

AKCIT-FN at CheckThat! 2025: Switching Fine-Tuned SLMs and LLM Prompting for Multilingual Claim Normalization

Notebook for the CheckThat! Lab at CLEF 2025

Fabrycio Leite Nakano Almada^{1,2,†}, Kauan Divino Pouso Mariano^{1,2,†},
Maykon Adriell Dutra^{1,2,†}, Victor Emanuel da Silva Monteiro^{1,2,†},
Juliana Resplande Sant’Anna Gomes^{1,2,*}, Arlindo Rodrigues Galvão Filho^{1,2} and
Anderson da Silva Soares^{1,2}

¹*Institute of Informatics, Federal University of Goiás, Brazil*

²*Advanced Knowledge Center in Immersive Technology (AKCIT), Federal University of Goiás, Brazil*

Abstract

Claim normalization, the transformation of informal social media posts into concise, self-contained statements, is a crucial step in automated fact-checking pipelines. This paper details our submission to the CLEF-2025 CheckThat! Task 2, which challenges systems to perform claim normalization across twenty languages, divided into thirteen supervised (high-resource) and seven zero-shot (no training data) tracks.

Our approach, leveraging fine-tuned Small Language Models (SLMs) for supervised languages and Large Language Model (LLM) prompting for zero-shot scenarios, achieved podium positions (top three) in fifteen of the twenty languages. Notably, this included second-place rankings in eight languages, five of which were among the seven designated zero-shot languages, underscoring the effectiveness of our LLM-based zero-shot strategy. For Portuguese, our initial development language, our system achieved an average METEOR score of 0.5290, ranking third. All implementation artifacts, including inference, training, evaluation scripts, and prompt configurations, are publicly available at https://github.com/ju-resplande/checkthat2025_normalization.

Keywords

Claim Normalization, Disinformation, Multilingual NLP, Fact-Checking, Transformer Models, Zero-Shot Learning

1. Introduction

The proliferation of misinformation within social media ecosystems has intensified the demand for automated fact-checking pipelines that operate effectively across diverse languages, genres, and text characterized by significant noise. A crucial stage in such pipelines is claim normalization: the process of converting an informal, often multi-sentence social media post into a concise, self-contained statement suitable for subsequent evidence retrieval and veracity assessment. Without normalization, downstream modules must contend with extraneous elements such as redundancy, hashtags, emojis, and idiosyncratic phrasing, which collectively diminish both retrieval recall and factual accuracy [1].

The CLEF-2025 CheckThat! Lab confronts this bottleneck through Task 2 – Claim Normalization. Systems are required to generate normalized claims for twenty languages under two experimental conditions: (i) a monolingual setting, providing training and development splits (annotated examples)

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

[†]These authors contributed equally.

✉ fabrycio@egresso.ufg.br (F. L. N. Almada); kauan@discente.ufg.br (K. D. P. Mariano); maykonadriell@discente.ufg.br (M. A. Dutra); victor_emanuel@discente.ufg.br (V. E. d. S. Monteiro); juliana.resplande@discente.ufg.br (J. R. S. Gomes); arlindogalvao@ufg.br (A. R. G. Filho); andersonsoares@ufg.br (A. d. S. Soares)

ORCID 0000-0003-3876-9635 (F. L. N. Almada); 0009-0008-9082-0876 (K. D. P. Mariano); 0009-0000-0813-3084 (M. A. Dutra); 0009-0008-9059-6843 (V. E. d. S. Monteiro); 0000-0001-6900-1931 (J. R. S. Gomes); 0000-0003-2151-8039 (A. R. G. Filho); 0000-0002-2967-6077 (A. d. S. Soares)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

for thirteen higher-resource languages, and (ii) a zero-shot setting, releasing only test data for seven lower-resource languages, for which no specific training data is provided [2, 3, 4].

Our methodological development initially focused on Portuguese—the native language of our development team, facilitating a more nuanced understanding and iterative refinement—before extending validated strategies to all target languages. For languages in category (i), our experimentation involved fine-tuning open-source Encoder-Decoder Small Language Models (SLMs) and, as a comparative approach, inference via prompting with Large Language Models (LLMs). For languages in category (ii), we exclusively employed a zero-shot prompting strategy with LLMs.

Our efforts culminated in strong results, securing third place in the Portuguese subset with an average METEOR score of 0.5290. Across all languages, we achieved top-three placements in fifteen of the twenty languages. This included second-place finishes in eight languages overall (with five of these being zero-shot languages) and third-place finishes in seven languages, highlighting the robustness of our approach in both data-rich and data-scarce scenarios.

2. Related Work

Over the past few years, claim normalization has gained prominence as a crucial preprocessing step in automated fact-checking, moving beyond mere claim extraction. Konstantinovskiy et al. proposed the first annotation schema for claim detection informed by experts and a benchmark for automated claim detection that is more consistent across time, topics, and annotators than previous approaches. Sundriyal et al. formalized claim normalization by converting informal social-media texts into self-contained claims, demonstrating significant improvements in downstream evidence retrieval and veracity classification tasks.

Previous editions of the CLEF CheckThat! Lab (2018–2024) have included tasks such as check-worthiness detection from political debates or speeches [6, 7, 8, 9], and from political or COVID-19-related tweets [8, 9].

The construction of multilingual fact-checking corpora represents another active research direction. For instance, MuMiN automatically links 21 million tweets to 13 thousand fact-checked claims across 41 languages using LaBSE embeddings [10]. Similarly, Singh et al. created MMTweets by scraping “debunked narratives,” retrieving tweets via multilingual keyword queries, and applying detailed human annotation, resulting in a multimodal dataset for cross-lingual retrieval.

3. Task Definition

Task 2 of the CLEF-2025 CheckThat! Lab presents a comprehensive multilingual claim normalization challenge designed to evaluate system performance across diverse linguistic contexts and resource availability scenarios. The core objective involves transforming informal social media posts into concise, self-contained, and verifiable statements while preserving all factual content and systematically removing subjective opinions, redundant expressions, and extraneous material [2, 3, 4].

The task employs two experimental paradigms designed to assess system robustness under varying data availability conditions:

- a. **Monolingual setting:** Complete training, development, and test datasets are provided for thirteen languages: Arabic (AR), English (EN), French (FR), German (DE), Hindi (HI), Indonesian (ID), Marathi (MR), Polish (PL), Portuguese (PT), Punjabi (PA), Spanish (ES), Tamil (TA), and Thai (TH).
- b. **Zero-shot setting:** For seven languages, Only test splits are released for seven languages, requiring systems to generalize from cross-lingual knowledge or leverage multilingual pre-training: Bengali (BN), Czech (CS), Dutch (NL), Greek (EL), Korean (KO), Romanian (RO), and Telugu (TE).

Table 1 illustrates the complexity of this transformation through an English example that demonstrates the typical challenges: redundancy removal, formalization of informal language, and extraction of verifiable factual content from noisy social media text.

Table 1

Example of claim normalization from an informal social media post (English dataset), demonstrating redundancy removal, formalization, and extraction of verifiable facts.

Original Post	Normalized Claim
I guess the left is okay with this I guess the left is okay with this I guess the left is okay with this Dr. Rachel Levine @DrRachel Levine Thank you Vanity Fair for honoring me on the cover of your magazine this March. My dream of becoming @POTUS one day just took a step forward. THE SKY THE LIMIT. "Madam President Levine A LEADER IN THE MAKING 8:12 AM Feb 1, 2021 Twitter Web App	US assistant health secretary Rachel Levine appears on the cover of Vanity Fair's March 2021 issue

Table 2 provides a comprehensive overview of dataset statistics, revealing significant variation in resource availability across languages. This heterogeneity presents both opportunities and challenges: while resource-rich languages like English provide substantial training data (11,374 examples), smaller datasets for languages like Tamil (102 training examples) require careful consideration of overfitting risks and generalization strategies.

Table 2

Dataset statistics for Task 2: Claim Normalization, showing sample distribution across languages and splits.

(a) Monolingual setting				(b) Zero-shot setting	
Language	Train	Dev	Test	Language	Test
English	11374	1171	1285	Korean	274
Spanish	3458	439	439	Dutch	177
Portuguese	1735	223	225	Greek	156
French	1174	147	148	Romanian	141
Hindi	1081	50	100	Czech	123
Indonesian	540	137	100	Telugu	116
Arabic	470	118	100	Bengali	81
Punjabi	445	50	100		
German	386	101	100		
Thai	244	61	100		
Polish	163	41	100		
Marathi	137	50	100		
Tamil	102	50	100		

Participant submissions were evaluated using the METEOR score [12], a metric commonly employed for machine translation evaluation that assesses translation quality by aligning system output with reference texts based on exact word matches, stemming, and synonymy. For this task, official scores were calculated by averaging METEOR results across all test examples for a given language/setting. Punctuation was removed during pre-processing for the evaluation script to standardize inputs and mitigate its impact on scores.

4. Methodology

To address the claim normalization task, we employed a dual-strategy approach tailored to the different experimental settings. For the monolingual setting, where training, development, and test splits were

available, we investigated both: (i) fine-tuning of open-source Encoder-Decoder Small Language Models (SLMs), and (ii) inference using Large Language Models (LLMs) with few-shot prompting. For the zero-shot setting, characterized by the absence of training data for specific languages, our efforts exclusively focused on zero-shot inference with LLMs (using prompts without in-context examples specific to the task for those languages).

Initial experimentation, including preliminary model selection, hyperparameter tuning, and qualitative analysis, was conducted on the Portuguese dataset. This choice was motivated by the team’s native proficiency in the language, facilitating a more nuanced understanding of model behavior and output quality before extending the approach to other languages.

The subsequent subsections detail our data cleaning pipeline, exploratory data analysis insights, the specifics of our modeling techniques, and the evaluation framework.

4.1. Data Cleaning

A recurrent issue observed across multiple languages was the presence of triplicated sentences within original posts, often appended with a None placeholder. This pattern, likely an artifact of automated data collection or formatting, is exemplified in Table 3 for a Portuguese instance.

Table 3

Example of an original Portuguese post exhibiting triplicated content and a trailing None, alongside its normalized version and English translations. The cleaned original post segment (after deduplication by Algorithm 1) is underlined. This pattern was observed across multiple languages.

Portuguese	English Translation
Post Original: Na Holanda, a ministra da Saúde trabalha duas (2) horas diariamente como agente de limpeza antes de ir ao seu escritório. Gostei muito. Na Holanda, a ministra da Saúde trabalha duas (2) horas diariamente como agente de limpeza antes de ir ao seu escritório. Gostei muito. Na Holanda, a ministra da Saúde trabalha duas (2) horas diariamente como agente de limpeza antes de ir ao seu escritório. Gostei muito. None Saída Normalizada: Na Holanda, a ministra da Saúde trabalha duas horas diariamente como agente de limpeza antes de ir ao seu escritório.	Original Post: In the Netherlands, the Minister of Health works two (2) hours daily as a cleaning worker before going to her office. I really liked that. In the Netherlands, the Minister of Health works two (2) hours daily as a cleaning worker before going to her office. I really liked that. In the Netherlands, the Minister of Health works two (2) hours daily as a cleaning worker before going to her office. I really liked that. None Normalized Claim: In the Netherlands, the Minister of Health works two hours daily as a cleaning worker before going to her office.

To rectify this, we implemented a preprocessing routine (Algorithm 1) designed to first remove any trailing None tokens. Subsequently, it identifies and condenses repeated textual sequences by searching for the smallest repeating pattern that constitutes the entire post. If such a pattern is found, only a single instance of it is retained.

Algorithm 1 Preprocessing for Repetitive Content and Placeholder Removal

Require: Raw post P **Ensure:** Cleaned post P_{clean}

```
 $P_{\text{clean}} \leftarrow P$ 
if  $P$  ends with “None” then                                ▷ Remove trailing placeholder
     $P_{\text{clean}} \leftarrow P_{\text{clean}}[: -4].\text{strip}()$ 
end if
 $P_{\text{words}} \leftarrow \text{tokenize}(P_{\text{clean}})$                                 ▷ Tokenize into words
for  $s = 1$  to  $\lfloor |P_{\text{words}}|/2 \rfloor$  do                                ▷ Check for repetitive patterns
     $W \leftarrow P_{\text{words}}[0 : s]$                                     ▷ Extract candidate pattern
     $\text{num\_repeats} \leftarrow \lfloor |P_{\text{words}}|/s \rfloor$ 
     $\text{repeated\_sequence} \leftarrow W$  repeated  $\text{num\_repeats}$  times
    if  $P_{\text{words}}[0 : s \cdot \text{num\_repeats}] = \text{repeated\_sequence}$  then
         $P_{\text{clean}} \leftarrow \text{join}(W, \text{spaces})$                                 ▷ Return single instance
        return  $P_{\text{clean}}$ 
    end if
end for
return  $P_{\text{clean}}$                                 ▷ No repetition found
```

Additionally, we implemented cross-split deduplication to handle identical posts appearing in multiple dataset partitions. To preserve evaluation integrity, duplicates were systematically removed by prioritizing retention in test sets, then development sets, and finally training sets.

4.2. Exploratory Data Analysis (EDA)

We analyzed word count distributions for original posts and their corresponding normalized claims within the Portuguese subset following preprocessing (Algorithm 1). As depicted for the training and development sets in Figures 1a and 1b, original posts (mean ≈ 75 words, STD ≈ 107 words; highly skewed) are substantially longer and more variable than normalized claims (mean ≈ 15 words, STD ≈ 6.8 words). This demonstrates that normalization effectively reduces verbosity and structural noise, leading to more compact and verifiable statements.

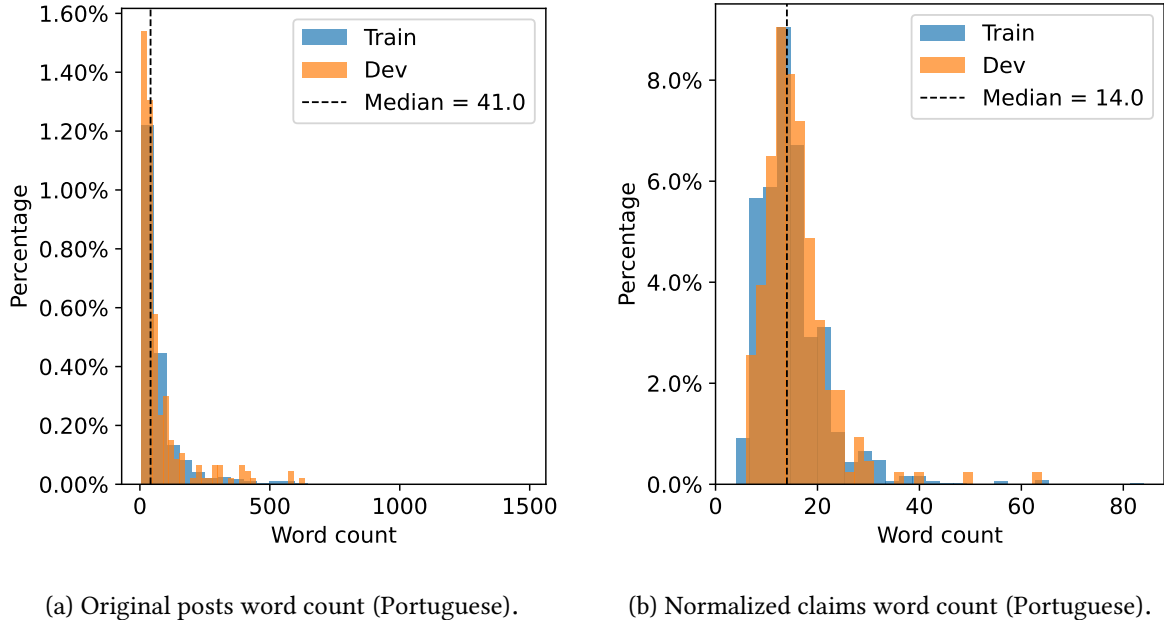


Figure 1: Histograms of word counts for the Portuguese training and development sets after preprocessing.

The dataset’s heterogeneity across languages, as indicated in Table 2, presents further challenges. While some languages offer substantial training samples, others, particularly those in the zero-shot

group, provide only minimal test data. This disparity necessitates careful consideration of model generalization and poses a risk of overfitting in resource-rich scenarios and underperformance in low-resource ones.

5. Experiments

Our experimental design directly follows the dual-strategy approach detailed in Section 4, addressing both monolingual (with training and development data) and zero-shot (test data only for specific languages) settings. All input posts were preprocessed according to Algorithm 1.

5.1. Fine-tuning of Encoder-Decoder SLMs

This approach involved adapting existing pre-trained encoder-decoder Transformer models for the specific task of claim normalization, utilizing the available training data for the thirteen supervised languages. We primarily sourced these models from the Hugging Face Hub¹. Our strategy prioritized monolingual models, selecting those pre-trained specifically for each target language. The selected monolingual models included:

- **Portuguese:** PTT5 (small, base, large) [13], PTT5-mMARCO (base) [14], PTT5-v2 (small, base, large) [15], Mono-PTT5 (small, base, large) [15], Portuguese Bart (base) [16].
- **Arabic:** AraT5 (base) [17].
- **French:** T5 French (base) [18].
- **German:** T5 German (small) [19].
- **Indic languages - Hindi, Marathi, Punjabi, Tamil:** Varta T5 (base) [20].
- **Indonesian:** Indonesian T5 Summarization Base [21].
- **Polish:** PLT5 (base) [22].
- **Spanish:** T5S (base) [23].
- **Thai:** ThaiT5 Instruct (base) [24].

In addition to these language-specific models, we also experimented with fine-tuning the following multilingual encoder-decoder architectures on the combined training data of all supervised languages: Flan-T5 (small, base, large) [14], mBART (large) [25], and UMT5 (base) [26].

Fine-tuning was conducted on three distinct hardware platforms: Kaggle kernels equipped with two NVIDIA T4 GPUs, Google Colab Pro sessions with a single NVIDIA T4 GPU, and an on-premise server hosting a single NVIDIA A100 GPU. The A100 GPU, with its substantial memory capacity, was crucial for fine-tuning larger models like PTT5 (Large), which would exceed the memory limits of the T4 GPUs.

Table 4

The hyperparameter search space used for fine-tuning. The final configuration for each language was selected based on the best METEOR or BERTScore on the development set.

Hyper-parameter	Value
Epochs	{3, 5, 10, 20}
Learning rate	$\{3 \times 10^{-3}, 1 \times 10^{-3}, 5 \times 10^{-4}\}$
Warm-up steps	90
Effective Batch size	32 (via gradient accumulation)
Generation Max Length	{128, dynamic}
Optimizer	{Adafactor, AdamW}
Number of beams	15

¹<https://huggingface.co/models>

We used hyperparameter search space, summarized in Table 4, in which a gradient accumulation was employed to achieve an effective batch size of 32, even on GPUs with 16 GB of VRAM. The maximum generation length for each batch was dynamically set to the length of the longest target sequence in that batch plus two tokens, minimizing unnecessary padding.

5.2. Inference with LLMs

For the monolingual setting, as an alternative to fine-tuning SLMs, we investigated the capabilities of LLMs using few-shot in-context learning in two distinct scenarios: (1) few-shot in-context learning as an alternative to fine-tuning smaller language models (SLMs) in monolingual settings, and (2) zero-shot inference for cross-lingual transfer to languages without available training data. The following LLMs were experimented via their respective APIs:

- **Google Gemini [27]:** Gemini 2.0 Flash Lite, Gemini 2.0 Flash Thinking.
- **OpenAI GPT [28]:** GPT-4o, GPT-4o mini, GPT-4.1 mini.
- **OpenAI Reasoning [29, 30]:** o1, o3 mini.
- **Mistral Pixtral²:** Pixtral Large (124B).
- **Alibaba Qwen [31]:** Qwen 2.5 Instruct (3B).

All models were used without post-processing, reclassification, or ensemble methods. Fewer than 0.5% of API requests returned empty strings, which were retained as-is in our submissions to preserve the authenticity of model outputs.

5.2.1. Zero-shot Prompting (Zero-shot Setting)

For zero-shot inference, we developed language-specific prompts that define the normalization task without providing examples. The English prompt (Figure 2) served as a template, which was then translated into other languages to ensure consistent task framing. The complete set of prompts is available in Appendix A.

You have received an informal and disorganized social media post. Summarize this post into a clear and concise statement, without adding any new information.

Post: {original_post}

Normalized statement:

Figure 2: The English prompt for zero-shot inference. It instructs the model to normalize the input, inserted at the {original_post} placeholder, by summarizing it concisely while preserving its meaning.

5.2.2. Few-shot Prompting (Monolingual Setting)

In the few-shot setting (applied to monolingual languages as an alternative to SLM fine-tuning), we investigated the impact of varying the number of in-context demonstrations by evaluating prompts with **3, 5, and 10 examples** (shots). These examples were selected from the training data of the respective language using several distinct strategies to assess their influence on model performance:

- **Random Selection:** Examples were drawn uniformly at random without replacement.
- **Mixed Difficulty:** Examples combined 'easy' and 'hard' posts (as defined below) to foster robustness.
- **Hard Only:** Examples exclusively featured 'hard' posts (as defined below) to test model performance on challenging cases.
- **HDBSCAN Top-k Prototypes:** Examples comprised prototypes from the k largest HDBSCAN clusters (e.g., $k = 3$ or $k = 5$), aiming for diverse semantic coverage.

²<https://mistral.ai/news/pixtral-large>

For the difficulty-stratified strategies, example difficulty was determined by calculating the METEOR score of the original post against its normalized version for every example in the training dataset. Training examples with the lowest METEOR scores (relative to other examples in the same dataset) were heuristically classified as 'hard', hypothesized to be more challenging for the model or to represent lower-quality reference outputs. Conversely, 'easy' examples were those with the highest METEOR scores.

For HDBSCAN-based strategies, semantically diverse prototypes were selected by first converting posts into sentence embeddings using the 'paraphrase-multilingual-MiniLM-L12-v2' model, a distilled Transformer architecture optimized for multilingual sentence-level representations [32, 33]. HDBSCAN (with `min_cluster_size=5`) then clustered these embeddings. The post closest to each cluster's centroid was designated as a prototype. The 'Top-k Prototypes' strategy used prototypes from the k largest clusters, supplemented with random examples if the number of clusters was less than k . This approach targets broad semantic coverage with minimal curation.

6. Results

This section presents the performance of team AKCIT-FN in the CLEF-2025 CheckThat! Task 2 on Claim Normalization. Our language-adaptive framework achieved podium finishes (top three) in 15 out of the 20 languages evaluated. Specifically, our submissions secured:

- **Second place** in 8 languages: Tamil, Thai, Punjabi, Telugu, Greek, Romanian, Dutch, and Korean.
- **Third place** in 7 languages: Portuguese, Spanish, French, Indonesian, Bengali, Polish, and German.

Table 5 provides a detailed breakdown of our best submission for each language, including the strategy employed, model specifications, average METEOR score, and the official ranking assigned by the organizers.

Analysis of our submissions reveals distinct trends based on data availability. For the **monolingual setting** (Table 5(a)), where in-domain training data was available, fine-tuned Small Language Models (SLMs) consistently outperformed few-shot Large Language Model (LLM) prompting strategies. This is evidenced by all our best-evaluated submissions in this setting utilizing SLM fine-tuning.

Notably, the Varta T5 base model (395M parameters) proved highly effective for Indic languages, securing 2nd place for Tamil (0.5197 METEOR) and Punjabi (0.3038 METEOR), and was also the model of choice for Hindi and Marathi (both achieving 5th place). Furthermore, SLM fine-tuning led to 3rd place finishes for Portuguese, Spanish, Indonesian, French, Polish, and German, employing various language-specific T5-based models with parameter counts ranging from 60M (T5 German small) to approximately 275M (PLT5 Base).

Conversely, in the **zero-shot setting** (Table 5(b)), characterized by the absence of language-specific training data, inference with pre-trained Large Language Models (LLMs) demonstrated strong generalization capabilities. Our top performances in this category were achieved using models such as Qwen 2.5 Instruct (3B parameters), which secured 2nd place for Telugu (0.5176 METEOR) and 3rd for Bengali (0.2916 METEOR), and GPT-4o Mini, which achieved 2nd place for Greek, Romanian, Dutch, and Korean. These results underscore the utility of LLMs for rapid adaptation to new languages where specialized training data is scarce.

7. Conclusion

This paper detailed AKCIT-FN's participation in the CLEF-2025 CheckThat! Task 2 on multilingual claim normalization. Our approach involved a language-adaptive strategy: fine-tuning language-specific or multilingual Small Language Models (SLMs) for languages with training data, and employing zero-shot prompting with Large Language Models (LLMs) for languages without such data.

Table 5

Comprehensive performance summary showing our best submission per language with strategy, model specifications, METEOR scores, and official rankings. Podium finishes (top 3) are highlighted in bold.

Setting	Language	Best Submission Details			
		Strategy	Model (Parameters)	Avg. METEOR	Rank
Monolingual (training data available)	Portuguese	SLM Fine-tuning	Mono PTT5 base (220M)	0.5290	3rd
	Spanish	SLM Fine-tuning	T5S base (220M)	0.5213	3rd
	Tamil	SLM Fine-tuning	Varta T5 base (395M)	0.5197	2nd
	English	SLM Fine-tuning	Flan-T5 base (250M)	0.4058	4th
	Indonesian	SLM Fine-tuning	Indonesian T5 Summarization Base (250M)	0.3866	3rd
	French	SLM Fine-tuning	T5 French base (250M)	0.3811	3rd
	Arabic	SLM Fine-tuning	AraT5 Base (220M)	0.3277	6th
	Thai	SLM Fine-tuning	Thai T5 Base (245M)	0.3179	2nd
	Punjabi	SLM Fine-tuning	Varta T5 base (395M)	0.3038	2nd
	Polish	SLM Fine-tuning	plT5 Base (275M)	0.2798	3rd
	Hindi	SLM Fine-tuning	Varta T5 base (395M)	0.2706	5th
	German	SLM Fine-tuning	T5 German small (60M)	0.2652	3rd
	Marathi	SLM Fine-tuning	Varta T5 base (395M)	0.2181	5th
Zero-shot (no training data)	Telugu	LLM Inference	Qwen 2.5 Instruct (3B)	0.5176	2nd
	Bengali	LLM Inference	Qwen 2.5 Instruct (3B)	0.2916	3rd
	Greek	LLM Inference	GPT-4o Mini	0.2567	2nd
	Romanian	LLM Inference	GPT-4o Mini	0.2516	2nd
	Dutch	LLM Inference	GPT-4o Mini	0.1922	2nd
	Czech	LLM inference	GPT-4o Mini	0.1734	4th
	Korean	LLM Inference	GPT-4o Mini	0.1209	2nd

Our submissions demonstrated strong performance, achieving podium finishes (top three) in 15 out of the 20 languages. Notably, fine-tuned SLMs excelled in supervised settings, while LLMs proved effective for zero-shot generalization, securing five second-place and one third-place finish among the seven zero-shot languages. For Portuguese, our primary development language, our best system (Mono PTT5 base) ranked third with a METEOR score of 0.5290.

Overall, these results underscore the complementary strengths of SLM fine-tuning when in-domain data is available and the powerful generalization capabilities of LLMs for rapid deployment in zero-shot scenarios for complex NLP tasks like claim normalization. A limitation of our work, however, is the lack of a qualitative error analysis or a discussion of failure cases, which would be a valuable direction for future investigation. Our code and configurations are publicly available to facilitate further research.

Acknowledgments

This work has been fully funded by the project Computational Techniques for Multimodal Data Security

and Privacy supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPPI.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 2.5 Pro [27] (gemini-2.5-pro-exp-03-25), Claude Sonet 4 [34] (claude-sonnet-4-20250514), GPT-4o [28] (gpt-4o-2024-05-13) in order to: Paraphrase and reword, Improve writing style, Abstract drafting, and Peer review simulation. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] M. Sundriyal, T. Chakraborty, P. Nakov, From chaos to clarity: Claim normalization to empower fact-checking, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6594–6609.
- [2] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venkatesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.
- [3] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonello (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.
- [4] M. Sundriyal, T. Chakraborty, P. Nakov, Overview of the CLEF-2025 CheckThat! lab task 2 on claim normalization, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.
- [5] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, Digital Threats 2 (2021). URL: <https://doi.org/10.1145/3412869>. doi:10.1145/3412869.
- [6] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Márquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, in: P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2018, pp. 372–387.
- [7] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2019, p. 301–321. URL: https://doi.org/10.1007/978-3-030-28577-7_25. doi:10.1007/978-3-030-28577-7_25.
- [8] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, Z. S. Ali, Overview of checkthat! 2020: Automatic identification and verification of claims in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings, Springer-

Verlag, Berlin, Heidelberg, 2020, p. 215–236. URL: https://doi.org/10.1007/978-3-030-58219-7_17. doi:10.1007/978-3-030-58219-7_17.

- [9] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer International Publishing, Cham, 2021, pp. 264–291.
- [10] D. S. Nielsen, R. McConville, Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset, in: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, Association for Computing Machinery, New York, NY, USA, 2022, p. 3141–3153. URL: <https://doi.org/10.1145/3477495.3531744>. doi:10.1145/3477495.3531744.
- [11] I. Singh, C. Scarton, X. Song, K. Bontcheva, Breaking language barriers with mmtweets: Advancing cross-lingual debunked narrative retrieval for fact-checking, 2024. URL: <https://arxiv.org/abs/2308.05680>. arXiv:2308.05680.
- [12] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: <https://aclanthology.org/W05-0909/>.
- [13] D. Carmo, M. Piau, I. Campiotti, R. Nogueira, R. Lotufo, Ptt5: Pretraining and validating the t5 model on brazilian portuguese data, 2020. URL: <https://arxiv.org/abs/2008.09144>. arXiv:2008.09144.
- [14] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tai, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, *J. Mach. Learn. Res.* 25 (2024).
- [15] M. Piau, R. Lotufo, R. Nogueira, ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language, in: A. Paes, F. A. N. Verri (Eds.), *Intelligent Systems*, Springer Nature Switzerland, Cham, 2025, pp. 324–338.
- [16] Adalberto Ferreira Barbosa Junior, bart-base-portuguese (revision 149de72), 2024. URL: <https://huggingface.co/adalbertojunior/bart-base-portuguese>. doi:10.57967/hf/3264.
- [17] E. M. B. Nagoudi, A. Elmadany, M. Abdul-Mageed, AraT5: Text-to-text transformers for Arabic language generation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 628–647. URL: <https://aclanthology.org/2022.acl-long.47/>. doi:10.18653/v1/2022.acl-long.47.
- [18] guillaumephd, T5-french-base model: A t5 model trained on french data only, 2024. URL: <https://huggingface.co/guillaumephd/t5-french-base>.
- [19] Shahm, t5-seven-epoch-base-german, 2023. URL: <https://huggingface.co/Shahm/t5-small-german>.
- [20] R. Aralikatte, Z. Cheng, S. Doddapaneni, J. C. K. Cheung, Varta: A large-scale headline-generation dataset for Indic languages, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 3468–3492. URL: <https://aclanthology.org/2023.findings-acl.215/>. doi:10.18653/v1/2023.findings-acl.215.
- [21] C. Wirawan, Indonesian T5 Summarization Base Model, <https://huggingface.co/cahya/t5-base-indonesian-summarization-cased>, 2021. Accessed: 2025-05-29.
- [22] A. Chrabrowa, Ł. Dragan, K. Grzegorzczak, D. Kajtoch, M. Koszowski, R. Mroczkowski, P. Rybak, Evaluation of transfer learning for Polish with a text-to-text model, in: N. Calzolari, F. Béchet,

- P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2022, pp. 4374–4394. URL: <https://aclanthology.org/2022.lrec-1.466/>.
- [23] V. Araujo, M. M. Trusca, R. Tufiño, M.-F. Moens, Sequence-to-sequence Spanish pre-trained language models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ELRA and ICCL, Torino, Italia, 2024, pp. 14729–14743. URL: <https://aclanthology.org/2024.lrec-main.1283/>.
- [24] Peenipat, Thai5-instruct, 2025. URL: <https://huggingface.co/Peenipat/ThaiT5-Instruct>.
- [25] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, *Transactions of the Association for Computational Linguistics* 8 (2020) 726–742. URL: <https://aclanthology.org/2020.tacl-1.47/>. doi:10.1162/tacl_a_00343.
- [26] H. W. Chung, N. Constant, X. Garcia, A. Roberts, Y. Tay, S. Narang, O. Firat, Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining, 2023. URL: <https://arxiv.org/abs/2304.09151>. arXiv:2304.09151.
- [27] Gemini Team, Google, Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, Technical Report, Google DeepMind, 2025. URL: https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.
- [28] OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mądry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valldares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O’Connell, I. O’Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljube, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan,

- M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunningham, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, Y. Malkov, Gpt-4o system card, 2024. URL: <https://arxiv.org/abs/2410.21276>. arXiv: 2410.21276.
- [29] OpenAI, :, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O’Connell, I. Osband, I. C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J. Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M. Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R. G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S. R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, Z. Li, Openai o1 system card, 2024. URL: <https://arxiv.org/abs/2412.16720>. arXiv: 2412.16720.
- [30] OpenAI, OpenAI o3 and o4-mini System Card, Technical Report, OpenAI, 2025. URL: <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.
- [31] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang, et al., Qwen2.5-coder technical report, arXiv preprint arXiv:2409.12186 (2024).

- [32] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2140–2151. URL: <https://aclanthology.org/2021.findings-acl.188/>. doi:10.18653/v1/2021.findings-acl.188.
- [33] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410/>. doi:10.18653/v1/D19-1410.
- [34] Anthropic, System Card: Claude Opus 4 & Claude Sonnet 4, Technical Report, Anthropic, 2025. URL: <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>.

A. Zero-shot prompts

This appendix lists the prompts used for the zero-shot claim normalization task. The English prompt in Figure 2 served as the template and was translated into the seven target languages. The {post_text} placeholder is replaced with the social media post text during inference.

• Czech

Dostanete neformální a neuspořádaný příspěvek ze sociálních sítí. Shrňte jej do jasného a stručného tvrzení, bez přidávání dalších informací.

Příspěvek: {post_text}

• Greek

Σου δίνεται μια ανοργάνωτη και ανεπίσημη ανάρτηση στα κοινωνικά δίκτυα. Περίληψέ την σε μια σαφή και συνοπτική δήλωση, χωρίς να προσθέσεις επιπλέον πληροφορίες.

Ανάρτηση: {post_text}

• Dutch

Je ontvangt een informeel en ongeorganiseerd bericht op een sociaal netwerk. Vat het samen in een duidelijke en beknopte verklaring zonder extra informatie toe te voegen.

berichten:{post_text}

• Korean

비정형적이고 비공식적인 소셜 미디어 게시물이 주어집니다. 이를 명확하고 간결한 주장으로 요약하십시오. 추가 정보는 포함하지 마십시오.

게시물:{post_text}

• Romanian

Prompt: Primești o postare informală și dezorganizată de pe o rețea socială. Rezum-o într-o afirmație clară și concisă, fără a adăuga informații.

Postare: {post_text}

• Tegulu

మీకు ఒక అసంఘటితమైన, అనౌపచారికమైన నోట్ మీడియాలో పోస్ట్ ఇవ్వబడుతుంది. దీనిని సుస్పష్టమైన మరియు సంక్షిప్తమైన వరకటనగా మార్చండి, అదనపు సమాచారం ఇవ్వకుండా.

పోస్ట్:{post_text}

- **Bengali**

আপনি একটি অগোছালো এবং অনানুষ্ঠানিক সোশ্যাল মিডিয়া পোস্ট পাচ্ছেন। এটি একটি স্পষ্ট এবং সংক্ষিপ্ত দাবিতে রূপান্তর করুন, কোনো অতিরিক্ত তথ্য ছাড়াই। **পোস্ট:** {post_text}