# Cross-Lingual Fact Verification: Analyzing LLM Performance Patterns Across Languages

**Hanna Shcharbakova[1], Tatiana Anikina[2], Natalia Skachkova[2], Josef van Genabith[1,2]**
[1]Saarland University
[2]German Research Center for Artificial Intelligence (DFKI)
hash00004@stud.uni-saarland.de
{tatiana.anikina, natalia.skachkova, josef.van_genabith}@dfki.de

## Abstract

Fact verification has emerged as a critical task in combating misinformation, yet most research remains focused on English-language applications. This paper presents a comprehensive analysis of multilingual fact verification capabilities across three state-of-the-art large language models: Llama 3.1, Qwen 2.5, and Mistral Nemo. We evaluate these models on the X-Fact dataset that includes 25 typologically diverse languages, examining both seen and unseen languages through various evaluation scenarios. Our analysis employs few-shot prompting and LoRA fine-tuning approaches, revealing significant performance disparities based on script systems, with Latin script languages consistently outperforming others. We identify systematic cross-lingual instruction following failures, particularly affecting languages with non-Latin scripts. Surprisingly, some officially supported languages, such as Indonesian and Polish, which are not high-resource languages, achieve better performance than high-resource languages like German and Spanish, challenging conventional assumptions about resource availability and model performance. The results highlight critical limitations in current multilingual LLMs for the fact verification task and provide insights for developing more inclusive multilingual systems.

## 1 Introduction

Fact-checking has emerged as a critical defense against the proliferation of misinformation in the digital age. While the broader fact-checking process involves multiple stages including claim detection and evidence gathering, fact verification, which is our primary focus, addresses the final crucial step of determining claim truthfulness when supporting evidence is available (Vykopal et al., 2024). The rapid spread of misinformation across digital platforms has made automated fact verification systems increasingly essential for maintaining information reliability (Fung et al., 2022; Aïmeur et al., 2023).

Recent developments in NLP have been significantly shaped by LLMs and transformer architectures, which have demonstrated remarkable capabilities across various tasks (Kotonya and Toni, 2020; Wang et al., 2023). However, the research landscape remains heavily skewed toward English-language applications (Guo et al., 2022; Vykopal et al., 2024; Wang et al., 2024). This linguistic imbalance creates substantial challenges for global misinformation detection, as false information frequently crosses language boundaries and impacts diverse communities worldwide.

Although powerful LLMs have demonstrated impressive performance across various NLP tasks, the degree to which they work well with particular languages and specific tasks varies significantly (Bang et al., 2023; Huang et al., 2023; Ignat et al., 2024). Fact verification represents a particularly challenging task requiring nuanced understanding of claims, contextual reasoning, and the ability to distinguish between different degrees of truthfulness —capabilities that may not transfer uniformly across languages (Dmonte et al., 2024). Understanding how effectively current LLMs address this critical challenge across different linguistic contexts remains an essential but understudied question.

Recent multilingual datasets such as X-Fact (Gupta and Srikumar, 2021), MultiClaim (Pikuliak et al., 2023), and (Quelle et al., 2025) have begun to address this gap by providing fact verification resources across multiple languages. However, systematic analysis of how state-of-the-art LLMs perform across different languages and scripts in fact verification tasks remains limited.

This paper presents a comprehensive analysis of LLMs' performance on multilingual fact verification, focusing on language-specific challenges and patterns. We evaluate state-of-the-art LLMs Llama



*Proceedings of Recent Advances in Natural Language Processing*,pages 1137–1147
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-131

3.1 (Dubey et al., 2024), Qwen 2.5 (Yang et al., 2024), and Mistral Nemo (Mistral AI Team, 2024) across 25 languages using the X-Fact dataset, employing both few-shot prompting and fine-tuning approaches.

Our key contributions are:

- **A comprehensive multilingual performance analysis and taxonomy** across 25 languages, revealing significant disparities based on script systems with systematic challenges identified for non-Latin writing systems, providing important insights for developing more effective multilingual fact verification systems.
- **A cross-lingual instruction following investigation** identifying specific failure patterns where models struggle to produce requested outputs across languages, particularly affecting under-represented languages.

These findings have important implications for deploying fact verification systems globally and highlight the need for more inclusive approaches to multilingual NLP system development.

## 2 Related Work

### 2.1 Multilingual Fact Verification

Early multilingual fact verification efforts focused primarily on dataset creation and basic cross-lingual transfer methods. Some notable multilingual datasets include FakeCovid (Shahi and Nandini, 2020), which spans 40 languages focusing on COVID-19 related claims, NewsPolyML (Mohtaj et al., 2024) covering over 32K fact-checked claims in five European languages, and MultiClaim (Pikuliak et al., 2023) providing 28K claims across 27 languages. However, these datasets vary significantly in size, language coverage, and annotation schemes, making consistent cross-lingual evaluation challenging.

Several studies have explored multilingual fact verification using traditional transformer models. Gupta and Srikumar (2021) achieved an F1 score of 41.9% on in-domain data but showed significant performance degradation on out-of-domain (16.2%) and zero-shot scenarios (16.7%), though detailed language-specific breakdowns were not provided. Shcharbakova et al. (2025) demonstrated that specialized smaller models (XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021)) substantially outperformed larger general-purpose LLMs on the X-Fact (Gupta and Srikumar, 2021) dataset despite having fewer parameters.

Recent work has increasingly focused on LLMs for multilingual fact verification. Pelrine et al. (2023) demonstrated that GPT-4 could outperform prior methods across multiple datasets and languages, achieving superior classification results with GPT-4 Score Optimized performing best at 68.1% F1 on English dataset LIAR (Wang, 2017) and showing strong performance on German data (57.6% accuracy) even without changing English prompts. On the NewsPolyML dataset, Mohtaj et al. (2024) showed that mBERT achieved F1 scores of up to 75.1% across English, German, French, Spanish, and Italian, with performance varying significantly by language.

### 2.2 Cross-lingual Transfer in LLMs for Fact Verification

Cross-lingual transfer learning has emerged as a promising approach to address data scarcity in multilingual fact verification, though its effectiveness varies significantly across language pairs and task complexities. Zhang et al. (2024) conducted a comprehensive analysis of Chinese fact-checking, showing that direct translation from Chinese to English resulted in inaccuracies, particularly with idiomatic expressions, and that models trained specifically on Chinese data outperformed both translation-based and multilingual approaches by over 10%.

Du et al. (2021) proposed CrossFake, a cross-lingual fake news detector. The authors applied a monolingual model (English) cross-lingually via translation, demonstrating that this strategy can outperform generic multilingual encoders for domain-specific tasks like COVID-19 fake news detection.

Cekinel et al. (2024) conducted a comprehensive evaluation of cross-lingual transfer for Turkish fact-checking, comparing zero-shot and few-shot prompting with fine-tuning approaches using LLaMA-2 models (Touvron et al., 2023). Their experiments revealed that while few-shot learning provided modest improvements over zero-shot approaches, fine-tuning on native Turkish data yielded substantially better results compared to cross-lingual transfer methods. This finding underscores the importance of language-specific training data even when leveraging powerful multilingual models. The study also explored machine

translation as a bridge for cross-lingual transfer, finding that translating Turkish claims to English and applying English-trained models achieved better results than the reverse direction. However, translation-based approaches introduced their own limitations, particularly in preserving cultural and contextual nuances essential for accurate fact verification.

The challenge of cross-lingual fact verification is further complicated by the need to handle diverse writing systems and cultural contexts. Research has consistently shown that model effectiveness is closely tied to language representation in pre-training data, with high-resource languages like English and Spanish typically showing better performance than low-resource languages (Hendy et al., 2023; Ahuja et al., 2023; Asai et al., 2024). Script-related challenges have been identified as a significant factor affecting model performance, with non-Latin scripts often presenting additional processing difficulties (Bang et al., 2023).

Despite these advances, several gaps remain in our understanding of LLMs performance in multilingual fact verification. First, most studies focus on a limited number of languages or specific language pairs, leaving the broader multilingual landscape underexplored. Second, systematic analysis of how script systems and resource levels affect fact verification performance is lacking. Finally, the specific challenges faced by LLMs in cross-lingual instruction following for fact verification tasks have not been thoroughly investigated. Our work addresses these gaps by providing a comprehensive analysis of LLM performance across 25 languages, examining the interplay between script systems, resource levels, and models, in the context of fact verification.

## 3 Data

We conduct our multilingual fact verification analysis using the X-Fact dataset (Gupta and Srikumar, 2021), comprising 31,189 claims across 25 languages from 11 language families. The data consists of claims, accompanying evidence, and metadata collected from fact-checking websites, ensuring real-world applicability. The metadata includes language information, source website, claimant details, claim dates, review dates, and links to original evidence sources. Claims are classified into seven veracity categories: *true, mostly true, partly true/misleading, mostly false, false, complicated/hard to categorise,* and *other*.

The dataset is structured into multiple evaluation subsets designed to test different aspects of cross-lingual generalization. The training data contains 19,079 claims across 13 languages. The test subset includes 3,826 claims from the same 13 languages, enabling evaluation of model performance on familiar languages. The zero-shot subset comprises 3,381 claims across 12 different languages not seen during training, testing cross-lingual transfer capabilities to completely unfamiliar languages. While an out-of-domain evaluation set exists as well, it falls outside the scope of our research on language-specific analysis.

X-Fact exhibits significant imbalances in terms of both language and label distribution. These imbalances extend to the evaluation subsets, with uneven representation across different script systems and resource levels (Figure 1). Such imbalances may affect model calibration and performance, particularly for underrepresented languages and less frequent veracity categories, potentially leading to biased predictions toward dominant languages and frequent labels.

To systematically analyze these language-specific challenges, we categorize the 25 languages along two key dimensions:

**Script systems:** *Latin script* languages (Azerbaijani, German, Indonesian, Italian, Polish, Portuguese, Romanian, Serbian, Spanish, Turkish, Albanian, Dutch, French, Norwegian), *Arabic script* languages (Arabic, Persian), *Devanagari script* languages (Hindi, Marathi) and *Other scripts* (Georgian, Tamil, Bengali, Gujarati, Punjabi, Russian, Sinhala).

**Resource levels:** while Joshi et al. (2020) proposes a six-class categorization based on the data availability, we simplify this into a ternary classification for our analysis, distinguishing between *well-represented* (German, Spanish, French, Arabic), *moderately-represented* (Portuguese, Italian, Dutch, Polish, Turkish, Persian, Hindi, Russian, Serbian), and *under-represented* languages (Indonesian, Romanian, Georgian, Tamil, Bengali, Punjabi, Marathi, Albanian, Azerbaijani, Gujarati, Norwegian, Sinhala).

## 4 Experimental Setting

We evaluate three state-of-the-art multilingual LLMs across 25 languages to analyze their fact verification capabilities in terms of resource levels

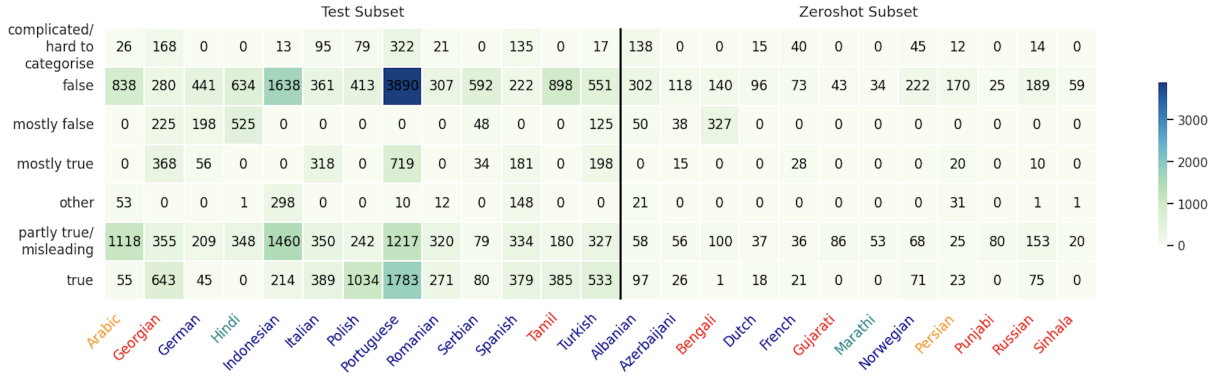


Figure 1: X-Fact test and zero-shot evaluation subsets details. The languages are color-coded based on the script systems they use (Arabic, Latin, Devanagari, and Other).

and script systems. We focus on large generative decoder-only models as they are expected to excel at reasoning tasks and evidence assessment, which are crucial components of fact verification. We selected three instruction-tuned LLMs based on their parameter sizes and language coverage: Llama 3.1 (8B) (Dubey et al., 2024) officially supports eight languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. Qwen 2.5 (7B) (Yang et al., 2024) offers the broadest coverage with 29 languages, including strong representation of Asian languages. Mistral Nemo (12B) (Mistral AI Team, 2024) is the largest model in our selection, supporting 11 languages: English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi.

It is important to note that the varying official language support across models creates a potential confounding factor in our analysis. When comparing performance patterns across resource levels and languages, differences may reflect not only traditional factors like training data availability, but also model-specific optimizations for officially supported languages.

### 4.1 Experimental Approach

We evaluate each model under two configurations: **few-shot prompting** and **LoRA fine-tuning** (Hu et al., 2022), both using claim-evidence pairs. For few-shot prompting, we used 7 examples corresponding to the number of veracity categories, ensuring representation across different languages and scripts. We developed a structured prompt providing clear task instructions in English for the seven-way veracity classification and descriptions of each category. The fine-tuning experiments employed LoRA targeting all attention and feed-forward components. Training data was selected randomly and balanced across all languages and veracity labels to prevent bias toward overrepresented categories.

### 4.2 Evaluation Protocol

We evaluate model performance using macro-F1 scores. Performance analysis is conducted at multiple levels: language-specific analysis for each of the 25 languages, script system comparison (Latin, Arabic, Devanagari, and Other scripts), and resource level analysis (well-, moderately-, and under-represented languages).

We evaluate models on both the test subset (languages seen during training) and the zero-shot subset (languages absent from training data) to analyze cross-lingual transfer effectiveness. All models were instructed to provide veracity labels in English, with robust output processing implemented to handle diverse response formats through text normalization and label mapping procedures.

## 5 Results

Our results reveal significant performance disparities across languages, script systems, and resource levels (Figure 2). Overall performance remains relatively low across all models and languages, with the highest-performing language-model combination (Norwegian with Qwen 2.5 fine-tuning) achieving 0.34 macro-F1. Most languages perform substantially below this level, indicating the challenging nature of multilingual fine-grained fact verification for current LLMs.

Qwen 2.5 achieves the highest scores across most languages, consistently outperforming Mistral Nemo and Llama 3.1 across different language categories and script systems. Mistral Nemo shows

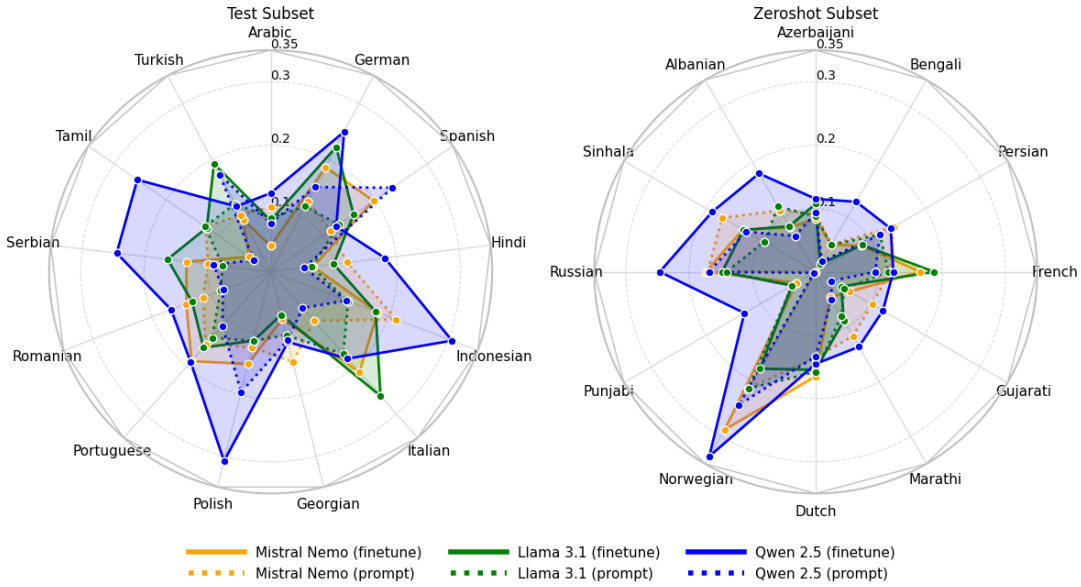


Figure 2: Comparative performance of Mistral Nemo, Llama 3.1, and Qwen 2.5 models on test and zero-shot subsets, measured by macro-F1 scores. Solid lines denote LoRA fine-tuned models, dotted lines few-shot prompted models.

competitive results for some European languages but demonstrates difficulties with non-European languages and non-Latin scripts. Despite having the largest parameter count (12B), it does not consistently outperform the smaller Qwen 2.5 model. Llama 3.1 exhibits variable performance patterns with notable strengths in certain well-represented languages but some significant weaknesses in cross-lingual transfer scenarios.

Fine-tuning consistently outperforms few-shot prompting across most languages, with particularly pronounced performance improvements in top-performing settings such as Polish and Indonesian with Qwen 2.5, and Norwegian in the zero-shot evaluation, where fine-tuning provides substantial gains over prompting approaches.

### 5.1 Performance on Test Subset

Polish and Indonesian demonstrate the strongest performance with Qwen 2.5 achieving around 0.31 macro-F1 in the fine-tuning configuration. Portuguese and Italian achieve moderate performance with scores predominantly around 0.18-0.22 macro-F1. Spanish and German show similar moderate performance across different models. Arabic and Georgian consistently show the poorest performance, with both languages scoring below 0.13 macro-F1 across all models and configurations.

### 5.2 Performance on Zero-shot Subset

Norwegian achieves exceptional performance with Qwen 2.5 reaching 0.34 macro-F1 in the fine-tuning configuration, surpassing most seen languages from the test subset. French and Dutch demonstrate relatively strong cross-lingual transfer, while Russian shows moderate transfer performance across most scenarios. South Asian languages show poor zero-shot transfer performance. Bengali, Gujarati, Punjabi, and Marathi consistently score below 0.13 macro-F1 across all models.

### 5.3 Resource Level Performance

Resource-level analysis does not show a clear pattern, though this is complicated by varying official language support across models. Well-represented languages demonstrate mixed results across both evaluation scenarios. German achieves performance around 0.12-0.25 macro-F1 in the test subset across different LLMs and settings, while Spanish shows similar performance. French demonstrates relatively strong zero-shot transfer performance around 0.19 macro-F1.

Moderately-represented languages exhibit highly variable performance patterns. Polish achieves exceptional performance as one of the top performers despite its moderate resource status, notably supported by Qwen 2.5. Many languages from this group, including Italian and Hindi, show

moderate performance.

Under-represented languages show the most inconsistent relationship between resource availability and performance. Indonesian achieves one of the best scores on the test subset, contradicting expectations based on resource limitations but aligning with its official support in Qwen 2.5. Conversely, Georgian and Romanian show performance more aligned with traditional resource constraints and lack broad official support across models.

### 5.4 Script System Performance

As shown in Figure 4 for few-shot prompting, Latin script languages demonstrate the highest median performance across Llama 3.1 and Qwen 2.5 models. All three models achieve similar median performance for Arabic script languages, with Mistral Nemo showing slightly higher variance in this category.

Devanagari script languages demonstrate the most constrained performance across all models, with consistently low median scores and minimal variance. This pattern indicates systematic challenges in processing languages using the Devanagari writing system regardless of the model architecture.

The Other script category shows the highest performance variance, particularly for Qwen 2.5 and Mistral Nemo. While some languages in this category achieve relatively high performance, others perform poorly, resulting in wide interquartile ranges and numerous outliers.

### 5.5 Cross-lingual Instruction Following

Analysis of fine-tuned model outputs reveals systematic failures in cross-lingual instruction following, with models frequently unable to produce valid English labels as instructed.

These failures exhibit two major patterns: complete output failure (empty responses) and language code-switching (responding in the same or different from the input language rather than English) (see Figure 3).

Qwen 2.5 demonstrates the most robust instruction-following capabilities. In the test subset, it produced only 2 invalid examples (0.05% of the test dataset) due to same-language responses. In the zero-shot subset, it had 13 invalid examples (0.38%), including 3 outputs in unintended languages and 10 same-language responses. Among the same-language failures, 6 occurred in Gujarati.

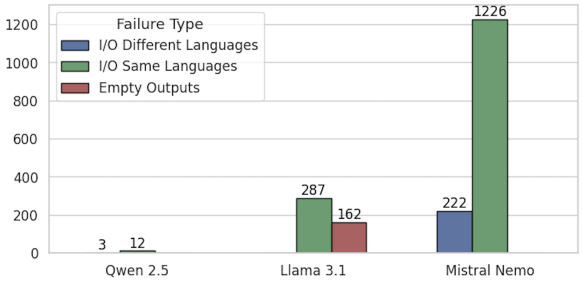


Figure 3: Instruction-following failure counts across three LLMs on the test and zero-shot subsets.

Llama 3.1 shows moderate instruction-following difficulties. In the test subset, 224 instances (5.85% of the test dataset) failed to produce valid labels: 9 complete failures (4 Georgian, 5 Tamil) and 215 cases of responding in the input language (136 Georgian, 79 Tamil). The zero-shot subset reveals more severe challenges, with 225 failures (6.66% of the zero-shot dataset), including 153 empty outputs (111 Bengali, 19 Gujarati, 22 Punjabi) and 72 same-language responses (65 Bengali, 2 Gujarati, 5 Punjabi).

Mistral Nemo exhibits the most significant challenges across both subsets. In the test subset, 767 instances (20.05% of the test dataset) failed to produce valid labels: 198 responses in different languages (161 Georgian) and 569 same-language responses (174 Arabic, 159 Tamil, 90 Hindi). In the zero-shot subset, 681 failures (20.14% of the zero-shot dataset) occurred, including 24 different-language responses and 657 same-language responses (325 Bengali, 93 Persian, 78 Punjabi).

## 6 Discussion

### 6.1 Script System as a Performance Predictor

The most striking finding from our analysis is the significant impact of script systems on model performance. Languages using Latin scripts consistently demonstrate superior performance across all three models, with median macro-F1 scores generally higher than other script categories. This pattern suggests that current LLMs, despite their multilingual training, maintain inherent biases toward Latin-based writing systems that dominate their training corpora.

The systematic challenges faced by models when processing non-Latin scripts extend beyond simple character recognition issues. Our analysis reveals that Devanagari script languages (Hindi, Marathi) show the most constrained performance with mini-

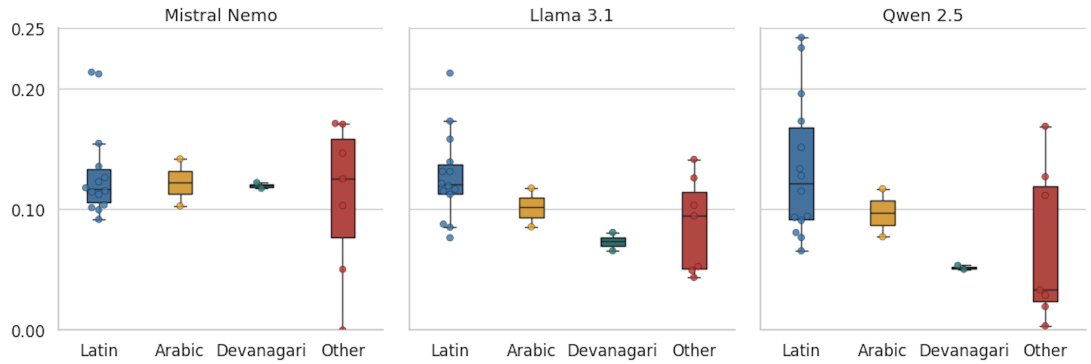


Figure 4: Comparison of macro-F1 scores across Mistral Nemo, Llama 3.1, and Qwen 2.5 models using few-shot prompting, grouped by script system. Each subplot represents a script (Latin, Arabic, Devanagari, and Other)., with boxplots showing median and interquartile ranges, and swarm plots indicating individual language performances.

mal variance across models. Similarly, languages in the *Other scripts* category exhibit high performance variance, suggesting that some scripts may be better represented in training data than others, leading to inconsistent cross-lingual transfer.

### 6.2 Cross-lingual Transfer Patterns

Our evaluation on the zero-shot subset reveals some interesting patterns in cross-lingual transfer. Norwegian's exceptional performance (0.34 macro-F1 with Qwen 2.5) in the zero-shot setting, surpassing many languages seen during training, suggests that certain linguistic features may facilitate better transfer than others. This finding indicates that other factors may be more important for fact verification transfer than simple training data availability.

The strong performance of Germanic languages (Norwegian, Dutch) in zero-shot scenarios, combined with the relatively good transfer to French, demonstrates clear patterns in cross-lingual transfer effectiveness. However, the poor performance of South Asian languages (Bengali, Gujarati, Punjabi, Marathi) in zero-shot scenarios highlights the complex interplay between script systems and resource levels.

Particularly noteworthy is the dramatic performance degradation for languages absent from training data. While Norwegian achieves remarkable zero-shot performance, most unseen languages struggle significantly, with Bengali, Gujarati, and Punjabi consistently scoring below 0.13 macro-F1. This contrast suggests that successful cross-lingual transfer in fact verification depends on complex factors that are not uniformly distributed across different languages.

### 6.3 Resource Levels and Official Language Support

The exceptional performance of Indonesian (under-represented according to our classification) and Polish (moderately-represented) compared to well-represented languages like German and Spanish suggests that factors beyond data availability drive multilingual fact verification capabilities. Official language support by model developers (Indonesian and Polish for Qwen 2.5) provides a more reliable predictor of success.

The inconsistent relationship between resource availability and performance suggests that data quality may be more critical than quantity for enabling knowledge transfer. This assumption is supported by Norwegian's exceptional zero-shot performance (0.34 macro-F1), which surpasses most training languages despite being absent from training data. Norwegian's success likely stems from cross-lingual transfer from linguistically similar Germanic languages (German, Dutch) with substantial reasoning-focused training data.

### 6.4 Cross-lingual Instruction Following Failures

A particularly interesting finding is the systematic failure of models to follow cross-lingual instructions, with models frequently unable to produce requested English labels when processing non-English inputs. These failures become obvious in two primary patterns: complete output failure and language code-switching. The concentration of these failures in specific languages rather than uniform distribution across all unfamiliar languages indicates systematic model limitations rather than random processing errors. Bengali's prominence

in instruction following failures across multiple models (325 same-language responses with Mistral Nemo, 111 empty outputs with Llama 3.1) suggests that certain languages present fundamental challenges to current model architectures. A detailed error analysis examining prediction patterns across veracity categories and their varying impact severity is provided in our GitHub repository[1].

## 7 Limitations and Future Directions

Our analysis focuses on a specific fact verification dataset and task formulation, which may limit generalizability to other multilingual reasoning tasks. While X-Fact's inherent language and label imbalances may influence our observed performance patterns, these reflect real-world multilingual data challenges. Additionally, the LLMs investigated are relatively modest in scale compared to state-of-the-art models, and future work should evaluate larger models to determine whether these patterns persist at scale. The development of more balanced multilingual fact verification datasets, particularly for under-represented languages and scripts, would enable more comprehensive evaluation and support targeted training strategies that address the systematic biases we have identified.

Future work should explore more sophisticated approaches to cross-lingual transfer in fact verification, including investigation of the factors that enable successful cross-lingual transfer, which could inform more appropriate approaches to multilingual model architecture design.

While our analysis focuses primarily on script systems as a key organizing principle for understanding multilingual performance patterns, an alternative categorization based on language families would provide an interesting comparative perspective. Future research could systematically examine whether genealogical relationships between languages offer different insights into cross-lingual transfer effectiveness.

A critical limitation of the current analysis is the inability to distinguish between script-specific challenges in fact verification and more fundamental tokenization inefficiencies. The performance patterns we observe for non-Latin scripts could stem from suboptimal tokenization and under-representation of these languages in model vocabularies, rather than inherent limitations in cross-lingual capabilities. This aligns with findings from Ahia et al. (2023) and Ali et al. (2024) regarding tokenizer biases in current LLMs. Future research should investigate tokenization efficiency across different writing systems and its impact on downstream task performance to better isolate script-specific challenges from tokenizer-related limitations.

[1] https://github.com/Aniezka/cross-lingual-fact-verification.

## 8 Conclusion

This paper presents a comprehensive analysis of multilingual fact verification capabilities across 25 languages using three state-of-the-art LLMs. Our evaluation on the X-Fact dataset reveals significant performance disparities based on script systems, with Latin script languages consistently outperforming others across all models.

Key findings include the identification of systematic cross-lingual instruction following failures and the surprising result that some officially supported but not high-resource languages (such as Indonesian and Polish) achieve better performance than traditionally high-resource languages like German and Spanish, challenging conventional assumptions about the relationship between resource availability and model performance. The dramatic variation in cross-lingual transfer effectiveness, exemplified by Norwegian's strong zero-shot performance against poor results for South Asian languages, highlights the complex factors affecting multilingual capabilities.

These findings underscore critical limitations in current multilingual LLMs for complex reasoning tasks and emphasize the need for more inclusive approaches to multilingual model development. Future work should focus on addressing script system biases and developing more fair fact verification systems that serve diverse communities effectively.

## 9 Acknowledgments

This project was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005), which provided funding for Josef van Genabith and Tatiana Anikina. This work was also co-funded by the Erasmus Mundus Masters Programme in Language and Communication Technologies (EU grant no. 2019-1508).

## References


Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2024. Cross-lingual learning vs. low-resource fine-tuning: A case study with fact-checking in Turkish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4127–4142, Torino, Italia. ELRA and ICCL.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Jiangshu Du, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and Philip S Yu. 2021. Cross-lingual covid-19 fake news detection. In *2021 international conference on data mining workshops (ICDMW)*, pages 859–862. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yi R Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4790–4791.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual 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 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

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.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan C. Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. 2024. Has it all been solved? open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia. ELRA and ICCL.

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.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mistral AI Team. 2024. Mistral nemo. https://mistral.ai/en/news/mistral-nemo. Accessed: 14-Feb-2025.

Salar Mohtaj, Ata Nizamoglu, Premtim Sahitaj, Vera Schmitt, Charlott Jakob, and Sebastian Möller. 2024. Newspolyml: Multi-lingual european news fake assessment dataset. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, MAD '24, page 82–90, New York, NY, USA. Association for Computing Machinery.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, Singapore. 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. 2023. 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.

Dorian Quelle, Yi-Xiang Chen, Alexandre Bovet, and Scott A. Hale. 2025. Lost in translation: using global fact-checks to measure multilingual misinformation prevalence, spread, and evolution. *EPJ Data Sci.*, 14(1):22.

Gautam Kishore Shahi and Durgesh Nandini. 2020. *FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19*. ICWSM.

Hanna Shcharbakova, Tatiana Anikina, Natalia Skachkova, and Josef Van Genabith. 2025. When scale meets diversity: Evaluating language models on fine-grained multilingual claim verification. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 69–84, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative large language models in automated fact-checking: A survey. *arXiv preprint arXiv:2407.02351*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Xinyu Wang, Wenbo Zhang, and Sarah Rajtmajer. 2024. Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024. Do we need language-specific fact-checking models? the case of chinese. In *Conference on Empirical Methods in Natural Language Processing*.