# When Scale Meets Diversity: Evaluating Language Models on Fine-Grained Multilingual Claim Verification

**Hanna Shcharbakova**[1],
hash00004@stud.uni-saarland.de
**Tatiana Anikina**[2], **Natalia Skachkova**[2], **Josef van Genabith**[1,2]
{tatiana.anikina, natalia.skachkova, josef.van_genabith}@dfki.de

[1]Saarland University
[2]German Research Center for Artificial Intelligence (DFKI)

## Abstract

The rapid spread of multilingual misinformation requires robust automated fact verification systems capable of handling fine-grained veracity assessments across diverse languages. While large language models have shown remarkable capabilities across many NLP tasks, their effectiveness for multilingual claim verification with nuanced classification schemes remains understudied. We conduct a comprehensive evaluation of five state-of-the-art language models on the X-Fact dataset, which spans 25 languages with seven distinct veracity categories. Our experiments compare small language models (encoder-based XLM-R and mT5) with recent decoder-only LLMs (Llama 3.1, Qwen 2.5, Mistral Nemo) using both prompting and fine-tuning approaches.[1] Surprisingly, we find that XLM-R (270M parameters) substantially outperforms all tested LLMs (7-12B parameters), achieving 57.7% macro-F1 compared to the best LLM performance of 16.9%. This represents a 15.8% improvement over the previous state-of-the-art (41.9%), establishing new performance benchmarks for multilingual fact verification. Our analysis reveals problematic patterns in LLM behavior, including systematic difficulties in leveraging evidence and pronounced biases toward frequent categories in imbalanced data settings. These findings suggest that for fine-grained multilingual fact verification, smaller specialized models may be more effective than general-purpose large models, with important implications for practical deployment of fact-checking systems.

## 1 Introduction

The rapid spread of misinformation on the internet has become a critical challenge in today's digital age (Scheufele and Krause, 2019; Fung et al., 2022). With the increasing amount of false information being shared across different languages and platforms, automated fact verification systems have emerged as useful tools for maintaining information reliability.

The field of automated fact verification has seen significant progress in recent years, particularly with the advent of large language models and transformer-based architectures (Guo et al., 2022). However, most of these advancements have been predominantly focused on English-language content (Singhal et al., 2024; Dmonte et al., 2024), creating a significant gap in addressing misinformation in other languages.

Multilingual fact verification presents fundamental challenges for NLP (Dmonte et al., 2024; Wang et al., 2024; Zhang et al., 2024), particularly when employing fine-grained classification schemes that better capture the nuanced nature of truth assessment (Gupta and Srikumar, 2021; Pelrine et al., 2023; Mohtaj et al., 2024). While existing datasets and approaches employ various classification systems, classification beyond binary (*true/false*) and ternary (*true/false/other*) categories remains understudied across multiple languages.

The multi-category nature of this task bears conceptual similarity to Natural Language Inference (NLI) tasks (Poliak et al., 2018), though claim verification differs in its specific objectives. While NLI focuses on determining entailment relationships (*entails, contradicts, neutral*) between premise and hypothesis, our task requires assessing veracity across different distinct truth categories that reflect professional fact-checking standards.

In this work, we examine the performance of diverse model architectures and sizes on multilingual claim verification with fine-grained truth categories. We benchmark language model performance on the X-Fact dataset (Gupta and Srikumar, 2021) spanning multiple languages with seven distinct veracity categories, contrasting encoder-based model XLM-

[1]We consider a large language model (LLM) to be any model with more than 1B parameters, and correspondingly, small language model (SLM) to have less than 1B parameters.

R base (Conneau et al., 2020), encoder-decoder architecture mT5 base (Xue et al., 2021), and recent decoder-only models Llama 3.1 8B (Dubey et al., 2024), Qwen 2.5 7B (Yang et al., 2024), and Mistral Nemo 12B (Mistral AI Team, 2024).[2] For smaller models, we employ standard fine-tuning, while for larger models, we use both parameter-efficient fine-tuning with LoRA (Hu et al., 2022a) and carefully engineered few-shot prompting approaches. We evaluate models under two conditions: using claims alone and using claims with accompanying evidence text, which allows us to assess both inherent verification capabilities and evidence-augmented reasoning across models using a classification scheme that better reflects the nuanced assessments made by professional fact-checkers.

Our contributions include:

- We conduct comprehensive benchmarking of five state-of-the-art language models on the challenging seven-category multilingual X-Fact dataset, achieving new state-of-the-art results with a 15.8% improvement in macro-F1 score over previous best performance reported by Gupta and Srikumar (2021). We reveal a substantial performance gap between encoder-based and decoder-only architectures despite the latter's greater size and general capabilities.

- We provide analysis of model behaviors and error patterns across architectures, identifying several factors that appear to influence multilingual fact verification performance. These observations may help inform future research on verification approaches for diverse languages.

## 2 Related Work

### 2.1 Multilingual Fact Verification Datasets

While a substantial portion of fact verification research has centered on English-language content (Guo et al., 2022; Singhal et al., 2024; Dmonte et al., 2024), several datasets have emerged to address the multilingual dimensions of this challenge. These datasets vary significantly in size, language coverage, and labeling schemes.

Multilingual datasets include FakeCovid (Shahi and Nandini, 2020), covering 5K claims across 40 languages, and MM-COVID (Li et al., 2020), which provides 11K articles in English, Spanish, Portuguese, Hindi, French, and Italian. The MultiClaim dataset (Pikuliak et al., 2023) contains 28K social media posts in 27 languages that can be leveraged for fact verification tasks. FbMultiLingMisinfo (Barnabò et al., 2022) offers 7K news articles spanning 37 languages, while NewsPolyML (Mohtaj et al., 2024) includes 32K claims across English, German, French, Spanish, and Italian. The X-Fact dataset (Gupta and Srikumar, 2021) provides 31K claims from fact-checking websites in 25 languages across 11 language families.

Labeling approaches range from binary classification (Li et al., 2020; Barnabò et al., 2022) to three-category systems (Nørregaard and Derczynski, 2021; Hu et al., 2022b; Ullrich et al., 2023) and more complex multi-class schemes including 11 categories in FakeCovid (Shahi and Nandini, 2020), 4 in NewsPolyML (Mohtaj et al., 2024), and 7 in X-Fact (Gupta and Srikumar, 2021). The diversity of annotation schemes, while enabling finer-grained veracity assessments, complicates cross-dataset training and evaluation for cross-lingual verification.

### 2.2 Methods for Fact Verification

The task of claim verification has evolved significantly with various methodological approaches emerging to tackle the complexities of determining claim veracity. Transformer-based architectures (Devlin et al., 2019) brought substantial advancements to fact verification. Gupta and Srikumar (2021) evaluated mBERT-based models on the X-Fact dataset spanning 25 languages with 7-way classification. Their best model achieved an F1 score of 41.9% on the in-domain test set, though performance dropped to 16.2% F1 on out-of-domain and 16.7% F1 on zero-shot test sets, highlighting cross-lingual generalization challenges.

Recent research has explored large language models for fact verification using various approaches. For prompting-based methods, Cao et al. (2023) investigated different prompting strategies for fact-checking, finding that carefully crafted prompts with explicit instructions about expected output formats and task definitions significantly improved performance. Hu et al. (2023) found that increasing few-shot examples beyond a certain threshold provides substantial gains, suggesting a threshold effect. Self-consistency methods

[2]Further details on the specific model versions are provided in Appendix A.

using majority voting from multiple LLM runs improved performance, while self-refinement strategies where models iteratively refine their answers showed gains over standard approaches.

Chain of Thought (CoT) approaches have shown promising results by enabling LLMs to articulate reasoning processes before reaching conclusions (Wei et al., 2022). Hu et al. (2023) found that CoT prompting significantly improved performance across all tested models on English data compared to standard prompting.

Pelrine et al. (2023) compared GPT-4 against traditional approaches across multiple datasets. For binary classification on English LIAR (Wang, 2017), GPT-4 variants outperformed traditional models like ConvBERT (Jiang et al., 2020) and BERT. However, in multi-way classification tasks, performance declined significantly with traditional models like DeBERTa (He et al., 2021) showing better results. The same study demonstrated GPT-4's cross-lingual capabilities on CT-FAN-22 (Shahi et al., 2021), with GPT-4 substantially outperforming RoBERTa-L (Liu et al., 2019) on English multi-way classification.

Cekinel et al. (2024) found that fine-tuning LLaMA-2 models (Touvron et al., 2023) on Turkish language data outperformed cross-lingual transfer methods for fact verification. Their fine-tuned model achieved strong performance on binary classification, while cross-lingual prompting with English data showed improvements but proved less effective than language-specific fine-tuning. Mohtaj et al. (2024) evaluated multiple models on the NewsPolyML dataset spanning five European languages with four veracity categories. Interestingly, mBERT achieved the highest performance, suggesting that model size does not necessarily correlate with performance in multilingual fact verification tasks.

## 3 Dataset

For our experiments, we use the X-Fact dataset (Gupta and Srikumar, 2021). This dataset was selected due to several advantages over other multilingual fact verification resources. X-Fact encompasses a broad range of topics from verified fact-checking websites, making it more representative of real-world misinformation challenges compared to specialized datasets like FakeCovid (Shahi and Nandini, 2020) that focus solely on COVID-19 related claims. Unlike datasets derived from social media platforms such as FbMultiLingMisinfo (Barnabò et al., 2022), X-Fact provides ready-to-use data without requiring access to platform-specific APIs, ensuring reproducibility of research findings.

X-Fact comprises 31,189 claims across 25 languages from 11 language families, including Indo-European, Afro-Asiatic, Austronesian, Kartvelian, Dravidian, and Turkic. The dataset was carefully constructed by identifying reliable fact-checking sources from the International Fact-Checking Network[3] and Duke Reporter's Lab[4], excluding websites that conduct fact-checks in English to avoid overlap with existing datasets. Each claim in X-Fact is accompanied by up to 5 pieces of evidence extracted from fact-checking articles, with an average of 4.75 non-empty evidence pieces per claim. The dataset also includes valuable metadata such as the language of the claim and evidence, the fact-checking site where the claim was derived from, links to the evidence where they were published, claim date, review date, and claimant information. Examples of claims, corresponding evidence, and associated metadata can be found in Appendix B.

To ensure consistent evaluation across different fact-checking standards, the dataset employs a standardized seven-label classification scheme: *true, mostly true, partly true/misleading, mostly false, false, complicated/hard to categorize*, and *other*. This fine-grained approach provides a more nuanced assessment of claim veracity compared to less fine-grained classification schemes used in many other datasets.

The dataset is divided into multiple subsets designed to evaluate different aspects of model performance (see Table 1). The training set contains 19,079 claims across 13 languages, while the development set comprises 2,535 claims spanning 13 languages. For testing, X-Fact provides three separate subsets: an in-domain test set with 3,826 claims from the same languages and sources as the training data; an out-of-domain test set containing 2,368 claims from the same languages but different sources; and a zero-shot test set featuring 3,381 claims from 12 languages not present in the training data. This evaluation framework supports a thorough assessment of models' generalization capabilities across both domains and languages.

The label distribution in the X-Fact exhibits

[3]https://www.poynter.org/ifcn/.
[4]https://reporterslab.org/.

| Dataset Subset | # Claims | # Languages |
|---|---|---|
| Training | 19079 | 13 |
| Development | 2535 | 13 |
| In-domain | 3826 | 13 |
| Out-of-domain | 2368 | 5 |
| zero-shot | 3381 | 12 |

Table 1: Overview of the X-Fact dataset subsets.

significant variation across all subsets (see Figure 1). The *false* label dominates the training set with 7,515 instances (39.4%), followed by *partly true/misleading* with 4,359 instances (22.8%). The least represented label is *other* with only 576 instances (1.9%).

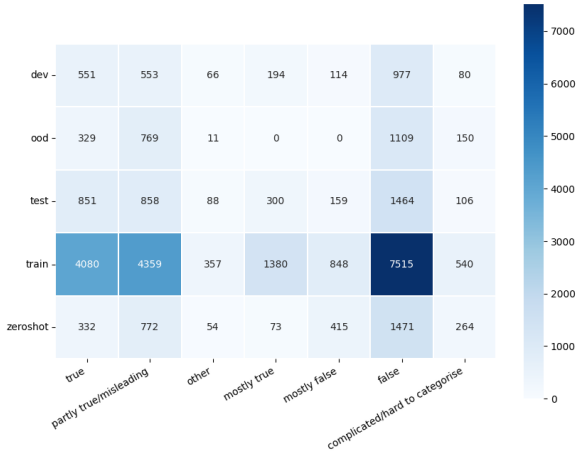


Figure 1: Distribution of data in X-Fact by label across subsets.

The language distribution also shows substantial variation across different subsets (see Figure 2). Portuguese dominates the training set with 5,601 claims (29.4%), followed by Indonesian with 2,231 claims (11.7%) and Arabic with 1,567 claims (8.2%). Serbian has the lowest representation with only 624 claims (3.3%).

These imbalances may potentially impact model learning, particularly for cross-lingual transfer, and present additional challenges for models to learn fine-grained veracity categories.

## 4 Experiments

### 4.1 Experimental Setup

Our evaluation focuses on benchmarking different language model architectures on the multilingual fact verification task, using X-Fact's seven-category classification scheme across multiple languages. We evaluate both small language models (SLMs, <1B parameters) and large language models (LLMs, >1B parameters) to determine their relative effectiveness for fine-grained multilingual verification. Table 2 provides an overview of the models evaluated in this study.

We selected these models based on their strong multilingual capabilities and architectural diversity. XLM-R was chosen for its robust pre-training on 100 languages and encoder-only architecture that has proven effective for classification tasks. MT5 represents the encoder-decoder paradigm, offering a different architectural approach while maintaining strong multilingual capabilities across 101 languages. For LLMs, we selected Llama 3.1 8B, Qwen 2.5 7B, and Mistral Nemo 12B to represent state-of-the-art decoder-only architectures with varying degrees of multilingual support.

We prioritized open-source models with moderate parameter sizes to ensure reproducibility and facilitate deployment in resource-constrained environments. This selection allows us to evaluate whether sophisticated reasoning in current LLMs transfers effectively to multilingual fact verification compared to smaller, specialized architectures like XLM-R.

| Model | # Par. | # Lang. | Architecture |
|---|---|---|---|
| XLM-R | 270 M | 100 | Encoder-only |
| mT5 | 580 M | 101 | Encoder-decoder |
| Llama 3.1 | 8 B | 8 | Decoder-only |
| Qwen 2.5 | 7 B | 29 | Decoder-only |
| Mistral Nemo | 12 B | 11 | Decoder-only |

Table 2: Models evaluated on multilingual fact verification using the X-Fact dataset. # Par. is the number of parameters and # Lang. is the number of languages supported by each model.

For the SLMs, we performed fine-tuning experiments, while for LLMs, we explored both direct prompting and parameter-efficient fine-tuning using LoRA. The models' implementation details can be found in the Appendix C.

### 4.2 Small Language Models Experiments

For XLM-R and mT5, we conducted two types of fine-tuning experiments:

- **Full Model Fine-tuning:** We performed complete fine-tuning of the models, allowing all parameters to be updated during training.

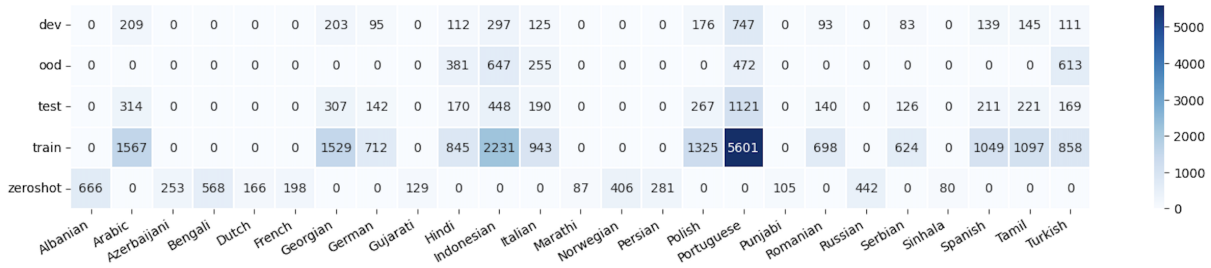


Figure 2: Distribution of data in X-Fact by language across subsets.

- **Classification Head Fine-tuning:** We fine-tuned only the classification head while keeping the base model frozen.

For both approaches, we provided the models with the claim text and evidence as input. We did not conduct experiments using only claim text without evidence, as preliminary experiments confirmed the X-Fact paper's finding that claim-only setups yield worse performance.

### 4.3 Large Language Models Experiments

For LLMs, we explored both few-shot prompting and parameter-efficient fine-tuning approaches. We evaluated each model in two input configurations: (1) claim-only, providing only the claim text; and (2) claim with evidence, providing both claim and evidence text. Our experimental setup included the following approaches:

- **Few-shot prompting:** We developed 7-shot prompts containing examples for each veracity category to guide prediction without training. Each prompt included clear instructions, category definitions, and was tested in both claim-only and claim+evidence variants.

- **LoRA fine-tuning:** We implemented parameter-efficient fine-tuning using LoRA for both claim-only and claim+evidence configurations.

The optimized prompt template is provided in the Appendix E.

## 5 Results

### 5.1 SLMs Performance

Table 3 presents the macro-F1 scores for small language models across three evaluation subsets. XLM-R with full fine-tuning achieves the highest performance with 57.7% macro-F1 on the test set, substantially outperforming the previous state-of-the-art mBERT baseline (41.9%) by 15.8% reported in Gupta and Srikumar (2021). XLM-R also demonstrates superior cross-domain and cross-lingual generalization, maintaining relatively strong performance across all evaluation subsets.

| Model | Test | OOD | Zero-shot |
|---|---|---|---|
| mBERT (baseline) | 41.9 | 16.2 | 16.7 |
| XLM-R frozen | 51.4 | 40.8 | 41.3 |
| XLM-R | **57.7** | **47.6** | **43.2** |
| mT5 | 47.6 | 22.2 | 19.2 |

Table 3: SLMs performance on the X-Fact dataset (macro-F1 scores). XLM-R frozen refers to fine-tuning the classification head only. mBERT performance is derived from (Gupta and Srikumar, 2021).

MT5 reaches 47.6% macro-F1 on the test set but shows poor generalization to out-of-domain (22.2%) and zero-shot (19.2%) scenarios. The performance gap between XLM-R and mT5 widens significantly on these evaluation sets, indicating that XLM-R's encoder-only architecture may be better suited for multilingual fact verification tasks.

### 5.2 LLMs Performance

Table 4 presents LLMs' results across different configurations. Despite their significantly larger size (7-12B parameters), all LLMs substantially underperform compared to SLMs. The best LLM configuration (Qwen claim-only fine-tuning) achieves only 16.9% macro-F1 on the test set - 40.8% points below XLM-R. For visualizations of models' performance across different evaluation subsets, refer to Appendix D.

Among the LLMs, Qwen 2.5 consistently demonstrates the best performance across most configurations. The model achieves its highest macro-F1 score of 16.9% with claim-only fine-tuning on the test set, compared to 15.9% with claim+evidence fine-tuning and 11.4%-12.7% with

| Method | Few-shot | | | | LoRA-based Finetune | | | |
|---|---|---|---|---|---|---|---|---|
| | Claim+Evidence | | Claim Only | | Claim+Evidence | | Claim Only | |
| | macro | micro | macro | micro | macro | micro | macro | micro |
| *Qwen 2.5* | | | | | | | | |
| Test | 12.7 | 24.9 | 11.4 | 18.6 | 15.9 | **39.5** | **16.9** | 29.6 |
| OOD | 13.0 | 29.6 | 11.2 | 27.4 | **15.1** | **47.1** | 11.1 | 31.3 |
| Zero-shot | 10.9 | 18.9 | 12.9 | 23.9 | **15.4** | **35.8** | 11.7 | 24.5 |
| *Mistral Nemo* | | | | | | | | |
| Test | **14.8** | 30.8 | 8.5 | 23.4 | 14.6 | **31.9** | 10.3 | 20.2 |
| OOD | **16.1** | **42.6** | 9.7 | 36.6 | 12.1 | 34.2 | 9.6 | 27.1 |
| Zero-shot | **15.1** | 28.7 | 10.6 | **29.6** | 12.9 | 25.7 | 8.2 | 15.6 |
| *Llama 3.1* | | | | | | | | |
| Test | 14.0 | **32.0** | 10.8 | 18.4 | 14.3 | 27.6 | **15.5** | 30.5 |
| OOD | 13.3 | **41.2** | 8.7 | 21.2 | 11.2 | 27.1 | **13.5** | 33.2 |
| Zero-shot | **12.9** | **30.1** | 8.7 | 17.6 | 9.6 | 17.5 | 12.1 | 29.4 |

Table 4: LLMs performance on the X-Fact dataset (macro-F1 and micro-F1 scores). Bold values indicate the highest macro- and micro-F1 scores for each model-subset combination.

few-shot prompting. LoRA-based fine-tuning consistently improves performance over few-shot inference across all models and configurations, with Qwen 2.5 showing the largest gains.

The impact of adding evidence to claims varies significantly across models and methods. For Qwen 2.5, fine-tuning with claim-only (16.9%) outperforms claim+evidence (15.9%) on the test set, showing a consistent pattern across all evaluation sets. In contrast, Mistral Nemo generally performs better with claim+evidence input in few-shot settings (14.8% vs 8.5% on test set) but shows mixed results with fine-tuning. Llama 3.1 demonstrates the most inconsistent performance across different configurations. While it achieves reasonable performance on the test set (15.5% macro-F1 with claim-only fine-tuning), it shows the largest performance drop on the zero-shot set, with the worst configuration (claim+evidence fine-tuning) falling to 9.6% macro-F1.

**LoRA Fine-tuning and Few-shot Prompting.** Fine-tuning consistently improves performance over few-shot prompting across all models. Qwen shows the most substantial improvement (from 12.7% in few-shot with claim+evidence setting to to 15.9% in fine-tuning with claim+evidence setting) on the test set. Mistral Nemo shows minimal differences between methods, with some configurations favoring few-shot prompting (16.1% vs 12.1% on out-of-domain with claim+evidence). Llama 3.1 generally benefits from fine-tuning, improving from 10.8% to 15.5% in the claim-only configuration on the test set.

**Performance Across Evaluation Subsets.** All models show declining performance from test to out-of-domain and zero-shot sets when fine-tuning. Qwen 2.5 maintains stable performance, with the sharpest drop by 5.2% from test to zero-shot. Mistral Nemo shows the least variation, performing best on out-of-domain (16.1%). Llama 3.1 exhibits the largest degradation, dropping from 15.5% on test to 9.6% on zero-shot in comparable configurations. Refer to the Appendix F for visualizations comparing LLMs performance across these evaluation subsets. For a combined view of claim+evidence configurations across all LLMs, refer to Appendix G, which directly compares the macro-F1 scores across evaluation subsets and highlights the best performing method for each model.

**Macro vs. Micro F1 Score.** The substantial gap between micro- and macro-F1 scores is consistent across all LLMs, with the largest gaps observed in fine-tuning configurations. Qwen 2.5 achieves 39.5% micro-F1 compared to 15.9% macro-F1 in its claim+evidence fine-tuning on the test set, a gap of 23.6%. Similarly, Mistral Nemo shows a 17.2% gap in its claim+evidence fine-tuning configuration on the test set.

Few-shot configurations generally show smaller gaps. For instance, Qwen's few-shot claim+evidence on the test set shows an 12.2% gap, while its fine-tuning equivalent shows a 23.6% gap. This pattern holds across all models

where fine-tuning configurations consistently exhibit gaps ranging from 15 to 32 percentage points, while few-shot configurations typically show gaps between 8 to 20 percentage points. The visualizations depicting the performance gaps between macro- and micro-F1 scores across LLMs can be found in Appendix H.

## 6 Discussion

Our comprehensive evaluation across 25 languages reveals several important findings that advance our understanding of how different architectures handle fine-grained veracity classification across languages.

**Performance Gap Between Model Types.** The most striking finding is XLM-R's superiority over all tested LLMs, achieving 57.7% macro-F1 compared to the best LLM performance of 16.9% from Qwen 2.5. This performance difference is particularly noteworthy given that LLMs contain many more parameters than XLM-R. XLM-R was pretrained on 100 languages using a masked language modeling objective that may align well with classification tasks, whereas LLMs use next-token prediction objectives optimized for text generation. These differences in pre-training approaches and objectives may contribute to the observed performance gap.

While our comparison involves different training methodologies (full fine-tuning for SLMs versus LoRA for LLMs), it is important to note that even when comparing more similar approaches, substantial performance gaps persist. Our frozen XLM-R configuration, which only updates the classification head similar to LoRA's parameter-efficient approach, still achieves 51.4% macro-F1 compared to the best LLM performance of 16.9%. This suggests that the performance differences extend beyond training methodology. Future work should include detailed per-label performance analysis to better understand model biases and identify which veracity categories prove most challenging across different architectures.

**Evidence Integration Patterns.** A clear pattern emerges in how LLMs handle evidence: surprisingly, incorporating additional evidence often does not enhance performance and can even lead to worse results. For instance, Qwen's claim-only fine-tuning (16.9%) outperforms its claim+evidence configuration (15.9%). This pattern persists across all Llama 3.1 configurations, suggesting systematic difficulties in leveraging additional context for verification decisions.

We hypothesize several factors that may contribute to this counterintuitive finding. First, the architectural limitations of decoder-only LLMs may hinder effective evidence integration. Unlike XLM-R's bidirectional attention that allows simultaneous consideration of all evidence elements against all claim components, LLMs' autoregressive attention can only consider previous tokens. This sequential processing creates a tendency to forget or ignore earlier information as sequences become longer, making balanced evidence evaluation more challenging.

Second, our input formatting may have contributed to this issue. While we used clear demarcation between claims and evidence in our prompts (as shown in Appendix E), we did not implement more sophisticated structuring techniques that might have helped LLMs better distinguish and compare these elements. Context window size was treated as a hyperparameter in our experiments, with LLMs tested at both 2048 and 4096 tokens, while SLMs were evaluated with context windows ranging from 256 to 512 tokens. With evidence pieces having median lengths of 25-35 words each and approximately 4.75 pieces per claim on average, the evidence was fully accommodated within the context windows of all models. Therefore, evidence truncation was not a contributing factor to the observed performance patterns.

This finding is particularly significant because evidence-based reasoning is fundamental to reliable fact verification. The fact that simply providing claims yields better results than including supporting evidence indicates that current LLMs may not be effectively utilizing the additional information or may be getting confused by the increased input complexity.

**Fine-Grained Classification Challenges.** The severe data imbalances in X-Fact likely contributes to the observed performance patterns. The dominance of *false* and *partly true/misleading* categories creates a challenging environment for models to learn effective representations for less frequent but equally important categories. This imbalance effect is aggravated in the seven-category setting, where models must not only distinguish between *true* and *false* but also navigate subtle gradations of partial truth. Furthermore, the language distribution imbalance (Portuguese comprising 29.4% of training data while Serbian represents only 3.3%) likely

impacts cross-lingual performance. Models may develop language-specific biases that hinder their ability to generalize across languages, particularly to those underrepresented in the training data.

The substantial disparity between micro- and macro-F1 scores across all LLMs reveals critical limitations in handling nuanced veracity categories. The micro-F1 scores being consistently higher than macro-F1 scores confirms that performance is driven primarily by accuracy on frequent categories, while rare categories remain poorly predicted. This pattern is particularly pronounced in LLMs, suggesting they may be more influenced by biases in the training data than the fine-tuned XLM-R.

**Cross-Lingual and Cross-Domain Generalization.** Performance degradation across evaluation subsets is consistent across all models but varies in magnitude. XLM-R demonstrates the most robust cross-lingual transfer, while LLMs show steeper drops. The relatively stable performance of XLM-R on unseen languages suggests that its multilingual pre-training provides effective cross-lingual representations for fact verification task. The sharper declines observed in LLMs may indicate that their multilingual capabilities are less robust when faced with languages not well represented in their training data or when transferring across different fact-checking domains.

Even when comparing XLM-R's frozen configuration (which only updates the classification head, similar to LoRA's parameter-efficient approach), we still observe substantial outperformance over LLMs (51.4% vs 16.9% best LLM performance). This suggests that the performance differences may stem not only from the fine-tuning methodology but also from other factors such as architectural advantages of encoder-based models for this specific task or the amount and quality of the pre-training data available in different languages.

## 7 Conclusion and Future Work

This work presents a comprehensive evaluation of diverse language model architectures (small and large; encoder, encoder-decoder, and decoder-only) on multilingual fact verification using the challenging seven-category X-Fact dataset. Our findings reveal several key insights that advance understanding of how different models handle fine-grained veracity classification across languages.

Fully fine-tuned XLM-R emerges as the clear winner, achieving 57.7% macro-F1 on the test set – a 15.8% improvement over previous state-of-the-art. Despite having significantly fewer parameters, XLM-R substantially outperforms all tested LLMs, with the best LLM (Qwen 2.5) reaching only 16.9% macro-F1. The magnitude of this performance gap persists even when comparing lightweight fine-tuning approaches (e.g., frozen XLM-R with a trained classification head: 51.4% vs best LLM: 16.9%), suggesting that factors beyond training methodology contribute to the observed differences. However, the exact nature of these factors requires further investigation.

Our analysis reveals problematic patterns in LLM behavior, particularly their inability to effectively utilize additional evidence. Models often perform worse when provided with claim-evidence pairs compared to claims alone, indicating systematic challenges in leveraging external information for verification decisions. This limitation is particularly problematic given that evidence-based reasoning is fundamental to reliable fact-checking, though the underlying causes of this behavior need deeper exploration.

The significant disparity between micro- and macro-F1 scores across the models reveals the challenge of handling imbalanced datasets with fine-grained categories. Models tend to learn shortcuts based on frequent categories while struggling with rare but equally important veracity labels. This bias appears more pronounced in LLMs, indicating they may be more vulnerable to dataset imbalances than smaller models that have been carefully fine-tuned.

These findings have important implications for the development of multilingual fact verification systems. While LLMs show promise for many NLP tasks, our results suggest that for fine-grained fact verification across languages, smaller specialized models may provide better performance while requiring fewer computational resources.

## 8 Limitations

Our study has several limitations that should be considered when interpreting the results.

**Training Methodology Differences.** Our comparison involves fundamentally different training approaches: XLM-R undergoes full fine-tuning with all parameters being updated, while LLMs utilize LoRA that freezes the majority of the original model parameters. This methodological difference could significantly impact the ability of LLMs to

adapt to the specific task requirements and may partially explain the observed performance gaps.

**Prompt Engineering Constraints.** Our prompt engineering approach may not be equally optimal across all languages in our multilingual evaluation. While we developed carefully engineered 7-shot prompts with examples balanced across the seven veracity categories, our prompt design focused primarily on ensuring representative coverage of each label rather than optimizing for linguistic diversity. This approach may have favored certain languages or language families that were better represented in our example selection. Language-specific prompt optimization could potentially narrow the performance gap, though this would require substantial additional engineering effort for each target language.

**Evidence Interpretation Limitations.** Given the relatively small performance differences between claim-only and claim+evidence configurations, we cannot definitively conclude that LLMs are incapable of evidence utilization. The limited performance gap may simply reflect the inherent difficulty of the task or limitations in our evaluation approach. It's possible that with more sophisticated prompting strategies, larger datasets, or alternative evidence presentation formats, LLMs might demonstrate improved evidence integration capabilities. The evidence quality in the X-Fact dataset may also play a role, as analysis reveals that search snippets may not always contain sufficient information for accurate verification (Gupta and Srikumar, 2021).

**Practical Computing Considerations.** Our comparison between fully fine-tuned XLM-R and LoRA-adapted LLMs reflects realistic scenario with limited computational resources. Full fine-tuning of billion-parameter models requires substantial computational resources that are often prohibitive for many researchers. In contrast, parameter-efficient methods like LoRA can be applied with modest computational resources, making them the more practical choice for deploying large models. This comparison addresses a critical question: given realistic computational constraints, which approach provides better performance for multilingual fact verification? Our results demonstrate that a smaller, fully fine-tuned model can significantly outperform much larger models adapted with parameter-efficient methods, suggesting that for specific tasks like multilingual fine-grained verification, specialized smaller models may be preferable to general-purpose large models.

**Output Analysis and Reproducibility.** To enhance reproducibility and enable further investigation of the observed performance patterns, we make our LLM outputs available in the repository[5], including detailed predictions and model responses. A comprehensive analysis of these outputs, including confusion matrices and detailed error patterns that could reveal potential parsing issues or systematic biases, represents important future work that could provide deeper insights into the substantial performance differences observed between model architectures.

## 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

Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. 2022. Fbmultilingmisinfo: Challenging large-scale multilingual benchmark for misinformation detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are large language models good fact checkers: A preliminary study. *Preprint*, arXiv:2311.17355.

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.

[5] https://github.com/Aniezka/xfact-fever.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *Preprint*, arXiv:2310.05177.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022b. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Convbert: Improving bert with span-based dynamic convolution. In *Advances in Neural Information Processing Systems*, volume 33, pages 12837–12848. Curran Associates, Inc.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *Preprint*, arXiv:2011.04088.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

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.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

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

Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. Overview of the clef-2021 checkthat! lab task 3 on fake news detection. *Working Notes of CLEF*.

Aryan Singhal, Thomas Law, Coby Kassner, Ayushman Gupta, Evan Duan, Aviral Damle, and Ryan Luo Li. 2024. Multilingual fact-checking using LLMs. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 13–31, Miami, Florida, USA. 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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Herbert Ullrich, Jan Drchal, Martin Rýpar, Hana Vincourová, and Václav Moravec. 2023. Csfever and ctkfacts: acquiring czech data for fact verification. *Language Resources and Evaluation*, 57(4):1571–1605.

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. *Preprint*, arXiv:2410.18390.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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, and 1 others. 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1899–1914, Miami, Florida, USA. Association for Computational Linguistics.

## A Model Implementation Details

For our experimental evaluation, we used the following model versions:

**XLM-R base**. We used *FacebookAI/xlm-roberta-base* (270 million parameters) model from Hugging Face, which has been pre-trained on text in 100 languages. The model was tested in two configurations: (1) with frozen parameters and only the classification head fine-tuned, and (2) with full fine-tuning of all parameters.

**mT5 base**. We employed the *google/mt5-base* model (580 million parameters) from Hugging Face, which follows an encoder-decoder architecture and has been pre-trained on multilingual text.

**Llama 3.1 8B**. We used the instruction-tuned version of Llama 3.1 with 8 billion parameters. This model officially supports seven languages: French, German, Hindi, Italian, Portuguese, Spanish, and Thai, in addition to English.

**Qwen 2.5 7B**. We employed the instruction-tuned Qwen 2.5 model with 7 billion parameters. This model supports 29 languages and has demonstrated strong performance in both English and multilingual tasks.

**Mistral Nemo 12B**. We used the Mistral Nemo model with 12 billion parameters. This model supports 11 languages: English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi.

All experiments with LLMs were conducted using the Unsloth library (Daniel Han and team, 2023) to efficiently implement and optimize the fine-tuning and inference processes, ensuring faster training times and reduced memory usage without compromising model performance.

## B Details on X-Fact

For examples from the X-Fact dataset, please refer to the Figure 3.

## C Hyperparameter Details

### C.1 Small Language Models

For our small language models (XLM-R and mT5), we employed Bayesian hyperparameter optimization through Weights&Biases, conducting 90 sweeps for the classification head approach and 60

| | |
|---|---|
| **Claim** | *Muslimische Gebete sind Pflichtprogramm an katholischer Schule.*<br>Muslim prayers are compulsory in Catholic schools. |
| Label | Mostly-False (*Grösstenteils Falsch*) |
| Claimant | Freie Welt |
| Language | German |
| Source | de.correctiv.org |
| Claim Date | March 16, 2018 |
| Review Date | March 23, 2018 |
| **Claim** | *Temos, hoje, a despesa de Previdência Social representando 57% do orçamento.*<br>Today, we have Social Security expenses representing 57% of the budget. |
| Label | Partly-True (*Exagerado*) |
| Claimant | Henrique Meirelles |
| Language | Portuguese (Brazilian) |
| Source | pt.piaui.folha.uol.com.br |
| Claim Date | None |
| Review Date | May 2, 2018 |

Figure 3: Details of the X-Fact dataset. Examples from X-Fact as presented in the original paper by Gupta and Srikumar (2021). For reference, translations are also shown.

sweeps each for the full fine-tuning experiments. We used an AdamW optimizer with a polynomial learning rate scheduler. To prevent overfitting, we implemented early stopping. Table 5 shows the key hyperparameter values for each model variant.

| Model | Learning Rate | Batch Size |
|---|---|---|
| XLM-R frozen | 5.7e-04 | 8 |
| XLM-R | 1.82e-05 | 6 |
| mT5 | 2.2e-05 | 8 |

Table 5: Key hyperparameter values for SLMs.

### C.2 Large Language Models

For large language models, we used parameter-efficient fine-tuning with LoRA. Through systematic experimentation, we identified optimal LoRA configurations with a rank of 16 and adapter alpha of 32. We targeted both attention components (query, key, value, and output projections) and feed-forward layers (gate projections and up/down projections).

Lower rank values (r = 2, 4, 8) and alpha values (8, 16) produced inferior results, while increasing these parameters beyond our chosen values (r > 16, alpha > 32) provided negligible performance gains while substantially increasing memory requirements.

For prompt engineering, we tested various temperature settings and found that temperatures between 0.3 and 0.5 provided the best balance between confident predictions and appropriate uncertainty handling. Lower temperatures led to overly deterministic outputs that failed to capture nuanced veracity judgments, while higher temperatures resulted in inconsistent classifications.

All LLM experiments were conducted using 4-bit quantization to enable efficient processing on GPUs while maintaining performance.

## D Performance Comparison across Models

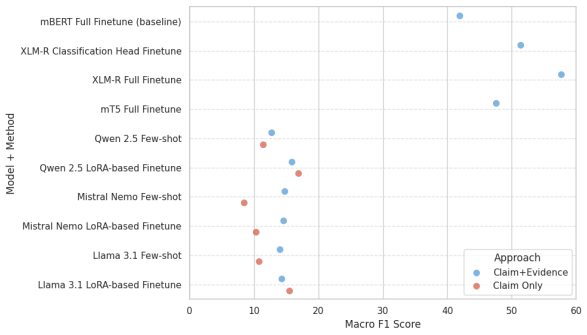


Figure 4: Macro-F1 scores across test subset by model.

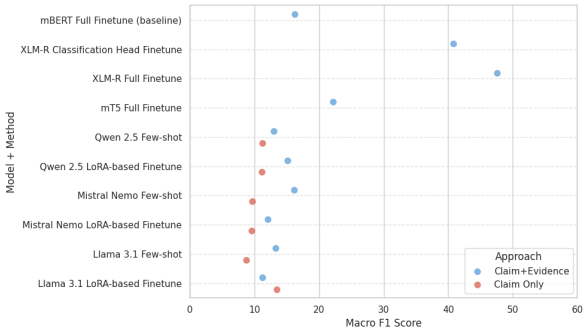


Figure 5: Macro-F1 scores across OOD subset by model.

## E Prompt Template

In Figure 7 we provide a prompt template used to instruct LLMs.

## F LLMs Performance Comparison Visualizations

In Figure 8 we provide a comparison of macro- and micro F1 scores across LLMs, evaluation subsets, and training methods.

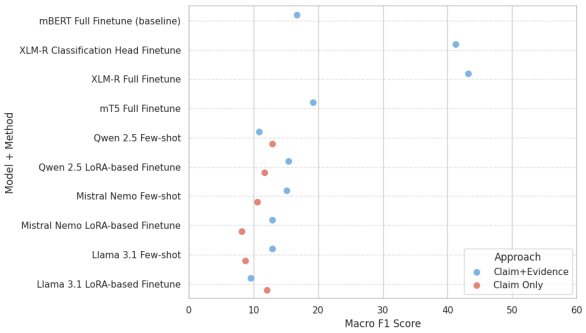


Figure 6: Macro-F1 scores across zero-shot subset by model.

## G LLMs Performance Summary Table

In Table 6 we present a combined comparison of macro-F1 scores for all evaluated models using claim+evidence configurations across the three evaluation subsets (Test, OOD, Zero-shot). This table extracts the claim+evidence results from Table 4 and combines them with the small language model performance to facilitate direct performance comparison.

## H Micro- and Macro-F1 Scores Comparison across LLMs

In Figure 9 we provide a comparison of average macro- and micro-F1 scores across LLMs for each evaluation subset.

| Method | Test | OOD | Zero-shot |
|---|---|---|---|
| mBERT (SLM) | 41.9 | 16.2 | 16.7 |
| XLM-R frozen (SLM) | 51.4 | 40.8 | 41.3 |
| XLM-R (SLM) | **57.7** | **47.6** | **43.2** |
| mT5 (SLM) | 47.6 | 22.2 | 19.2 |
| Qwen 2.5 Few-shot (LLM) | 12.7 | 13.0 | 10.9 |
| Qwen 2.5 LoRA (LLM) | **15.9** | **15.1** | **15.4** |
| Mistral Nemo Few-shot (LLM) | **14.8** | **16.1** | **15.1** |
| Mistral Nemo LoRA (LLM) | 14.6 | 12.1 | 12.9 |
| Llama 3.1 Few-shot (LLM) | 14.0 | **13.3** | **12.9** |
| Llama 3.1 LoRA (LLM) | **14.3** | 11.2 | 9.6 |

Table 6: Macro-F1 performance comparison across evaluation subsets for claim+evidence configurations. Bold values indicate the highest macro-F1 score for each LLM model across the two training methods (Few-shot vs LoRA). SLMs results included for reference.

Your task is to evaluate the given claim and evidence, then provide a verdict using one of the following labels: false (completely incorrect), true (completely correct), mostly true (mainly correct with minor issues), mostly false (mainly incorrect with minor true elements), partly true/misleading (mix of true and false elements), complicated/hard to categorise (cannot be verified with given evidence) or other (doesn't fit other categories).

***Q:*** Claim: In Ungheria le tasse sulle imprese sono al 9 per cento e sulle persone fisiche al 15 per cento, e l'Ungheria cresce del 5 per cento.\nEvidence: L'Ungheria, insieme ad altri paesi della Ue (Lussemburgo, Belgio, Olanda, ... Puntando su una tassazione dei redditi di forte vantaggio (9% per le società e 15% per le ... Per bilanciare la bassa imposizione fiscale su imprese e persone fisiche, ... L'Iva è generalmente al 27% anche se esistono aliquote al 18% e al 5%.
***A:*** Label: true
***Q:*** Claim: Das Coronavirus enthält HIV-Anteile, wurde also im Labor erschaffen.\nEvidence: Apr 26, 2020 — Paris – Es klingt wie eine wilde Verschwörungstheorie – und doch hat es der französische Virologe Luc Montagnier bei einer Fernsehdisk. Das Coronavirus enthält HIV-Anteile, wurde also im Labor erschaffen. Feb 6, 2020 — Im Internet kursieren wilde Theorien über den Ursprung des Virus. Dazu tragen auch fragwürdige „Forscher" bei. Schnelle Studien enthalten oft ...
***A:*** Label: false
***Q:*** Claim: "La velocidad promedio de Internet en 2015 era apenas de 4,5 megabits por segundo, hoy la triplicamos".\nEvidence: Mar 5, 2019 — Macri: "La velocidad promedio de Internet en 2015 era apenas de 4,5 megabits por segundo, hoy la triplicamos". ¿Es así? Leé el chequeo acá: ...
***A:*** Label: mostly true
***Q:*** Claim: „Trenutno se radi na popisu državne imovine..\nEvidence: Državna imovina u RH klasificira se, evidentira i vrednuje na neodgovarajući način. • Glavna knjiga Državne riznice ne ... prosinca svake godine provesti sveobuhvatni popis državne imovine kojom ... rad na izradi aplikacijskog rješenja za drugu fazu ISUDIO je u tijeku. (dovršenje se ... trenutno važećem Zakonu):. Poseban ...
***A:*** Label: mostly false
***Q:*** Claim: "ევროპული ღირებულებები" - იტალიის სამაშველო სამსახურებს მიგრანტების ჩაძირული გემების დახმარება აეკრძ...\nEvidence: Oct 4, 2018 — იტალიის სამაშველო სამსახურებს მიგრანტების ჩაძირული გემების დახმარება ულტრა-მემარჯვენე შინაგან საქმეთა ...
***A:*** Label: partly true/misleading
***Q:*** Claim: A evasão do Pronatec foi de 80%.\nEvidence: Palavras-Chave: Políticas públicas; Avaliação; Implementação; Pronatec ... os cursos FIC foi de 618, o que corresponde a 80,26 % do total de vagas ofertadas. ... Outra questão apontada como causa da evasão, foi a dificuldade, por boa parte ...
***A:*** Label: complicated/hard to categorise
***Q:*** Claim: Yoris Raweyai Bantah Terkait Tuntutan Pembubaran Banser.\nEvidence: Aug 25, 2019 — Anggota DPD terpilih Yorrys Raweyai menyebut hanya menerima selebaran tujuh poin tuntutan warga di Sorong, yang salah satunya meminta ...
***A:*** Label: other
***Q:*** *[claim and evidence]*
***A:*** Label:

Figure 7: Prompt template used for LLMs.

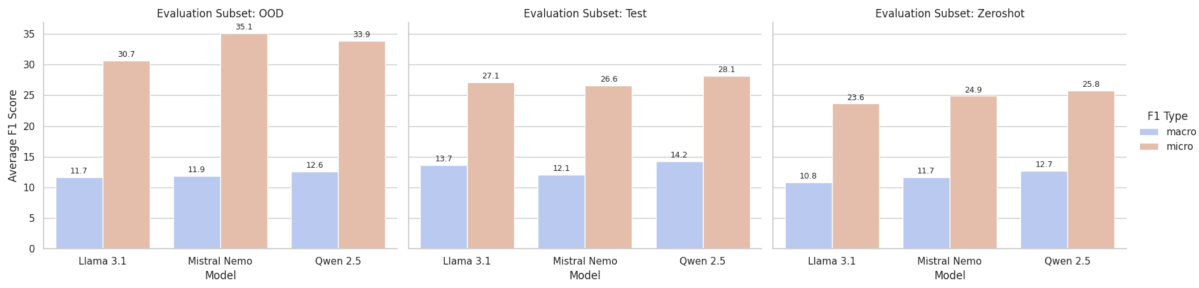


Figure 8: Comparison of macro- and micro F1 scores across LLMs, evaluation subsets, and training methods.

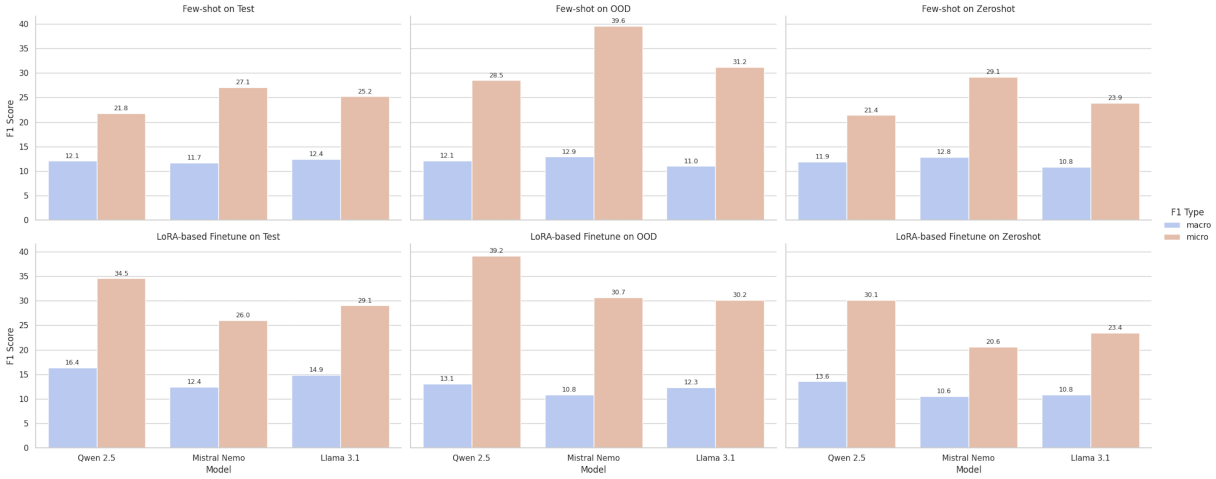


Figure 9: Comparison of average macro- and micro-F1 scores across LLMs for each evaluation subset.