49	45.22	55.34	52.98	48.54	51.49	64.27	47.59	63.64	61.85	60.22	55.45
Gemma3 27B	ZS	70.94	48.76	62.39	54.7	57.0	54.21	48.8	50.83	41.65	57.5	36.76	55.18	49.11	42.07	56.61	64.84	49.55	61.28	58.27	53.46	53.69
	ZS + Task	64.53	45.61	59.38	49.31	47.19	46.4	49.64	42.65	36.72	54.18	34.42	46.49	44.3	40.66	52.64	57.72	51.56	58.97	59.07	49.43	49.54
	FS + Task	75.74	64.68	85.36	60.06	60.67	60.55	61.74	66.66	53.21	65.75	57.7	65.46	60.26	63.28	64.9	72.48	54.53	64.1	71.37	70.33	64.94
	CoT	63.6	44.36	57.97	48.37	45.94	46.24	46.44	39.05	31.72	49.06	31.75	42.46	43.22	27.17	51.99	55.99	46.37	56.82	53.49	45.48	46.37
Average		68.70	50.85	66.28	53.11	52.70	51.85	51.66	49.80	40.83	56.62	40.16	52.40	49.22	43.30	56.54	62.76	50.50	60.29	60.55	54.68	53.64
Llama3.1 8B	ZS	62.12	52.19	47.31	57.28	49.67	46.89	47.05	45.94	29.98	53.72	31.53	45.24	27.06	31.28	35.86	60.57	38.11	52.82	45.03	51.01	45.53
	ZS + Task	77.38	76.04	66.75	72.6	67.34	65.43	56.44	69.46	50.93	71.69	57.07	67.58	57.75	61.65	57.84	78.03	69.62	67.33	71.05	67.66	66.48
	FS + Task	50.65	51.2	48.69	51.02	51.32	52.79	48.37	53.07	53.1	49.97	47.53	48.24	54.4	53.31	47.71	47.37	48.65	52.97	51.62	49.3	50.56
	CoT	71.89	56.59	57.3	64.54	65.99	58.45	48.83	62.55	49.83	71.05	52.02	63.36	48.91	53.2	49.3	70.1	51.9	54.73	60.77	64.96	58.82
Average		65.51	59.01	55.01	61.36	58.58	55.89	50.17	57.76	45.96	61.61	47.04	56.11	47.03	49.86	47.68	64.02	52.07	56.96	57.12	58.23	55.35
Llama3.1 70B	ZS	77.95	72.07	69.07	67.35	72.78	65.42	52.79	64.58	57.61	78.0	58.58	66.58	56.56	61.38	55.0	74.51	52.52	66.33	71.61	70.66	65.57
	ZS + Task	75.06	73.61	71.07	67.41	65.71	65.74	56.99	61.08	55.74	73.45	59.93	64.23	58.58	63.25	58.3	75.75	49.56	70.91	73.63	75.44	65.77
	FS + Task	65.84	65.64	48.25	64.97	62.04	51.34	56.38	49.23	56.76	55.98	46.67	50.98	51.67	49.81	49.04	56.89	49.43	48.78	53.21	53.92	54.34
	CoT	77.33	74.25	71.21	73.46	69.07	65.53	56.56	67.4	65.46	73.04	60.77	72.89	67.34	66.48	53.3	77.55	49.37	62.32	74.03	75.89	67.66
Average		74.05	71.39	64.90	68.30	67.40	62.01	55.68	60.57	58.89	70.12	56.49	63.67	58.54	60.23	53.91	71.18	50.22	62.09	68.12	68.98	63.34

Table 14: Macro F1 performance with English instruction in cross-lingual settings across six LLMs and four prompting techniques. The average is calculated across all languages for each model and technique.

Model	Technique	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Qwen3 8B	ZS	-3.16	-6.06	0.00	-2.71	2.73	0.00	-2.47	5.14	0.95	-6.50	-5.14	-5.66	-0.91	-5.25	-20.79	-8.90	-17.21	3.23	-8.52	-18.15	-4.97
	ZS + Task	2.02	-0.66	0.00	0.49	2.26	0.00	4.02	-2.26	-0.43	-2.81	4.90	-0.08	3.48	3.34	-16.01	1.78	-11.74	-4.12	-8.10	-10.99	-1.75
	FS + Task	7.74	14.28	0.00	6.78	4.75	0.00	10.24	7.09	-7.30	-7.15	5.79	-5.31	-11.24	8.09	-6.00	-9.00	-11.70	-5.68	0.99	7.72	0.51
	CoT	-9.03	-17.98	0.00	-5.45	-5.44	0.00	-7.91	3.32	-8.34	-4.56	-2.33	-10.80	-10.18	-3.48	-21.66	-5.34	-32.27	-16.79	-3.45	-8.66	-8.52
Average		-0.61	-2.61	0.00	-0.22	1.07	0.00	0.97	3.32	-3.78	-5.25	0.80	-5.46	-4.71	0.68	-16.12	-5.36	-18.23	-5.84	-4.77	-7.52	-3.68
Qwen3 14B	ZS	-5.23	-28.14	0.00	-8.06	-1.32	0.00	-8.52	-1.34	-3.79	-9.73	-20.13	-5.22	-1.19	-3.94	-4.82	-13.25	-23.21	6.59	-2.23	-19.79	-7.67
	ZS + Task	-2.59	7.23	0.00	1.03	1.58	0.00	-2.92	3.12	6.89	0.94	3.39	1.37	4.74	-0.03	-6.02	-1.58	-16.76	-5.84	1.31	-1.85	-0.30
	FS + Task	2.34	1.91	0.00	1.15	-2.02	0.00	1.02	-0.16	-0.89	0.98	2.51	0.12	-0.21	-1.44	-0.51	3.47	-4.55	5.62	0.70	1.63	0.58
	CoT	-8.51	-3.24	0.00	-7.24	-7.41	0.00	-10.40	3.60	3.72	-0.25	-2.33	-4.88	-2.57	-3.40	-45.90	-5.48	-18.73	2.80	0.15	-3.79	-5.69
Average		-3.50	-5.56	0.00	-3.28	-2.29	0.00	-5.21	1.30	1.48	-2.02	-4.14	-2.15	0.19	-2.20	-14.31	-4.21	-15.81	2.29	-0.02	-5.95	-3.27
Gemma3 12B	ZS	-1.82	-11.50	0.00	0.74	-4.23	0.00	-0.37	-4.98	11.45	5.52	0.32	6.57	4.79	4.47	-21.41	-20.48	-24.04	-2.46	-14.20	-26.58	-4.91
	ZS + Task	2.42	11.46	0.00	2.03	0.97	0.00	-0.65	-0.74	9.22	6.61	11.72	2.38	0.18	16.46	4.87	-1.22	-6.46	7.46	-3.77	0.60	3.18
	FS + Task	0.00	7.31	0.00	-1.67	-0.50	0.00	2.17	3.13	-2.71	2.64	1.65	-4.38	-10.98	-2.52	7.07	7.42	7.82	-0.99	-2.20	4.81	0.90
	CoT	-1.10	3.98	0.00	-3.40	9.70	0.00	0.50	7.89	5.29	6.17	5.56	-3.84	-2.50	7.43	0.00	-15.60	-19.74	0.44	6.65	1.04	0.42
Average		-0.12	2.81	0.00	-0.57	1.49	0.00	0.41	1.32	5.81	5.23	4.81	0.18	-2.13	6.46	-2.37	-7.47	-10.60	1.11	-3.38	-5.03	-0.10
Gemma3 27B	ZS	0.92	-5.19	0.00	-3.14	-3.58	0.00	-1.25	-4.17	5.49	1.82	-3.51	3.61	1.90	-0.06	-11.90	-5.17	-21.93	-3.78	-10.53	-7.77	-3.41
	ZS + Task	2.16	13.16	0.00	1.69	8.45	0.00	1.67	9.12	0.00	-1.20	9.65	1.69	2.50	2.06	5.67	2.42	1.90	5.99	-2.57	7.89	3.61
	FS + Task	-3.85	3.05	0.00	-0.69	-0.38	0.00	-2.30	-3.35	-3.38	0.44	1.76	-5.25	-1.28	3.21	0.73	-0.92	-4.66	-0.96	-2.89	-1.73	-1.12
	CoT	-1.73	10.25	0.00	-3.79	13.97	0.00	1.74	16.60	3.00	2.49	5.23	1.78	-2.73	3.03	2.02	-2.75	-0.72	2.86	7.22	6.39	3.24
Average		-0.62	5.32	0.00	-1.48	4.61	0.00	-0.03	4.55	1.28	0.89	3.28	0.46	0.10	2.06	-0.87	-1.60	-6.35	1.03	-2.19	1.20	0.58
Llama3.1 8B	ZS	-4.93	-11.13	0.00	-4.41	-0.81	0.00	-11.25	7.94	19.51	-5.30	8.06	1.30	6.03	-0.34	-3.07	-5.74	-11.87	11.11	-12.56	-35.51	-2.65
	ZS + Task	-3.83	-26.70	0.00	-2.63	-0.52	0.00	1.46	-1.45	14.13	-4.80	6.00	-5.67	0.61	1.92	-14.69	-52.87	-19.70	3.02	3.23	-11.93	-5.72
	FS + Task	7.16	14.81	0.00	9.86	6.53	0.00	4.07	1.99	-2.01	2.39	7.24	4.84	-1.85	4.48	2.56	3.33	-3.07	12.93	1.32	2.72	3.97
	CoT	-11.19	-21.48	0.00	-8.18	-1.09	0.00	-7.09	0.32	11.17	-2.21	6.95	-5.71	-0.10	0.18	-14.46	-10.59	-9.79	0.24	-0.43	-6.20	-3.98
Average		-3.20	-11.12	0.00	-1.34	1.03	0.00	-3.20	2.20	10.70	-2.48	7.06	-1.31	1.17	1.56	-7.42	-16.47	-11.11	6.82	-2.11	-12.73	-2.10
Llama3.1 70B	ZS	-1.68	1.19	0.00	-6.05	-5.12	0.00	-0.80	-0.61	4.11	-12.28	-1.74	5.05	17.41	0.20	-9.87	-4.86	47.48	-5.40	-1.25	-16.73	0.45
	ZS + Task	2.01	8.90	0.00	0.44	6.48	0.00	3.08	5.81	13.35	6.68	8.74	4.68	7.19	4.49	-5.68	-0.96	50.44	-21.55	-6.49	-27.52	3.00
	FS + Task	-7.17	-10.28	0.00	-2.34	-3.99	0.00	-10.93	4.49	-10.96	-8.07	6.31	1.02	0.64	-0.70	-4.68	-2.97	-0.84	-10.55	0.69	3.83	-2.82
	CoT	-10.82	-3.62	0.00	-6.89	4.52	0.00	-5.77	6.76	3.98	4.79	1.86	-5.59	1.66	3.74	-4.41	-15.21	-0.14	-0.80	1.17	-12.91	-1.88
Average		-2.28	-0.06	0.00	-2.65	-0.88	0.00	-2.88	3.23	2.16	-4.56	4.43	3.59	8.42	1.33	-6.74	-2.93	32.36	-12.50	-2.35	-13.47	0.21

Table 15: Difference in Macro F1 score when using post-language instructions versus English in cross-lingual settings. Each column represents a language pair (post–claim language), where the first language indicates the instruction language. Positive values (green) indicate improved performance when instructions are given in the post language. The table presents results for six models and four prompting techniques, with average values computed for each model–prompting combination as well as for each language pair.

Model	Technique	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Qwen3 8B	ZS	0.00	0.00	1.75	0.00	0.00	-8.86	3.72	8.58	0.62	0.00	-4.45	0.00	-1.11	0.42	-0.56	0.00	-13.95	1.29	-8.97	-6.88	-1.42
	ZS + Task	0.00	0.00	-8.15	0.00	0.00	-8.57	-5.97	-2.00	5.15	0.00	2.10	0.00	2.34	1.76	-11.01	0.00	-11.74	-6.68	4.51	-9.84	-2.40
	FS + Task	0.00	0.00	5.06	0.00	0.00	4.77	4.35	3.01	-11.66	0.00	8.19	0.00	-4.01	2.38	-6.75	0.00	0.06	2.57	1.82	9.74	0.98
	CoT	0.00	0.00	-10.48	0.00	0.00	-0.50	-6.10	7.38	-9.10	0.00	-5.78	0.00	-6.42	-6.65	-8.02	0.00	-23.56	-8.94	0.67	-8.09	-4.28
Average		0.00	0.00	-2.95	0.00	0.00	-3.29	-1.00	4.24	-3.75	0.00	0.01	0.00	-2.30	-0.52	-6.59	0.00	-12.30	-2.94	-0.49	-3.77	-1.78
Qwen3 14B	ZS	0.00	0.00	-8.33	0.00	0.00	-17.89	-7.83	-10.59	2.84	0.00	3.03	0.00	-19.43	3.23	-0.57	0.00	-7.10	-8.00	-16.27	-11.94	-4.94
	ZS + Task	0.00	0.00	5.32	0.00	0.00	-1.89	-4.54	1.21	8.40	0.00	3.61	0.00	4.62	1.43	2.03	0.00	-10.64	-0.60	5.19	1.96	0.81
	FS + Task	0.00	0.00	-1.02	0.00	0.00	-2.03	-7.28	1.93	-0.88	0.00	1.69	0.00	4.71	-7.13	3.58	0.00	-7.46	3.86	2.54	-0.17	-0.38
	CoT	0.00	0.00	-5.75	0.00	0.00	-5.91	-15.33	-1.58	-1.46	0.00	-3.62	0.00	-1.50	0.40	-8.35	0.00	-16.72	-2.91	0.73	-0.18	-3.11
Average		0.00	0.00	-2.44	0.00	0.00	-6.93	-8.74	-2.26	2.22	0.00	1.17	0.00	-2.90	-0.52	-0.83	0.00	-10.48	-1.91	-1.95	-2.58	-1.91
Gemma3 12B	ZS	0.00	0.00	-2.78	0.00	0.00	-1.46	-1.57	3.16	11.62	0.00	8.22	0.00	-5.35	-8.98	0.13	0.00	-26.98	-4.63	-7.85	-5.03	-2.08
	ZS + Task	0.00	0.00	10.90	0.00	0.00	2.10	1.28	-0.53	2.12	0.00	5.74	0.00	11.85	0.00	8.85	0.00	1.18	-1.68	11.49	-1.36	2.60
	FS + Task	0.00	0.00	-3.35	0.00	0.00	-2.03	0.33	-1.69	-3.07	0.00	-3.39	0.00	-5.64	-5.38	-1.78	0.00	-0.06	1.25	1.98	-3.44	-1.31
	CoT	0.00	0.00	6.24	0.00	0.00	-6.82	-3.47	0.43	3.56	0.00	11.17	0.00	2.14	4.69	5.97	0.00	-2.91	-2.86	2.21	-3.35	0.85
Average		0.00	0.00	2.75	0.00	0.00	-2.05	-0.86	0.34	3.56	0.00	5.43	0.00	0.75	-2.42	3.29	0.00	-7.19	-1.98	1.96	-3.30	0.01
Gemma3 27B	ZS	0.00	0.00	-1.26	0.00	0.00	-3.21	0.79	10.19	10.15	0.00	11.40	0.00	-5.53	0.72	-0.76	0.00	-4.61	-0.92	-2.56	-4.44	0.50
	ZS + Task	0.00	0.00	8.97	0.00	0.00	4.39	2.22	10.78	6.24	0.00	6.98	0.00	10.78	9.73	6.15	0.00	3.13	4.38	8.87	0.52	4.16
	FS + Task	0.00	0.00	-0.91	0.00	0.00	-0.15	-3.30	-3.11	0.55	0.00	2.26	0.00	3.04	5.28	-1.91	0.00	13.21	1.91	0.72	-0.90	0.83
	CoT	0.00	0.00	10.69	0.00	0.00	2.91	1.23	0.89	5.18	0.00	-0.72	0.00	3.40	16.60	4.53	0.00	1.07	-7.23	2.97	0.34	2.09
Average		0.00	0.00	4.37	0.00	0.00	0.99	0.23	4.69	5.53	0.00	4.98	0.00	2.92	8.08	2.00	0.00	3.20	-0.47	2.50	-1.12	1.90
Llama3.1 8B	ZS	0.00	0.00	5.85	0.00	0.00	-17.65	-25.71	7.68	9.09	0.00	18.61	0.00	8.46	10.36	14.57	0.00	-16.72	0.68	2.05	-14.85	0.12
	ZS + Task	0.00	0.00	10.21	0.00	0.00	-4.07	-1.69	1.52	7.12	0.00	-9.61	0.00	1.63	-0.21	-1.47	0.00	-28.40	10.46	-1.90	-9.75	-1.31
	FS + Task	0.00	0.00	3.01	0.00	0.00	-2.98	2.33	2.89	-2.17	0.00	4.71	0.00	5.35	3.07	9.33	0.00	-1.35	1.33	10.20	2.70	1.92
	CoT	0.00	0.00	2.25	0.00	0.00	0.79	3.42	-2.67	3.23	0.00	-0.95	0.00	16.06	-1.73	3.04	0.00	-14.75	1.43	10.46	0.43	1.05
Average		0.00	0.00	5.33	0.00	0.00	-5.98	-5.41	2.35	4.32	0.00	3.19	0.00	7.88	2.87	6.37	0.00	-15.31	3.48	5.20	-5.37	0.45
Llama3.1 70B	ZS	0.00	0.00	-4.71	0.00	0.00	-4.67	-1.86	-7.44	9.53	0.00	-1.34	0.00	11.41	-3.10	-1.07	0.00	-10.27	-5.97	0.79	-9.69	-1.42
	ZS + Task	0.00	0.00	-6.85	0.00	0.00	2.92	-0.28	9.52	7.48	0.00	2.99	0.00	11.56	7.31	-9.06	0.00	-6.78	6.88	7.18	-3.62	1.46
	FS + Task	0.00	0.00	-3.34	0.00	0.00	4.09	-6.30	6.28	-8.85	0.00	2.07	0.00	-0.66	2.60	-6.74	0.00	-5.06	0.02	4.58	-3.47	-0.74
	CoT	0.00	0.00	-11.97	0.00	0.00	-9.77	-4.04	1.54	-2.48	0.00	1.37	0.00	-1.11	4.76	-8.32	0.00	-13.07	-1.66	2.08	-4.82	-2.37
Average		0.00	0.00	-6.72	0.00	0.00	-1.86	-3.12	2.47	1.42	0.00	1.27	0.00	5.30	2.89	-6.30	0.00	-8.80	-0.18	3.66	-5.40	-0.77

Table 16: Difference in Macro F1 score when using claim-language instructions versus English in cross-lingual settings. Each column represents a language pair (post–claim language), with the second language indicating the instruction language. Positive values (green) reflect improved performance when using claim-language instructions. The table presents results for six models and four prompting techniques, with average values computed for each model–prompting combination as well as for each language pair.

Technique	Thinking	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Zero-Shot	✗	77.18	72.26	63.19	64.14	77.41	75.52	73.27	59.14	74.88	69.9	74.51	69.12	65.65	67.8	57.96	70.75	69.73	64.97	71.05	75.48	69.7
	✓	81.73	90.87	72.10	73.08	80.47	80.19	83.39	60.37	80.94	75.1	82.23	84.9	68.63	74.24	67.82	72.1	76.93	74.76	76.65	87.31	77.19
Zero-Shot + Task Description	✗	88.31	88.95	68.97	78.9	82.31	78.26	80.39	63.24	79.79	78.22	76.54	76.02	82.47	75.56	67.48	78.13	75.13	76.22	80.63	80.61	77.81
	✓	86.98	92.18	79.69	79.97	84.13	84.83	87.3	73.12	84.88	79.54	83.28	87.7	79.6	80.77	75.91	80.92	84.0	80.31	83.96	85.66	82.74
Few-Shot + Task Description	✗	81.22	79.9	65.33	66.93	74.05	67.22	79.05	64.24	77.18	67.09	73.75	64.57	72.1	68.27	61.49	71.12	68.67	64.83	76.11	70.86	70.7
	✓	89.55	90.49	75.49	80.47	76.77	82.45	84.07	71.74	85.48	80.39	84.78	82.29	72.78	83.43	74.56	80.14	76.66	81.87	85.48	86.51	81.27

Table 17: Comparison of Macro F1 performance using English for the instruction with (✓) and without (✗) thinking mode for the Qwen3 8B in a monolingual setting.

Technique	Thinking	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Zero-Shot	✗	-2.79	-10.89	0.43	8.16	-4.90	0.00	-2.93	-0.98	-0.40	-2.30	-17.20	-9.05	-21.36	-6.55	-5.36	-6.29	-6.70	1.89	-8.33	-24.67	-6.01
	✓	-0.08	-8.17	0.41	2.97	-5.84	0.00	-4.79	0.20	-0.42	3.73	-0.38	-3.29	-17.27	0.61	-3.92	-4.23	-3.48	4.78	-4.11	-6.69	-2.50
Zero-Shot + Task Description	✗	-1.13	-0.18	2.45	2.79	-1.68	0.00	3.79	2.50	-3.14	-3.19	-6.89	-3.44	-19.68	4.85	3.68	-2.71	-1.11	-3.58	3.13	-6.42	-1.50
	✓	3.81	0.58	0.89	0.66	-1.83	0.00	-1.93	-0.48	0.48	0.34	-2.93	-4.35	-14.52	1.90	1.15	-1.73	-3.73	-2.52	-0.13	-0.70	-1.25
Few-Shot + Task Description	✗	4.36	5.84	-3.90	6.86	-10.00	0.00	-4.76	2.42	7.17	3.11	-9.59	1.04	-22.66	9.88	-2.11	4.62	10.37	-0.83	2.83	10.65	0.76
	✓	-3.10	1.79	0.28	-1.78	4.31	0.00	-0.17	-1.37	0.65	5.47	-4.73	-3.15	-11.55	-3.35	-1.36	-0.62	3.51	-1.12	-2.35	-0.64	-0.96

Table 18: Impact of the thinking mode on performance in monolingual settings across three prompting techniques for the `Qwen3 8B`. The table shows the difference in Macro F1 score when using target language instructions versus English. Each row compares performance with (✓) and without (✗) "thinking mode", across 20 languages. Positive values (green) indicate improved performance with target language instructions. Average scores are reported in the final column.

Technique	Thinking	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Zero-Shot	✗	68.93	70.63	63.07	61.95	62.53	62.79	56.81	63.81	55.33	65.66	57.51	63.18	50.41	59.14	50.12	62.21	45.24	65.4	66.88	53.31	60.25
	✓	78.39	69.36	65.59	70.61	68.47	67.75	55.17	66.7	63.5	75.65	60.61	71.96	57.94	61.89	55.74	75.2	57.93	69.0	69.37	64.9	66.29
Zero-Shot + Task Description	✗	80.93	84.45	78.01	71.11	72.93	69.69	64.19	73.3	57.43	75.06	60.87	69.56	64.84	68.57	65.03	71.24	60.67	74.09	74.31	71.73	70.4
	✓	85.16	81.52	76.53	80.37	76.11	76.04	62.43	75.78	70.29	82.69	67.03	76.15	69.05	71.93	66.85	78.39	66.16	72.17	79.14	74.73	74.43
Few-Shot + Task Description	✗	68.88	71.49	80.8	62.24	59.65	57.25	58.79	67.79	57.58	67.1	56.85	60.57	66.09	61.95	58.3	68.14	49.81	77.29	67.25	58.95	63.84
	✓	80.62	84.92	77.84	76.14	75.41	70.72	66.43	75.18	69.1	76.2	70.8	74.5	70.95	74.65	67.45	78.43	58.97	68.21	76.14	71.75	73.22

Table 19: Macro F1 performance using English for the instruction with (✓) and without (✗) thinking mode for the `Qwen3 8B` in a cross-lingual setting.

Technique	Thinking	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Zero-Shot	✗	-3.16	-6.06	0.00	-2.71	2.73	0.00	-2.47	5.14	0.95	-6.50	-5.14	-5.66	-0.91	-5.25	-20.79	-8.90	-17.21	3.23	-8.52	-18.15	-4.97
	✓	-1.67	4.03	0.00	-3.89	4.50	0.00	-4.16	0.61	1.65	-0.36	-0.36	-4.53	1.23	-2.01	-9.06	-0.58	-22.69	-0.37	-2.84	-5.15	-2.28
Zero-Shot + Task Description	✗	2.02	-0.66	0.00	0.49	2.26	0.00	4.02	-2.26	-0.43	-2.81	4.90	-0.08	3.48	3.34	-16.01	1.78	-11.74	-4.12	-8.10	-10.99	-1.75
	✓	-3.26	6.52	0.00	-6.35	0.38	0.00	-0.53	-1.77	3.42	-1.76	4.94	0.53	-0.54	2.86	-10.78	-2.14	-13.33	-0.22	-5.95	-1.12	-1.45
Few-Shot + Task Description	✗	7.74	14.28	0.00	6.78	4.75	0.00	10.24	7.09	-7.30	-7.15	5.79	-5.31	-11.24	8.09	-6.00	-9.00	-11.70	-5.68	0.99	7.72	0.51
	✓	0.75	2.95	0.00	2.02	1.72	0.00	8.06	3.36	1.06	2.23	3.02	-1.16	3.37	3.65	-18.42	-0.78	-11.33	2.71	-0.69	-5.46	-0.15

Table 20: Impact of the thinking mode on performance in cross-lingual settings across three prompting techniques for the `Qwen3 8B`. The table shows the difference in the Macro F1 score when using post-language instructions (the first language in the column name) versus English. Each row compares performance with (✓) and without (✗) "thinking mode", across 20 language pairs. Positive values (green) indicate improved performance with post-language instructions. Average scores are reported in the final column.

Technique	Thinking	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Zero-Shot	✗	0.00	0.00	1.75	0.00	0.00	-8.86	3.72	8.58	0.62	0.00	-4.45	0.00	-1.11	0.42	-0.56	0.00	-13.95	1.29	-8.97	-6.88	-1.42
	✓	0.00	0.00	-0.48	0.00	0.00	-3.91	-0.83	-2.00	0.48	0.00	-0.15	0.00	-2.04	2.92	-2.71	0.00	-7.44	-4.18	2.94	-4.44	-1.09
Zero-Shot + Task Description	✗	0.00	0.00	-8.15	0.00	0.00	-8.57	-5.97	-2.00	5.15	0.00	2.10	0.00	2.34	1.76	-11.01	0.00	-11.74	-6.68	4.51	-9.84	-2.40
	✓	0.00	0.00	-2.89	0.00	0.00	-6.57	-4.21	-1.96	-4.00	0.00	0.88	0.00	4.20	-1.75	-15.15	0.00	-7.18	-0.92	0.90	-3.03	-2.08
Few-Shot + Task Description	✗	0.00	0.00	5.06	0.00	0.00	4.77	4.35	3.01	-11.66	0.00	8.19	0.00	-4.01	2.38	-6.75	0.00	0.06	2.57	1.82	9.74	0.98
	✓	0.00	0.00	0.91	0.00	0.00	7.12	-2.78	2.17	0.44	0.00	-1.77	0.00	4.19	2.14	-15.56	0.00	-1.04	10.75	1.38	-3.70	0.21

Table 21: Impact of the thinking mode on performance in cross-lingual settings across three prompting techniques for the `Qwen3 8B`. The table shows the difference in the Macro F1 score when using claim-language instructions (the second language in the column name) versus English. Each row compares performance with (✓) and without (✗) "thinking mode", across 20 language pairs. Positive values (green) indicate improved performance with claim-language instructions. Average scores are reported in the final column.

Model	Technique	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Qwen3 8B	ZS	0.79	0.85	0.68	0.58	0.84	0.76	0.80	0.61	0.76	0.84	0.90	0.66	0.78	0.74	0.61	0.64	0.90	0.78	0.71	0.86	0.75
	ZS + Task	0.91	0.95	0.73	0.77	0.86	0.72	0.83	0.65	0.82	0.87	0.86	0.67	0.94	0.81	0.68	0.77	0.84	0.84	0.80	0.89	0.81
	FS + Task	0.86	0.91	0.71	0.62	0.81	0.57	0.87	0.68	0.81	0.75	0.85	0.59	0.92	0.68	0.62	0.67	0.82	0.73	0.75	0.79	0.75
	CoT	0.97	0.99	0.93	0.92	0.97	0.88	0.95	0.87	0.96	0.95	0.97	0.91	0.98	0.90	0.89	0.88	0.95	0.94	0.92	0.97	0.94
Average		0.88	0.93	0.76	0.72	0.87	0.73	0.86	0.70	0.84	0.85	0.90	0.71	0.91	0.78	0.70	0.74	0.88	0.82	0.80	0.88	0.81
Qwen3 14B	ZS	0.78	0.94	0.63	0.68	0.89	0.78	0.80	0.62	0.81	0.77	0.91	0.79	0.85	0.73	0.65	0.67	0.87	0.83	0.80	0.95	0.79
	ZS + Task	0.88	0.93	0.68	0.75	0.85	0.70	0.84	0.64	0.81	0.76	0.84	0.72	0.93	0.78	0.65	0.72	0.82	0.75	0.79	0.91	0.79
	FS + Task	0.88	0.92	0.69	0.71	0.82	0.70	0.84	0.66	0.84	0.77	0.80	0.74	0.93	0.80	0.67	0.72	0.83	0.68	0.78	0.85	0.78
	CoT	0.95	0.99	0.90	0.92	0.96	0.87	0.94	0.85	0.96	0.92	0.96	0.92	0.99	0.92	0.85	0.92	0.95	0.94	0.95	0.95	0.93
Average		0.87	0.95	0.73	0.77	0.88	0.76	0.86	0.69	0.86	0.81	0.88	0.79	0.93	0.81	0.71	0.76	0.87	0.80	0.83	0.92	0.82
Gemma3 12B	ZS	0.66	0.60	0.28	0.28	0.59	0.45	0.65	0.26	0.59	0.48	0.65	0.55	0.65	0.38	0.29	0.36	0.44	0.45	0.46	0.78	0.49
	ZS + Task	0.67	0.69	0.43	0.33	0.58	0.39	0.61	0.32	0.62	0.51	0.61	0.48	0.77	0.41	0.36	0.42	0.53	0.47	0.49	0.74	0.52
	FS + Task	0.96	0.96	0.78	0.70	0.80	0.72	0.90	0.73	0.93	0.82	0.88	0.93	0.98	0.75	0.67	0.80	0.82	0.73	0.77	0.94	0.83
	CoT	0.70	0.66	0.42	0.35	0.61	0.46	0.57	0.33	0.65	0.54	0.63	0.47	0.74	0.43	0.36	0.42	0.52	0.45	0.50	0.67	0.52
Average		0.75	0.73	0.48	0.42	0.65	0.51	0.68	0.41	0.70	0.59	0.69	0.61	0.79	0.49	0.42	0.50	0.58	0.53	0.56	0.78	0.59
Gemma3 27B	ZS	0.65	0.66	0.40	0.35	0.65	0.47	0.62	0.26	0.65	0.54	0.69	0.55	0.71	0.39	0.31	0.43	0.47	0.51	0.50	0.81	0.53
	ZS + Task	0.62	0.65	0.40	0.27	0.56	0.35	0.57	0.23	0.58	0.47	0.57	0.41	0.74	0.36	0.30	0.33	0.46	0.47	0.47	0.67	0.47
	FS + Task	0.92	0.92	0.69	0.58	0.84	0.63	0.77	0.54	0.83	0.73	0.84	0.83	0.98	0.69	0.64	0.71	0.79	0.68	0.73	0.93	0.76
	CoT	0.64	0.53	0.35	0.20	0.51	0.33	0.50	0.27	0.56	0.42	0.53	0.38	0.67	0.29	0.28	0.41	0.34	0.36	0.42	0.56	0.43
Average		0.71	0.69	0.46	0.35	0.64	0.45	0.62	0.33	0.66	0.54	0.66	0.54	0.78	0.43	0.38	0.47	0.52	0.51	0.53	0.74	0.55
Llama3.1 8B	ZS	0.50	0.50	0.28	0.20	0.47	0.51	0.54	0.24	0.63	0.38	0.52	0.35	0.44	0.29	0.30	0.36	0.50	0.32	0.40	0.66	0.42
	ZS + Task	0.82	0.84	0.62	0.62	0.84	0.75	0.84	0.58	0.84	0.72	0.77	0.70	0.88	0.75	0.52	0.63	0.83	0.69	0.74	0.83	0.74
	FS + Task	0.89	0.94	0.87	0.87	0.89	0.83	0.88	0.84	0.87	0.90	0.95	0.91	0.96	0.87	0.90	0.88	0.94	0.87	0.84	0.94	0.89
	CoT	0.76	0.86	0.65	0.61	0.74	0.74	0.73	0.56	0.77	0.75	0.74	0.68	0.84	0.67	0.55	0.57	0.74	0.75	0.62	0.84	0.71
Average		0.74	0.79	0.61	0.58	0.74	0.71	0.75	0.56	0.78	0.69	0.75	0.66	0.78	0.65	0.57	0.61	0.75	0.66	0.65	0.82	0.69
Llama3.1 70B	ZS	0.82	0.92	0.69	0.69	0.84	0.77	0.82	0.64	0.92	0.78	0.86	0.82	0.78	0.71	0.63	0.68	0.81	0.80	0.78	0.91	0.78
	ZS + Task	0.81	0.94	0.66	0.64	0.82	0.69	0.78	0.59	0.87	0.74	0.83	0.73	0.90	0.68	0.60	0.67	0.81	0.74	0.76	0.92	0.76
	FS + Task	0.95	0.96	0.94	0.95	0.95	0.93	0.96	0.96	0.95	0.96	0.99	0.98	0.98	0.95	0.97	0.95	0.97	0.92	0.96	1.00	0.96
	CoT	0.88	0.48	0.82	0.82	0.86	0.78	0.84	0.73	0.82	0.81	0.93	0.80	0.80	0.82	0.77	0.75	0.92	0.87	0.75	0.95	0.81
Average		0.87	0.83	0.78	0.78	0.87	0.79	0.85	0.73	0.89	0.82	0.90	0.83	0.87	0.79	0.74	0.76	0.88	0.83	0.81	0.95	0.83

Table 22: The capabilities of LLMs in filtering irrelevant pairs based on TNR (higher is better) using English instruction in monolingual settings. The average TNR is calculated across all languages for each model and technique.

Model	Technique	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Qwen3 8B	ZS	-0.01	-0.10	0.01	0.17	-0.10	0.00	0.17	0.11	0.03	-0.05	-0.29	-0.13	-0.32	-0.11	0.02	0.12	0.04	0.07	-0.08	-0.35	-0.04
	ZS + Task	-0.02	0.00	0.05	0.14	-0.04	0.00	0.06	0.03	-0.05	-0.06	-0.09	-0.04	-0.18	0.10	0.09	-0.03	0.02	-0.01	0.09	-0.08	0.00
	FS + Task	0.03	0.05	-0.08	0.19	-0.21	0.00	0.08	0.09	0.11	0.09	-0.10	0.30	-0.36	0.22	0.01	0.11	0.11	-0.02	0.16	0.15	0.05
	CoT	-0.01	-0.02	-0.08	0.04	-0.03	0.00	-0.04	-0.10	-0.11	-0.05	-0.16	-0.12	-0.44	0.01	0.03	-0.01	-0.11	0.00	0.02	-0.03	-0.06
Average		0.00	-0.02	-0.03	0.13	-0.09	0.00	0.07	0.03	-0.01	-0.02	-0.16	0.00	-0.33	0.05	0.04	0.04	0.01	0.01	0.05	-0.08	-0.01
Qwen3 14B	ZS	0.05	0.00	0.03	0.00	0.00	0.00	0.04	0.07	-0.31	-0.05	-0.34	-0.17	-0.61	-0.04	-0.28	-0.09	-0.10	-0.04	-0.19	-0.40	-0.12
	ZS + Task	0.07	0.05	0.18	0.05	0.10	0.00	0.03	0.07	0.11	0.07	-0.03	0.06	-0.09	0.04	0.11	0.05	0.06	0.10	0.07	-0.03	0.05
	FS + Task	0.00	0.05	-0.01	-0.03	0.09	0.00	-0.03	0.05	0.01	-0.03	0.03	-0.02	-0.19	-0.02	0.05	-0.07	-0.01	0.07	0.04	0.02	0.00
	CoT	-0.02	-0.02	0.01	0.00	-0.01	0.00	-0.04	-0.02	-0.03	0.01	-0.09	-0.45	-0.29	-0.01	-0.03	-0.10	-0.01	0.00	-0.03	-0.01	-0.06
Average		0.03	0.02	0.05	0.01	0.04	0.00	0.00	0.04	-0.05	0.00	-0.11	-0.14	-0.30	-0.01	-0.04	-0.05	-0.01	0.03	-0.03	-0.11	-0.03
Gemma3 12B	ZS	0.03	0.16	0.26	0.00	-0.09	0.00	0.06	0.10	-0.26	0.02	-0.36	-0.24	-0.47	-0.03	0.05	0.07	0.08	0.17	0.12	-0.47	-0.04
	ZS + Task	0.22	0.10	0.04	0.04	0.16	0.00	0.07	0.06	0.13	0.08	0.07	0.12	-0.12	0.26	0.22	0.10	0.01	0.19	0.05	0.06	0.09
	FS + Task	-0.04	-0.04	-0.15	-0.14	-0.03	0.00	-0.07	-0.04	-0.01	-0.08	-0.08	-0.06	-0.20	0.05	-0.03	-0.14	-0.06	-0.08	0.00	-0.04	-0.06
	CoT	0.12	-0.01	0.07	0.23	-0.09	0.00	-0.01	0.13	0.06	0.12	-0.12	0.05	-0.35	0.22	0.07	-0.04	0.00	0.08	0.04	0.07	0.03
Average		0.08	0.05	0.05	0.04	-0.01	0.00	0.01	0.06	-0.02	0.03	-0.12	-0.03	-0.29	0.13	0.08	0.00	0.01	0.09	0.05	-0.09	0.01
Gemma3 27B	ZS	0.06	0.17	0.19	0.06	0.06	0.00	0.09	0.54	-0.11	0.01	-0.06	0.05	-0.23	0.05	0.01	0.11	0.27	0.13	0.18	-0.13	0.07
	ZS + Task	0.17	0.19	0.10	0.28	0.11	0.00	0.11	0.15	0.22	0.10	0.06	0.25	0.15	0.19	0.21	0.16	0.20	0.06	0.19	0.16	0.15
	FS + Task	-0.01	-0.04	-0.04	0.08	0.00	0.00	-0.02	-0.01	0.01	-0.06	-0.04	0.00	-0.02	0.06	0.04	-0.01	-0.01	-0.04	-0.03	-0.01	-0.01
	CoT	0.15	0.06	0.06	0.39	-0.04	0.00	-0.03	-0.04	0.15	0.14	-0.01	0.15	0.00	0.06	0.08	0.08	0.12	0.09	0.09	0.21	0.09
Average		0.09	0.10	0.08	0.20	0.03	0.00	0.04	0.16	0.07	0.05	-0.01	0.11	-0.03	0.09	0.09	0.09	0.14	0.06	0.11	0.06	0.08
Llama3.1 8B	ZS	0.30	-0.16	0.16	0.17	0.25	0.00	0.10	0.40	-0.28	-0.08	0.10	0.09	-0.07	-0.04	0.16	-0.10	0.32	0.40	0.04	-0.53	0.06
	ZS + Task	0.12	-0.06	0.11	0.33	0.11	0.00	0.08	-0.04	-0.28	0.13	-0.75	0.02	-0.12	0.17	0.20	0.01	-0.01	0.25	0.14	-0.03	0.02
	FS + Task	0.10	-0.06	-0.19	0.05	-0.15	0.00	0.03	-0.07	0.10	0.00	-0.05	-0.07	-0.24	-0.03	-0.05	0.03	-0.09	0.06	0.09	-0.07	-0.03
	CoT	0.13	0.03	0.10	0.16	0.00	0.00	0.03	0.06	-0.37	0.08	-0.02	0.00	-0.05	0.08	0.20	0.32	-0.09	0.11	0.04	0.07	0.04
Average		0.16	-0.06	0.05	0.18	0.05	0.00	0.06	0.09	-0.21	0.04	-0.18	0.01	-0.12	0.05	0.13	0.06	0.03	0.20	0.08	-0.14	0.02
Llama3.1 70B	ZS	0.13	0.03	0.14	0.05	0.04	0.00	0.03	0.30	0.05	0.06	-0.08	-0.10	0.15	0.06	0.15	0.04	0.07	0.13	0.06	-0.26	0.05
	ZS + Task	0.18	0.01	0.13	0.18	0.15	0.00	0.09	0.25	0.05	0.17	0.03	0.14	0.06	0.15	0.18	0.04	0.07	0.20	0.08	0.07	0.11
	FS + Task	-0.43	-0.23	-0.11	-0.13	-0.70	0.00	-0.05	-0.22	-0.22	-0.23	-0.01	-0.16	-0.04	-0.03	-0.15	-0.14	-0.15	-0.24	-0.24	-0.08	-0.18
	CoT	-0.15	0.42	-0.02	0.05	0.03	0.00	0.00	0.04	0.03	0.01	-0.21	0.00	0.19	0.07	-0.03	0.01	-0.10	0.08	-0.07	-0.02	0.02
Average		-0.07	0.05	0.03	0.04	-0.12	0.00	0.02	0.09	-0.02	0.00	-0.07	-0.03	0.09	0.06	0.04	-0.01	-0.02	0.04	-0.04	-0.07	0.00

Table 23: Difference in TNR when using target-language instructions versus English in monolingual settings. Positive values (green) reflect improved filtering when using target-language instructions. The table presents results for six models and four prompting techniques, with average values computed for each model–prompting combination as well as for each language pair.

Model	Technique	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Qwen3 8B	ZS	0.78	0.88	0.89	0.70	0.71	0.85	0.90	0.68	0.70	0.60	0.70	0.69	0.59	0.68	0.77	0.84	0.76	0.93	0.81	0.84	0.76
	ZS + Task	0.86	0.96	0.96	0.80	0.79	0.89	0.95	0.83	0.71	0.66	0.73	0.66	0.76	0.81	0.91	0.81	0.98	0.97	0.87	0.87	0.84
	FS + Task	0.73	0.92	0.97	0.71	0.67	0.78	0.91	0.77	0.76	0.58	0.65	0.60	0.80	0.70	0.88	0.71	1.00	0.98	0.80	0.75	0.78
	CoT	0.94	0.99	0.98	0.93	0.92	0.98	0.97	0.90	0.92	0.83	0.89	0.89	0.90	0.92	0.96	0.90	0.98	0.98	0.92	0.93	0.93
Average		0.83	0.94	0.95	0.79	0.77	0.88	0.93	0.80	0.77	0.67	0.74	0.71	0.76	0.78	0.88	0.82	0.93	0.97	0.85	0.85	0.83
Qwen3 14B	ZS	0.86	0.91	0.91	0.84	0.79	0.87	0.92	0.81	0.64	0.61	0.63	0.69	0.65	0.72	0.85	0.82	0.89	0.96	0.81	0.83	0.80
	ZS + Task	0.86	0.94	0.94	0.80	0.71	0.88	0.97	0.81	0.65	0.58	0.67	0.64	0.70	0.83	0.94	0.77	0.98	0.97	0.80	0.80	0.81
	FS + Task	0.80	0.97	0.98	0.79	0.75	0.87	0.97	0.85	0.69	0.60	0.67	0.63	0.72	0.84	0.94	0.73	0.96	0.99	0.83	0.80	0.82
	CoT	0.96	1.00	0.98	0.94	0.93	0.96	0.99	0.93	0.90	0.87	0.90	0.88	0.90	0.92	0.97	0.90	0.99	0.99	0.91	0.91	0.94
Average		0.87	0.96	0.95	0.84	0.80	0.90	0.96	0.85	0.72	0.67	0.72	0.71	0.74	0.83	0.93	0.81	0.96	0.98	0.84	0.84	0.84
Gemma3 12B	ZS	0.67	0.50	0.83	0.58	0.45	0.68	0.79	0.41	0.35	0.34	0.34	0.38	0.53	0.42	0.75	0.50	0.78	0.92	0.62	0.76	0.58
	ZS + Task	0.58	0.56	0.82	0.49	0.45	0.65	0.77	0.44	0.39	0.35	0.37	0.34	0.52	0.40	0.75	0.47	0.85	0.89	0.62	0.70	0.57
	FS + Task	0.84	0.98	0.98	0.84	0.82	0.91	0.92	0.85	0.81	0.64	0.81	0.74	0.79	0.84	0.98	0.86	1.00	0.99	0.88	0.90	0.87
	CoT	0.66	0.62	0.80	0.56	0.49	0.66	0.75	0.39	0.34	0.30	0.34	0.37	0.48	0.30	0.66	0.51	0.76	0.88	0.53	0.57	0.55
Average		0.69	0.67	0.86	0.62	0.55	0.73	0.81	0.52	0.47	0.41	0.47	0.46	0.58	0.49	0.79	0.59	0.85	0.92	0.66	0.73	0.64
Gemma3 27B	ZS	0.72	0.61	0.81	0.59	0.54	0.68	0.80	0.44	0.44	0.35	0.32	0.43	0.52	0.37	0.87	0.52	0.86	0.90	0.58	0.62	0.60
	ZS + Task	0.63	0.55	0.77	0.49	0.40	0.55	0.79	0.32	0.36	0.30	0.28	0.31	0.45	0.35	0.78	0.42	0.89	0.88	0.59	0.56	0.53
	FS + Task	0.83	0.93	0.96	0.68	0.62	0.81	0.93	0.74	0.67	0.52	0.68	0.60	0.70	0.75	0.97	0.74	0.96	0.98	0.81	0.86	0.79
	CoT	0.62	0.53	0.75	0.48	0.38	0.55	0.73	0.27	0.29	0.24	0.24	0.26	0.43	0.17	0.77	0.40	0.79	0.85	0.51	0.49	0.49
Average		0.70	0.66	0.82	0.56	0.49	0.65	0.81	0.44	0.44	0.35	0.38	0.40	0.53	0.41	0.85	0.52	0.88	0.90	0.62	0.63	0.60
Llama3.1 8B	ZS	0.59	0.66	0.61	0.61	0.46	0.56	0.74	0.37	0.26	0.33	0.24	0.30	0.21	0.23	0.47	0.47	0.59	0.80	0.41	0.62	0.48
	ZS + Task	0.82	0.92	0.88	0.84	0.75	0.90	0.91	0.75	0.59	0.65	0.63	0.65	0.70	0.76	0.85	0.79	0.98	0.94	0.83	0.86	0.80
	FS + Task	0.83	0.90	0.92	0.81	0.83	0.91	0.88	0.86	0.87	0.86	0.88	0.83	0.88	0.88	0.93	0.87	0.95	0.88	0.90	0.92	0.88
	CoT	0.77	0.75	0.82	0.72	0.71	0.80	0.78	0.63	0.61	0.62	0.65	0.64	0.55	0.60	0.74	0.71	0.90	0.84	0.73	0.83	0.72
Average		0.75	0.81	0.81	0.75	0.69	0.79	0.83	0.65	0.58	0.62	0.60	0.61	0.59	0.62	0.75	0.71	0.86	0.87	0.72	0.81	0.72
Llama3.1 70B	ZS	0.84	0.89	0.91	0.80	0.78	0.88	0.87	0.68	0.70	0.70	0.67	0.63	0.65	0.68	0.86	0.79	0.94	0.93	0.80	0.94	0.80
	ZS + Task	0.78	0.90	0.92	0.77	0.69	0.84	0.92	0.64	0.67	0.61	0.68	0.58	0.67	0.71	0.90	0.73	0.99	0.95	0.81	0.91	0.78
	FS + Task	0.91	0.98	0.98	0.95	0.93	0.98	0.98	0.93	0.94	0.88	0.98	0.94	0.95	0.98	0.98	0.94	0.98	0.97	0.97	1.00	0.96
	CoT	0.81	0.91	0.90	0.82	0.74	0.88	0.89	0.71	0.83	0.64	0.76	0.73	0.82	0.78	0.84	0.82	0.89	0.92	0.87	0.90	0.82
Average		0.84	0.92	0.93	0.84	0.79	0.90	0.92	0.74	0.79	0.71	0.77	0.72	0.77	0.79	0.90	0.82	0.95	0.94	0.86	0.94	0.84

Table 24: The capabilities of LLMs in filtering irrelevant pairs based on TNR (higher is better) using English instruction in cross-lingual settings. The average TNR is calculated across all languages for each model and technique.

Model	Technique	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Qwen3 8B	ZS	-0.05	-0.04	0.00	0.21	0.08	0.00	-0.03	0.11	0.09	0.04	-0.04	0.00	0.02	-0.05	-0.41	-0.40	-0.39	0.03	-0.13	-0.46	-0.07
	ZS + Task	0.01	0.00	0.00	0.02	0.13	0.00	0.03	0.07	0.01	-0.02	0.07	0.02	0.05	0.11	-0.14	-0.10	-0.14	-0.01	-0.16	-0.14	-0.01
	FS + Task	0.10	0.08	0.00	0.17	0.14	0.00	0.05	0.13	-0.08	-0.08	0.13	-0.12	-0.16	0.19	0.07	-0.19	-0.41	-0.02	0.03	0.16	0.01
	CoT	0.00	-0.08	0.00	-0.06	-0.01	0.00	-0.02	0.04	-0.01	-0.03	-0.01	-0.20	-0.12	-0.05	-0.18	-0.25	-0.42	0.01	-0.07	-0.10	-0.08
Average		0.01	-0.01	0.00	0.09	0.08	0.00	0.01	0.08	0.00	-0.02	0.04	-0.07	-0.05	0.05	-0.16	-0.24	-0.34	0.00	-0.08	-0.13	-0.04
Qwen3 14B	ZS	-0.05	-0.34	0.00	-0.03	-0.05	0.00	-0.13	-0.05	-0.07	-0.18	-0.33	-0.10	-0.02	-0.06	-0.18	-0.45	-0.54	0.02	-0.05	-0.46	-0.15
	ZS + Task	-0.01	0.04	0.00	0.01	0.03	0.00	-0.01	0.05	0.11	0.03	0.06	0.03	0.07	0.03	-0.03	-0.01	-0.11	0.01	0.01	-0.01	0.01
	FS + Task	0.04	0.00	0.00	-0.01	-0.02	0.00	0.00	-0.02	-0.01	0.01	0.05	0.01	0.00	-0.02	0.00	0.04	-0.03	0.00	-0.01	0.02	0.00
	CoT	-0.05	-0.02	0.00	-0.07	-0.06	0.00	-0.03	-0.01	0.01	-0.05	-0.05	-0.07	-0.03	-0.04	-0.61	-0.14	-0.23	-0.02	-0.01	-0.06	-0.08
Average		-0.02	-0.08	0.00	-0.02	-0.03	0.00	-0.04	-0.01	0.01	-0.05	-0.07	-0.03	0.00	-0.02	-0.20	-0.14	-0.23	0.00	-0.02	-0.13	-0.05
Gemma3 12B	ZS	0.01	-0.18	0.00	0.02	-0.06	0.00	-0.01	-0.07	0.18	0.07	0.00	0.10	0.08	0.07	-0.41	-0.28	-0.51	-0.02	-0.18	-0.40	-0.08
	ZS + Task	0.04	0.18	0.00	0.05	0.02	0.00	0.01	-0.02	0.15	0.10	0.18	0.03	0.00	0.27	0.07	-0.01	-0.13	0.07	-0.04	0.01	0.05
	FS + Task	0.00	0.00	0.00	-0.03	-0.10	0.00	0.00	-0.03	-0.12	-0.08	0.03	-0.10	-0.15	0.04	-0.01	-0.10	-0.02	-0.02	-0.08	0.02	-0.04
	CoT	0.00	0.06	0.00	-0.05	0.16	0.00	0.01	0.15	0.09	0.09	0.08	-0.05	-0.04	0.13	0.00	-0.22	-0.43	0.02	0.10	0.02	0.01
Average		0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.07	0.05	0.07	-0.01	-0.03	0.13	-0.09	-0.15	-0.27	0.01	-0.05	-0.09	-0.02
Gemma3 27B	ZS	0.03	-0.09	0.00	-0.02	-0.05	0.00	-0.03	-0.06	0.11	0.04	-0.06	0.05	0.04	0.01	-0.18	-0.05	-0.49	-0.04	-0.15	-0.03	-0.05
	ZS + Task	0.03	0.20	0.00	0.03	0.13	0.00	0.05	0.12	0.00	-0.02	0.14	0.03	0.04	0.03	0.07	0.03	0.03	0.05	-0.03	0.12	0.05
	FS + Task	-0.06	-0.01	0.00	-0.02	0.00	0.00	-0.02	-0.08	-0.07	-0.04	0.03	-0.09	-0.02	0.03	0.00	-0.08	0.04	-0.03	-0.06	0.01	-0.02
	CoT	0.01	0.17	0.00	-0.06	0.21	0.00	0.04	0.25	0.04	0.03	0.07	0.02	-0.04	0.04	0.03	-0.04	-0.02	0.04	0.11	0.10	0.05
Average		0.00	0.07	0.00	-0.02	0.07	0.00	0.01	0.06	0.02	0.00	0.04	0.00	0.00	0.03	-0.02	-0.03	-0.11	0.01	-0.03	0.05	0.01
Llama3.1 8B	ZS	-0.07	-0.15	0.00	-0.01	0.00	0.00	-0.24	0.13	0.46	0.20	0.17	0.04	0.09	0.00	-0.06	0.02	-0.24	0.13	-0.15	-0.53	-0.01
	ZS + Task	0.02	-0.30	0.00	0.07	0.18	0.00	0.01	0.19	0.36	0.24	0.13	-0.02	0.02	0.15	-0.20	-0.75	-0.11	0.05	0.06	-0.04	0.00
	FS + Task	0.06	0.09	0.00	0.03	0.06	0.00	0.04	0.06	0.09	0.05	0.03	-0.21	-0.16	0.00	-0.01	-0.01	-0.11	0.11	0.04	-0.03	0.01
	CoT	-0.13	-0.37	0.00	-0.10	0.08	0.00	-0.13	0.11	0.23	0.09	0.03	-0.09	0.01	0.07	-0.28	-0.22	-0.17	0.11	0.02	-0.06	-0.04
Average		-0.03	-0.18	0.00	0.00	0.08	0.00	-0.08	0.12	0.29	0.15	0.09	-0.07	-0.01	0.06	-0.14	-0.24	-0.16	0.10	-0.01	-0.16	-0.01
Llama3.1 70B	ZS	-0.02	0.06	0.00	-0.03	-0.06	0.00	0.01	0.00	0.23	0.07	0.07	0.15	0.23	0.00	-0.16	-0.09	0.06	0.05	0.05	-0.27	0.02
	ZS + Task	0.02	0.05	0.00	0.07	0.09	0.00	0.02	0.12	0.24	0.19	0.14	0.10	0.11	0.05	-0.03	0.01	0.01	0.05	0.14	0.09	0.07
	FS + Task	-0.16	-0.18	0.00	-0.03	-0.11	0.00	-0.21	-0.08	-0.29	-0.33	-0.12	-0.16	-0.11	-0.05	-0.20	-0.01	-0.03	-0.42	-0.25	-0.05	-0.14
	CoT	-0.10	-0.03	0.00	-0.08	0.08	0.00	-0.08	0.14	0.08	0.18	-0.01	-0.05	0.00	0.06	-0.07	-0.23	0.08	0.01	0.01	0.06	0.00
Average		-0.06	-0.03	0.00	-0.02	0.00	0.00	-0.06	0.04	0.07	0.03	0.02	0.01	0.06	0.02	-0.12	-0.08	0.03	-0.08	-0.01	-0.04	-0.01

Table 25: Difference in TNR when using post-language instructions versus English in cross-lingual settings. Each column represents a language pair (post–claim language), with the first language indicating the instruction language. Positive values (green) reflect improved filtering when using post-language instructions. The table presents results for six models and four prompting techniques, with average values computed for each model–prompting combination as well as for each language pair.

Model	Technique	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Qwen3 8B	ZS	0.00	0.00	-0.02	0.00	0.00	-0.03	0.03	0.24	0.03	0.00	-0.04	0.00	0.00	0.12	-0.04	0.00	-0.32	0.04	-0.12	-0.08	-0.01
	ZS + Task	0.00	0.00	-0.05	0.00	0.00	-0.07	-0.02	-0.02	0.08	0.00	0.01	0.00	0.03	0.09	-0.06	0.00	-0.14	-0.02	0.01	-0.10	-0.01
	FS + Task	0.00	0.00	0.02	0.00	0.00	0.13	0.03	0.15	-0.19	0.00	0.15	0.00	-0.03	0.13	-0.05	0.00	0.00	0.01	0.00	0.18	0.03
	CoT	0.00	0.00	-0.01	0.00	0.00	-0.03	-0.03	-0.07	-0.10	0.00	-0.07	0.00	-0.04	0.00	0.02	0.00	-0.20	-0.02	0.01	-0.05	-0.03
Average		0.00	0.00	-0.02	0.00	0.00	0.00	0.00	0.07	-0.05	0.00	0.02	0.00	-0.01	0.09	-0.03	0.00	-0.17	0.00	-0.02	-0.01	-0.01
Qwen3 14B	ZS	0.00	0.00	0.01	0.00	0.00	-0.11	-0.12	-0.07	0.06	0.00	0.14	0.00	-0.30	0.07	0.02	0.00	-0.21	-0.04	-0.23	0.00	-0.04
	ZS + Task	0.00	0.00	0.03	0.00	0.00	-0.01	-0.01	0.02	0.14	0.00	0.06	0.00	0.07	0.02	0.01	0.00	-0.02	0.01	0.07	0.03	0.02
	FS + Task	0.00	0.00	-0.01	0.00	0.00	-0.05	-0.03	-0.02	-0.03	0.00	0.02	0.00	0.07	-0.11	0.02	0.00	-0.03	0.00	0.03	0.01	-0.01
	CoT	0.00	0.00	-0.03	0.00	0.00	-0.04	-0.04	-0.07	-0.04	0.00	-0.06	0.00	-0.04	-0.01	-0.02	0.00	-0.15	-0.01	-0.01	-0.01	-0.03
Average		0.00	0.00	0.00	0.00	0.00	-0.05	-0.05	-0.03	0.03	0.00	0.04	0.00	-0.05	-0.01	0.00	0.00	-0.10	-0.01	-0.04	0.01	-0.01
Gemma3 12B	ZS	0.00	0.00	-0.02	0.00	0.00	0.03	-0.03	0.05	0.18	0.00	0.12	0.00	-0.08	-0.12	0.04	0.00	-0.56	-0.02	-0.08	0.02	-0.02
	ZS + Task	0.00	0.00	0.09	0.00	0.00	0.03	0.05	-0.01	0.03	0.00	0.08	0.00	0.18	0.00	0.12	0.00	0.02	-0.02	0.15	-0.02	0.04
	FS + Task	0.00	0.00	-0.01	0.00	0.00	-0.06	-0.02	-0.14	-0.13	0.00	-0.03	0.00	-0.05	-0.11	-0.01	0.00	0.00	-0.02	-0.01	-0.07	-0.03
	CoT	0.00	0.00	0.07	0.00	0.00	-0.10	-0.07	0.01	0.06	0.00	0.18	0.00	0.03	0.11	0.12	0.00	-0.07	-0.03	0.04	-0.05	0.01
Average		0.00	0.00	0.03	0.00	0.00	-0.02	-0.02	-0.03	0.04	0.00	0.09	0.00	0.02	-0.03	0.07	0.00	-0.15	-0.02	0.03	-0.03	0.00
Gemma3 27B	ZS	0.00	0.00	-0.01	0.00	0.00	0.06	0.02	0.19	0.17	0.00	0.38	0.00	-0.09	0.01	0.01	0.00	-0.11	-0.01	-0.03	0.16	0.04
	ZS + Task	0.00	0.00	0.09	0.00	0.00	0.08	0.04	0.16	0.10	0.00	0.10	0.00	0.16	0.18	0.08	0.00	0.04	0.04	0.12	0.01	0.06
	FS + Task	0.00	0.00	-0.01	0.00	0.00	0.01	0.01	-0.07	0.00	0.00	0.01	0.00	0.03	0.04	-0.01	0.00	0.02	-0.02	-0.01	-0.01	0.00
	CoT	0.00	0.00	0.12	0.00	0.00	0.06	0.05	0.02	0.08	0.00	-0.01	0.00	0.05	0.28	0.06	0.00	0.03	-0.11	0.04	0.01	0.03
Average		0.00	0.00	0.05	0.00	0.00	0.05	0.03	0.07	0.09	0.00	0.12	0.00	0.04	0.13	0.04	0.00	-0.01	-0.02	0.03	0.04	0.03
Llama3.1 8B	ZS	0.00	0.00	0.14	0.00	0.00	-0.27	-0.50	0.14	0.20	0.00	0.49	0.00	0.12	0.17	0.33	0.00	-0.32	0.01	0.08	-0.24	0.02
	ZS + Task	0.00	0.00	0.07	0.00	0.00	-0.09	-0.02	0.13	0.16	0.00	-0.12	0.00	0.06	0.16	0.09	0.00	-0.30	0.04	0.04	-0.11	0.01
	FS + Task	0.00	0.00	0.06	0.00	0.00	0.01	0.05	-0.06	-0.15	0.00	-0.13	0.00	-0.01	0.03	0.03	0.00	-0.05	0.04	-0.04	0.02	-0.01
	CoT	0.00	0.00	0.08	0.00	0.00	0.12	0.13	-0.01	0.04	0.00	-0.08	0.00	0.24	0.09	0.15	0.00	-0.32	0.02	0.19	0.11	0.04
Average		0.00	0.00	0.09	0.00	0.00	-0.06	-0.08	0.05	0.06	0.00	0.04	0.00	0.11	0.11	0.15	0.00	-0.25	0.03	0.07	-0.06	0.01
Llama3.1 70B	ZS	0.00	0.00	0.05	0.00	0.00	0.01	0.03	-0.05	0.19	0.00	0.30	0.00	0.18	-0.03	0.08	0.00	-0.24	-0.04	0.03	0.01	0.03
	ZS + Task	0.00	0.00	0.08	0.00	0.00	0.03	0.00	0.17	0.15	0.00	0.21	0.00	0.15	0.13	0.09	0.00	-0.24	0.02	0.10	0.04	0.05
	FS + Task	0.00	0.00	-0.33	0.00	0.00	-0.04	-0.10	-0.05	-0.16	0.00	-0.18	0.00	-0.10	-0.13	-0.32	0.00	-0.18	-0.08	-0.08	-0.08	-0.09
	CoT	0.00	0.00	0.01	0.00	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00	-0.05	0.08	-0.10	0.00	-0.32	-0.03	0.02	0.04	-0.02
Average		0.00	0.00	-0.05	0.00	0.00	0.00	-0.02	0.02	0.04	0.00	0.08	0.00	0.05	0.01	-0.06	0.00	-0.24	-0.03	0.02	0.00	-0.01

Table 26: Difference in TNR when using claim-language instructions versus English in cross-lingual settings. Each column represents a language pair (post–claim language), with the second language indicating the instruction language. Positive values (green) reflect improved filtering when using claim-language instructions. The table presents results for six models and four prompting techniques, with average values computed for each model–prompting combination as well as for each language pair.

Technique	Thinking	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Zero-Shot	✗	0.79	0.85	0.68	0.58	0.84	0.76	0.80	0.61	0.76	0.84	0.90	0.66	0.78	0.74	0.61	0.64	0.90	0.78	0.71	0.86	0.75
	✓	0.85	0.97	0.78	0.70	0.86	0.76	0.86	0.61	0.87	0.85	0.94	0.86	0.85	0.79	0.72	0.67	0.86	0.84	0.78	0.94	0.82
Zero-Shot + Task Description	✗	0.91	0.95	0.73	0.77	0.86	0.72	0.83	0.65	0.82	0.87	0.86	0.67	0.94	0.81	0.68	0.77	0.84	0.84	0.80	0.89	0.81
	✓	0.93	0.99	0.90	0.86	0.92	0.87	0.92	0.79	0.92	0.91	0.95	0.92	0.94	0.89	0.82	0.85	0.95	0.93	0.88	0.95	0.90
Few-Shot + Task Description	✗	0.86	0.91	0.71	0.62	0.81	0.57	0.87	0.68	0.81	0.75	0.85	0.59	0.92	0.68	0.62	0.67	0.82	0.73	0.75	0.79	0.75
	✓	0.97	0.99	0.91	0.91	0.96	0.88	0.93	0.83	0.97	0.93	0.96	0.92	0.95	0.95	0.87	0.86	0.95	0.93	0.94	0.95	0.93

Table 27: TNR using English for the instruction with (✓) and without (✗) thinking mode for the Qwen3 8B in amonolingual setting.

Technique	Thinking	ara	bul	ces	deu	ell	eng	fra	hbs	hin	hun	kor	msa	mya	nld	pol	por	ron	slk	spa	tha	Avg.
Zero-Shot	✗	-0.01	-0.10	0.01	0.17	-0.10	0.00	0.17	0.11	0.03	-0.05	-0.29	-0.13	-0.32	-0.11	0.02	0.12	0.04	0.07	-0.08	-0.35	-0.04
	✓	0.03	0.00	0.03	0.09	-0.07	0.00	0.01	0.02	0.03	0.03	0.00	-0.07	-0.28	-0.02	-0.03	-0.01	-0.01	0.05	-0.05	-0.05	-0.02
Zero-Shot + Task Description	✗	-0.02	0.00	0.05	0.14	-0.04	0.00	0.06	0.03	-0.05	-0.06	-0.09	-0.04	-0.18	0.10	0.09	-0.03	0.02	-0.01	0.09	-0.08	0.00
	✓	0.03	0.00	-0.01	0.05	-0.01	0.00	-0.01	-0.01	0.01	0.00	-0.01	-0.01	-0.17	0.03	0.09	-0.04	-0.03	0.00	0.02	0.01	0.00
Few-Shot + Task Description	✗	0.03	0.05	-0.08	0.19	-0.21	0.00	0.08	0.09	0.11	0.09	-0.10	0.30	-0.36	0.22	0.01	0.11	0.11	-0.02	0.16	0.15	0.05
	✓	0.01	0.01	0.00	0.02	0.00	0.00	-0.01	0.01	0.01	0.04	-0.01	0.01	-0.15	0.01	0.04	0.04	-0.01	0.01	0.03	-0.01	0.00

Table 28: Impact of the thinking mode on TNR in monolingual settings across three prompting techniques for the Qwen3 8B. The table shows the difference in TNR when using target language instructions versus English. Each row compares performance with (✓) and without (✗) "thinking mode", across 20 languages. Positive values (green) indicate improved performance with target language instructions. Average scores are reported in the final column.

Technique	Thinking	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Zero-Shot	✗	0.78	0.88	0.89	0.70	0.71	0.85	0.90	0.68	0.70	0.60	0.70	0.69	0.59	0.68	0.77	0.84	0.76	0.93	0.81	0.84	0.76
	✓	0.84	0.88	0.89	0.79	0.76	0.89	0.88	0.70	0.81	0.72	0.74	0.73	0.70	0.72	0.86	0.79	0.95	0.94	0.79	0.86	0.81
Zero-Shot + Task Description	✗	0.86	0.96	0.96	0.80	0.79	0.89	0.95	0.83	0.71	0.66	0.73	0.66	0.76	0.81	0.91	0.81	0.98	0.97	0.87	0.87	0.84
	✓	0.91	0.97	0.96	0.89	0.86	0.95	0.95	0.87	0.88	0.85	0.85	0.82	0.84	0.86	0.94	0.86	0.98	0.97	0.88	0.92	0.90
Few-Shot + Task Description	✗	0.73	0.92	0.97	0.71	0.67	0.78	0.91	0.77	0.76	0.58	0.65	0.60	0.80	0.70	0.88	0.71	1.00	0.98	0.80	0.75	0.78
	✓	0.92	0.98	0.97	0.90	0.88	0.95	0.97	0.91	0.92	0.83	0.86	0.83	0.88	0.92	0.96	0.88	0.96	0.98	0.92	0.90	0.91

Table 29: TNR using English for the instruction with (✓) and without (✗) thinking mode for the Qwen3 8B in a cross-lingual setting.

Technique	Thinking	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Zero-Shot	✗	-0.05	-0.04	0.00	0.21	0.08	0.00	-0.03	0.11	0.09	0.04	-0.04	0.00	0.02	-0.05	-0.41	-0.40	-0.39	0.03	-0.13	-0.46	-0.07
	✓	-0.05	0.04	0.00	-0.03	0.04	0.00	-0.04	0.04	0.00	-0.02	-0.02	-0.05	0.03	-0.01	-0.13	0.00	-0.43	0.01	-0.03	-0.14	-0.04
Zero-Shot + Task Description	✗	0.01	0.00	0.00	0.02	0.13	0.00	0.03	0.07	0.01	-0.02	0.07	0.02	0.05	0.11	-0.14	-0.10	-0.14	-0.01	-0.16	-0.14	-0.01
	✓	-0.02	0.01	0.00	-0.05	0.01	0.00	0.00	0.01	0.03	-0.05	0.04	-0.04	-0.02	0.05	-0.05	-0.08	-0.07	0.01	-0.05	0.01	-0.01
Few-Shot + Task Description	✗	0.10	0.08	0.00	0.17	0.14	0.00	0.05	0.13	-0.08	-0.08	0.13	-0.12	-0.16	0.19	0.07	-0.19	-0.41	-0.02	0.03	0.16	0.01
	✓	0.01	0.01	0.00	0.01	0.01	0.00	0.02	0.03	0.00	-0.01	0.05	-0.04	0.03	0.04	-0.05	-0.03	-0.11	0.02	0.01	-0.03	0.00

Table 30: Impact of the thinking mode on performance in cross-lingual settings across three prompting techniques for the `Qwen3 8B`. The table shows the difference in the TNR when using post-language instructions (the first language in the column name) versus English. Each row compares performance with (✓) and without (✗) "thinking mode", across 20 language pairs. Positive values (green) indicate improved performance with post-language instructions. Average scores are reported in the final column.

Technique	Thinking	spa-eng	hin-eng	eng-ara	fra-eng	deu-eng	eng-por	spa-por	deu-fra	slk-ces	slk-eng	pol-hbs	ces-eng	ces-pol	nld-deu	msa-ara	kor-eng	mya-msa	ara-fra	hun-pol	tha-por	Avg.
Zero-Shot	✗	0.00	0.00	-0.02	0.00	0.00	-0.03	0.03	0.24	0.03	0.00	-0.04	0.00	0.00	0.12	-0.04	0.00	-0.32	0.04	-0.12	-0.08	-0.01
	✓	0.00	0.00	0.01	0.00	0.00	-0.03	-0.01	0.02	0.01	0.00	-0.02	0.00	-0.02	0.05	0.00	0.00	-0.08	-0.01	0.00	-0.04	-0.01
Zero-Shot + Task Description	✗	0.00	0.00	-0.05	0.00	0.00	-0.07	-0.02	-0.02	0.08	0.00	0.01	0.00	0.03	0.09	-0.06	0.00	-0.14	-0.02	0.01	-0.10	-0.01
	✓	0.00	0.00	-0.01	0.00	0.00	-0.04	-0.03	-0.03	-0.02	0.00	-0.04	0.00	0.06	0.03	0.01	0.00	-0.02	0.00	0.04	-0.01	0.00
Few-Shot + Task Description	✗	0.00	0.00	0.02	0.00	0.00	0.13	0.03	0.15	-0.19	0.00	0.15	0.00	-0.03	0.13	-0.05	0.00	0.00	0.01	0.00	0.18	0.03
	✓	0.00	0.00	0.02	0.00	0.00	0.03	0.02	0.03	-0.04	0.00	0.02	0.00	0.03	0.01	0.00	0.00	-0.01	0.01	0.04	0.02	0.01

Table 31: Impact of the thinking mode on performance in cross-lingual settings across three prompting techniques for the `Qwen3 8B`. The table shows the difference in the TNR when using claim-language instructions (the second language in the column name) versus English. Each row compares performance with (✓) and without (✗) "thinking mode", across 20 language pairs. Positive values (green) indicate improved performance with claim-language instructions. Average scores are reported in the final column.

ited factual content.

Using `TurfTopic` (Kardos et al., 2025), we provide a further analysis on the topics of the posts, claims and top 1 ranked claims (both true positives and false positives). For this analysis we used the `Multilingual E5` and 1239 unique posts. For each post we analyzed the top-20 retrieved claims resulting in a total of 13.861 unique claims. We also extracted topics for the top-1 retrieved claim for every post, resulting in a set of 481 true positive (TP) claims and 758 false positive claims (FP).

Finally, to analyze the structure of the retrieved claim space, we encode each retrieved claim using the `paraphrase-multilingual-MiniLM-L12-v2` sentence encoder and project the resulting embeddings into two dimensions using t-SNE. Topic assignments are obtained using `TurfTopic` and used solely for visualization. Figure 7 shows that retrieved claims form dense, topically coherent clusters in the embedding space, with high-frequency topics such as COVID-19, geopolitics, and viral content occupying large and well-defined regions. This pronounced topical organization indicates that dense retrieval is driven primarily by topical similarity in the embedding space, increasing the likelihood that claims from dominant themes are repeatedly retrieved, even when they are not directly relevant to individual posts.

D Interaction Between Language Bias and Topic Concentration

Following the language resource taxonomy proposed by Joshi et al. (2020), we group languages into high-, mid-, and low-resource categories, since it is a common approach used by other studies as well. Then, using the `TurfTopic` modeling framework (Kardos et al., 2025), we analyze topic distributions separately for posts and fact-check claims based on the full AMC-16k dataset (Vykopal et al., 2025) that contains verified english translations for each multilingual post and claim. Tables 32 and 33 show that low-resource languages are substantially underrepresented in both posts and claims, reflecting well-known disparities in data availability.

This imbalance is further reflected in the topic concentration of the data. As shown in Table 35, posts originating from low-resource languages exhibit higher topic concentration than those from high- and mid-resource languages, with the most frequent topic accounting for 21.1% of posts compared to 12.1% in high-resource settings. A similar, though less pronounced, pattern is observed in the claims, where high-resource claims span a broader range of topics, while low-resource claims concentrate more heavily around a small set of dominant themes, primarily related to COVID-19 and health misinformation (Table 34).

Taken together, these results suggest that language bias and retrieval bias could interact through corpus structure: content in low-resource lan-

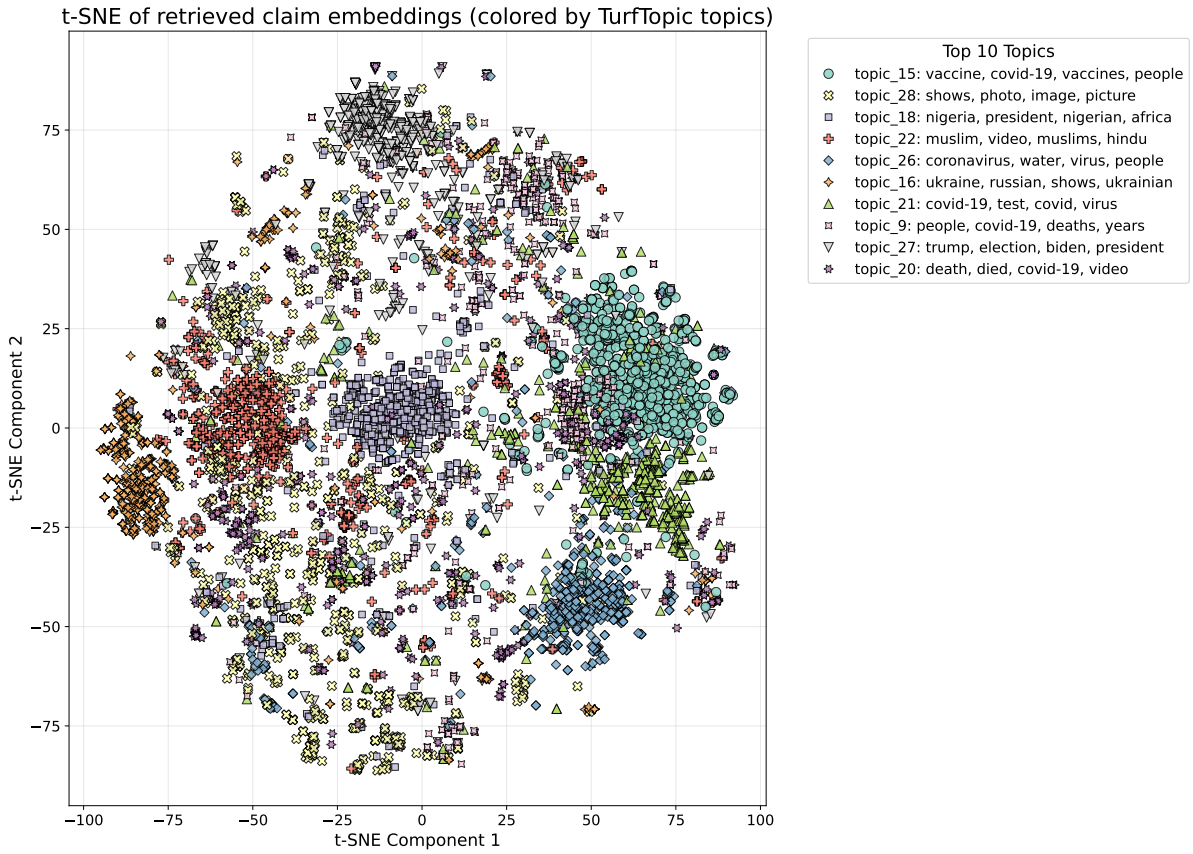

Figure 7: t-SNE projection of fact-check claim embeddings colored by their assigned topics from `TurfTopic`. The plot shows the 10 most frequent topics in the corpus. Each point represents a fact-checked claim, and clusters correspond to semantically coherent topic groupings. Despite overlapping content dimensions, visually separable clusters reflect the model's ability to distinguish major thematic areas such as health, geopolitics, and misinformation trends.

guages is not only scarcer but also topically narrower, making it more susceptible to topic concentration effects during retrieval. This provides a plausible mechanism by which dense retrieval systems may amplify existing language resource disparities, even when retrieval is performed using a single query language.

Resource Group	Languages (ISO 639-3)	Posts (N)	Claims (N)
High	English (eng), German (deu), Spanish (spa), French (fra), Portuguese (por), Arabic (ara), Hindi (hin)	6,400	6,791
Mid	Dutch (nld), Polish (pol), Czech (ces), Slovak (slk), Hungarian (hun), Greek (ell), Romanian (ron), Bulgarian (bul), Korean (kor), Thai (tha)	7,600	2,343
Low	Burmese (mya), Malay (msa), Serbo-Croatian (hbs)	2,000	1,056
Total	**20 languages**	**16,000**	**10,190**

Table 32: AMC-16k dataset composition of total languages, posts and claims; after grouping the languages in resource groups following (Joshi et al., 2020).

Language	eng	por	ara	fra	pol	msa	deu	hbs	spa	hin	ell	nld	ces	kor	slk	tha	ron	bul	hun	mya
# Claims	2651	1242	825	823	808	576	558	405	366	326	271	240	201	172	154	137	131	118	111	75
# Posts	111	91	71	69	61	60	56	55	55	55	50	48	45	43	42	40	40	40	40	40

Table 33: Language distribution for the unique posts and claims in the AMC-16k dataset corpus (ISO 639-3 codes).

Resource Group	Topic Keywords	Count	%
High	india, minister, muslims, people	586	8.6
	coronavirus, corona, virus, water	509	7.5
	photo, shows, photos, picture	439	6.5
	president, israeli, trump, against	428	6.3
	people, million, germany, died	421	6.2
Mid	ukraine, ukrainian, russian, shows	249	10.6
	people, million, germany, died	228	9.7
	covid-19, vaccine, vaccines, people	210	9.0
	covid-19, vaccines, vaccine, graphe	198	8.5
	photo, shows, photos, picture	145	6.2
Low	video, shows, covid-19, against	98	9.3
	photo, shows, photos, picture	97	9.2
	coronavirus, corona, virus, water	95	9.0
	covid-19, covid, virus, pandemic	94	8.9
	covid-19, vaccine, vaccines, people	79	7.5

Table 34: Top-5 topics per resource group based on the claims of the AMC-16k dataset.

Resource Group	Topic Keywords	Count	%
High	africans, africa, black, people	776	12.1
	japan, muslims, pakistan, india	508	7.9
	hospital, mother, make, says	470	7.3
	children, parents, must, other	451	7.0
	australia, australians, british, ja	431	6.7
Mid	vaccination, vaccine, mrna, vaccine	800	10.5
	ukraine, children, ukrainian, world	697	9.2
	children, parents, must, other	594	7.8
	pete, milk, cancer, children	509	6.7
	cough, sneezing, body, vaccine	491	6.5
Low	like, girls, army, flowers	421	21.1
	pete, milk, cancer, children	273	13.7
	japan, muslims, pakistan, india	153	7.6
	vaccination, vaccine, mrna, vaccine	134	6.7
	virus, fault, email, masks	123	6.2

Table 35: Top-5 topics per language resource group based on the posts of the AMC-16k dataset.