FactSpan: Multilingual and Epistemic Disparities in LLM Fact-Checking

Lorraine Saju[1,2], Arnim Bleier[2], Jana Lasser[3,4], Claudia Wagner[2,1]

[1]RWTH Aachen University, Germany, [2]GESIS – Leibniz Institute for the Social Sciences, Germany
[3]University of Graz, Austria, [4]Complexity Science Hub, Vienna, Austria
Correspondence: lorraine.saju@rwth-aachen.de

Abstract

As online platforms increasingly rely on large language models (LLMs) for automated content moderation and fact-checking, it remains unclear how reliably these systems operate across multilingual and socially diverse information environments. Existing evaluations largely overlook how linguistic and epistemic context shape model behaviour, particularly in low-resource settings. We introduce **FactSpan**, a multilingual benchmark of 61,514 real-world claims spanning 30 languages and 17 years (2007–2024), with claim-level annotations for epistemic framing and temporal alignment relative to each model's training cutoff. Using this benchmark, we evaluate five widely used LLMs (GPT-4o, GPT-3.5 Turbo, Llama 3.1 8B/70B, Mixtral 8x7B) on zero-shot veracity classification. Performance drops significantly for low-resource languages and for factually framed claims, which are consistently harder to verify than opinion-oriented content. Importantly, models rarely abstain more under uncertainty; instead, they confidently misclassify factual claims. We further observe stable or improved performance on post-cutoff claims, suggesting reliance on contextual and stylistic heuristics rather than grounded factual knowledge. Overall, our findings show that LLM-based fact-checking remains uneven across linguistic and epistemic contexts, raising concerns about the reliability of automated moderation systems in multilingual online environments.

1 Introduction

Multilingual NLP evaluation has expanded rapidly in recent years, yet most benchmarks remain anchored to a narrow slice of the linguistic world. Fact-checking is a particularly challenging evaluation setting: it requires models to reason across languages, topics, and time, and to handle claims that vary not only in content but in how they are framed. Whether a claim is asserted as an objective fact or expressed as a subjective opinion is itself a linguistic signal, and one whose effect on model behaviour remains poorly understood.

LLMs are increasingly deployed in automated content moderation pipelines where they assess claims across linguistically diverse, socially embedded environments (Hoes et al., 2023; Huang et al., 2024). Yet the evaluations underpinning these deployments treat language as decontextualised text, ignoring the structural, rhetorical, and temporal factors that shape both how claims are produced and how models respond to them. Recent efforts such as M4FC (Geng et al., 2025) and the SemEval-2025 MultiClaim benchmark (Pikuliak et al., 2025) have expanded multilingual coverage, but systematic analysis of *how* and *why* model reliability breaks down across these contextual dimensions is still lacking.

We address this gap by introducing **FactSpan**, a multilingual benchmark of 61,514 real-world fact-checked claims across 30 languages and 17 years (2007–2024), with annotated epistemic framing (factual vs. opinionated) and verified temporal alignment relative to each model's training cutoff. Based on this data we present preliminary results on how linguistic resources available for a given language, the epistemic framing of a claim, and the temporal relationship between a claim and a model's training data impact the fact-checking performance of LLMs. FactSpan is designed to be dynamically extensible: the collection and annotation pipeline is fully reproducible, allowing evaluation to keep pace as models evolve.

Using FactSpan, we structure our analysis around three research questions:

- **RQ1: Multilingual reliability.** How does LLM fact-checking performance vary across languages of different resource levels, and what failure modes emerge in low-resource settings?
- **RQ2: Epistemic framing.** Does the epis-

temic framing of a claim (factual assertion vs. opinion) systematically affect verifiability, and do models track stylistic plausibility in place of truth?

- **RQ3: Generalisation beyond training data.** How well do LLMs assess claims that postdate their training cutoff, and what does this reveal about the verification mechanism they employ?

Our contributions are as follows. We release FactSpan, an extensible multilingual fact-checking benchmark with claim-level annotations for epistemic framing and temporal alignment (Section 3). We provide a systematic zero-shot evaluation of five LLMs reporting accuracy and refusal rates across all evaluation dimensions (Section 5). Our preliminary results highlight consistent structural vulnerabilities: a fourfold error rate increase for low-resource languages, a paradoxical accuracy gap that makes factually framed claims harder to verify than opinionated ones, and evidence that models rely on surface-level linguistic heuristics rather than grounded knowledge retrieval (Section 6).

2 Related Work

Fact-checking benchmarks. Automated fact-checking is typically decomposed into claim detection, evidence retrieval, and claim verification (Vlachos and Riedel, 2014; Guo et al., 2022). Foundational benchmarks such as FEVER (Thorne et al., 2018) used Wikipedia-derived claims but are vulnerable to annotation artefacts (Schuster et al., 2019); real-world corpora such as MultiFC (Augenstein et al., 2019) and LIAR (Wang, 2017) remain predominantly English and domain-specific. Multilingual coverage is scarce: X-Fact (Gupta and Srikumar, 2021) covers 25 languages up to 2020; M4FC (Geng et al., 2025) and the SemEval-2025 MultiClaim benchmark (Pikuliak et al., 2025) expand multilingual and multimodal scope; and Quelle et al. (2025) show that claims frequently mutate across linguistic borders. However, systematic analysis of LLM reliability as a function of *language resource level* is absent from all of these, a gap we close using the taxonomy of Joshi et al. (2020). A parallel line of work examines cross-lingual factual inconsistency directly. Qi et al. (2023) show that multilingual models return inconsistent factual answers to semantically equivalent prompts across languages, and that scaling improves per-language accuracy without improving cross-lingual consistency. Wang et al. (2025) trace this to the final layers, where a language-independent representation is mapped back to a target language and failures in that transition yield incorrect answers even when the model answers correctly in another language. This offers a plausible account of the resource gradient we observe at the task level.

LLMs for fact-checking and epistemic framing. Zero-shot LLM fact-checking has shown promising accuracy in narrow settings (Hoes et al., 2023; Huang et al., 2024), but generalisation across topics and languages is poorly understood. Quelle and Bovet (2024) find that LLMs struggle most near the true/false boundary; Tai et al. (2025) find that model ratings are often accurate while the associated rationales are flawed, suggesting heuristic rather than evidential reasoning. Qazi et al. (2026) evaluate nine LLMs on 5,000 claims across 47 languages and find that model confidence and accuracy can diverge sharply, with the largest gaps for non-English and Global South claims. Wang et al. (2026) quantify cross-language inequality in LLM fact-checking across nine languages, finding stronger performance for richer-resource languages. Both operate on samples of a few thousand claims. We extend this direction with a corpus an order of magnitude larger and link the observed disparity to resource class, refusal behaviour, and claim framing. A largely unexplored source of difficulty is the epistemic framing of the claim itself. Lasser et al. (2023) show that political content frequently blends factual assertion with subjective framing, and Hafid et al. (2025) find that surface-level linguistic form degrades LLM performance on scientific claims relative to generic discourse. We extend this line of inquiry across a large multilingual benchmark, directly comparing model accuracy on fact-framed versus opinion-framed claims.

3 FactSpan: A Multilingual Fact-Checking Benchmark

FactSpan combines the X-Fact corpus (Gupta and Srikumar, 2021) with newly collected ClaimReview data, assembled in four sequential stages. First, we restricted ClaimReview[1] data to IFCN- or Duke Reporters' Lab-accredited organisations and

[1] https://schema.org/ClaimReview, the structured-data markup used by fact-checking organisations to publish machine-readable verdicts.

Table 1: Dataset Statistics

Characteristic	Value
Total Claim Count	61,514
Number of Languages	30
Languages	English, Portuguese, Spanish, German, Indonesian, Tamil, Hindi, Arabic, Turkish, Polish, Italian, Telugu, Dutch, Romanian, French, Persian, Serbian, Bengali, Georgian, Sinhala, Russian, Albanian, Norwegian, Kannada, Filipino, Hebrew, Chinese, Azerbaijani, Assamese, Khmer
Time Span	2007–2024
Veracity Distribution	False (80.0%), True (20.0%)
Claim Framing	Expressed as Opinion (60.1%), Stated as Facts (39.9%)

deduplicated on normalised claim text, yielding 43,938 post-2020 claims. Second, following Pelrine et al. (2023), we applied a two-pass verifiability filter to both sources: GPT-3.5 Turbo screened all claims and rejected those it deemed unverifiable; GPT-4 then re-examined all rejections and reinstated claims that were in fact verifiable, retaining 33,430 X-Fact claims and 34,065 from ClaimReview. Third, we merged the two cleaned sets and removed claims explicitly depending on non-textual media (video, image, audio) via a targeted GPT prompt, eliminating a further 5,204 claims. Fourth, we retained only languages with at least 100 claims, which dropped 777 claims and produced the final 30-language corpus of 61,514. Full prompt details are provided in the appendix.

3.1 Veracity Label Normalisation

Fact-checking organisations use heterogeneous verdict schemes. We normalised all labels into five categories following Gupta and Srikumar (2021), extended with manually curated entries for new labels in the ClaimReview data: *False*, *Mostly False*, *Partly False/Misleading*, *Mostly True*, *True*. For evaluation we further collapse to binary: the first three map to **False**; the latter two to **True**.

3.2 Epistemic Framing Annotation

The *epistemic framing* of a claim may impact how its veracity is evaluated (Lasser et al., 2023; Rathje et al., 2024). For example, *"More than one quarter of America's young adults are too fat to serve in the U.S. military"* is fact-framed, while *"It seems that more than a quarter of America's young adults are too fat to serve in the military."* is opinion-framed. Both statements assert the same verifiable proposition.

We use GPT-3.5 Turbo to identify if a claim is framed as a fact or as an opinion, independently of the verdict. A small human validation study with three annotators on a stratified sample of 20 claims yielded 90% agreement with the model (Cohen's $\kappa = 0.74$).

3.3 Dataset Statistics

FactSpan comprises 61,514 claims: 36,992 Opinion-Framed (60.1%) and 24,522 Fact-Framed (39.9%), spanning five topic areas (Politics & Governance 25.2%, Society & Culture 24.0%, Economy & Environment 21.0%, Health & Pandemics 16.7%, Conflict & Security 13.1%). Table 1 provides a summary. The language distribution is heavily skewed: English accounts for 37.7% of claims (23,215), the five largest languages for 64.3%, and the ten smallest for 5.0% combined, with per-language counts ranging from 118 (Khmer) to 23,215. Full per-language counts are included with the released dataset; our resource class assignment procedure is described in Appendix B.2.

FactSpan is engineered as a dynamically extensible resource. The collection and annotation pipeline is fully reproducible, and an update script ingests new ClaimReview data systematically, directly addressing the rapid obsolescence challenge identified by Tai et al. (2025) for descriptive LLM benchmarking. To keep results comparable as the corpus grows, releases are versioned and individually citable: the update pipeline publishes new Zenodo versions rather than modifying existing ones, so any reported result remains reproducible against the fixed release it was computed on.

4 Evaluation Framework

Task and metrics. We frame fact-checking as a binary decision under uncertainty, motivated by its role in platform moderation workflows where mod-

els must adjudicate claims using only parametric knowledge. Each model receives a claim and outputs one of three labels: *True*, *False*, or *No Verdict*. The three-way output separates errors (misclassification) from refusals (abstention), which aggregate accuracy conflates. We evaluate models on **accuracy** (computed only on claims with a definitive verdict) and **refusal rate** (proportion of *No Verdict* outputs).

Models. We evaluate five LLMs spanning closed and open weights and roughly an order of magnitude in parameter count. GPT-4o (`gpt-4o-2024-05-13`) and GPT-3.5 Turbo (`gpt-3.5-turbo-0125`) were accessed via the OpenAI API. Llama 3.1 8B (`meta-llama/Meta-Llama-3.1-8B`; Dubey et al., 2024) was deployed locally on NVIDIA GPUs. Llama 3.1 70B (`llama-3.1-70b-versatile`) and Mixtral 8x7B (`mixtral-8x7b-32768`; Jiang et al., 2024) were accessed via the Groq API. All models ran zero-shot at temperature 0 (or the minimum supported value) with a shared base prompt; minor syntactic adaptations were applied to Llama and Mixtral to enforce output format consistency. Claims were batched per model based on context window capacity. For the validation set described below, Llama 3.1 8B was additionally run via the Groq API (`llama-3.1-8b-instant`) as an instruction-tuned comparison to the local base checkpoint. Prompts and model version details are in Appendices A and B.

Analysis axes. Results are disaggregated along three contextual dimensions. **Language resource level**: each language is assigned a Joshi et al. (2020) class (0 = very low resource, 5 = high resource), enabling us to test whether performance tracks structural training-data inequalities. **Epistemic framing**: accuracy and refusal rates are compared between fact- and opinion-framed claims to determine whether harder performance on factual claims reflects selective caution or outright misclassification. **Temporal alignment**: claims are classified as pre- or post-cutoff relative to each model's official training cutoff (GPT-3.5 Turbo: Sep 2021; GPT-4o: Oct 2023; Llama 3.1: Dec 2023; Mixtral excluded as undisclosed). To validate that post-cutoff gains reflect genuine generalisation rather than timeless content, we additionally collected a separate set of 3,033 claims published between August 2024 and February 2025, postdating every model's training cutoff, and annotated each for date-relevance (*Date-Relevant* vs. not).

Table 2: Overall zero-shot performance on FactSpan. Accuracy is computed on claims with a definitive verdict; Refusal = % *No Verdict* outputs.

Model	Accuracy	Refusal
GPT-4o	73.3%	43.0%
GPT-3.5 Turbo	69.4%	16.6%
Llama 3.1 70B	54.6%	25.6%
Llama 3.1 8B	63.8%	10.4%
Mixtral 8x7B	53.4%	36.7%

Comparability across models. Three properties of this design limit direct comparison of absolute accuracy between models. First, the pre/post-cutoff partition is model-specific: because cutoffs range from September 2021 to December 2023, each model is evaluated on a different division of the same corpus. Second, accuracy is conditioned on answered claims, and refusal rates range from 10.4% to 43.0%, so the effective evaluation set differs across models in both size and composition. Third, refusal is informative behaviour rather than missing data, which is why we report it alongside accuracy throughout rather than folding it into a single score. We therefore treat cross-model accuracy differences as indicative. The claims we draw are within-model: the resource-class gradient, the fact/opinion gap, and the pre/post shift are each computed inside a single model's own answered set, and each holds in the same direction for every model tested.

5 Results

We evaluated five LLMs on 61,514 multilingual claims across 30 languages. Table 2 reports overall performance. Figures 1 and 2 present per-language and per-resource-class breakdowns; Figure 3 shows the fact vs. opinion split.

5.1 RQ1: Multilingual Reliability

Overall reliability varied widely. GPT-4o reached the highest accuracy (73.3%) but declined to classify 43.0% of claims. Among open-weights models, Llama 3.1 8B achieved the best balance (63.8% accuracy, 10.4% refusal rate), whereas Mixtral 8x7B performed near chance-level (53.4%), a spread of over 10 percentage points among open-weights models alone.

These aggregate metrics mask a systematic disparity linked to language resources (Figure 2). High-resource languages consistently yielded the

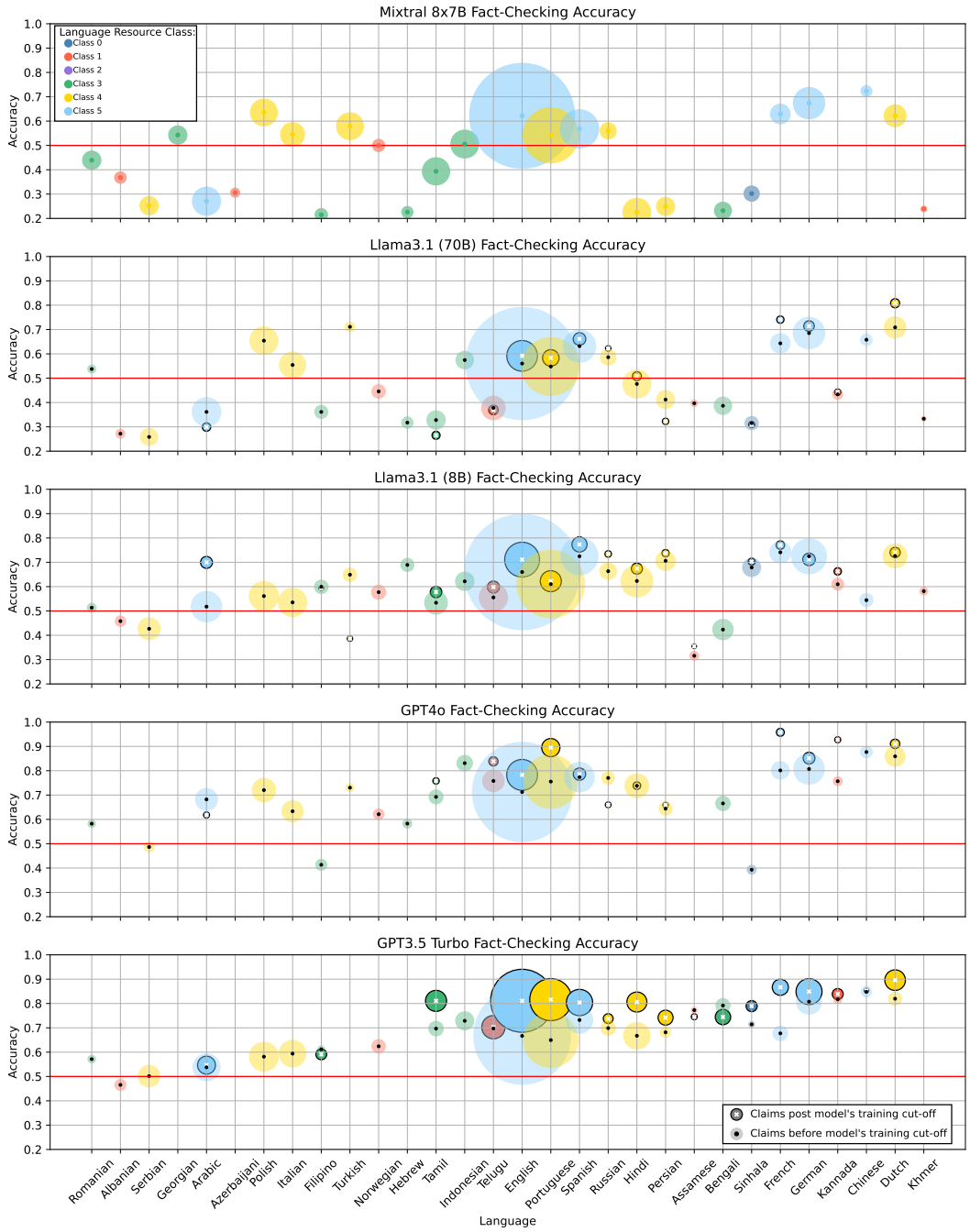

Figure 1: Fact-checking accuracy of five LLMs across 61,514 claims in 30 languages. Each subplot shows one model; x-axis: language; y-axis: accuracy on answered claims. Marker shape encodes temporal alignment (filled = pre-cutoff, open = post-cutoff); marker size encodes claim volume; colour encodes language resource class (Joshi et al., 2020) (Class 5 = high resource, Class 0 = low resource). The red dashed line at 0.5 indicates chance-level performance. Mixtral's training cutoff is undisclosed; pre/post distinctions are not shown for that model.

best performance: GPT-3.5 Turbo reached error rates as low as 12.5% for Dutch, and German showed similarly strong results across all five models, with error rates consistently below 15%. In contrast, lower-resource languages exhibited a dual failure mode: models either refused to engage (GPT-4o abstained from ≈80% of claims in Class 0 languages) or failed to verify accurately. Error rates spiked sharply for lower-resource languages, with Serbian (Class 1, 49.9%), Albanian (Class 2, 55.4%), and Arabic (Class 3, 46.0%) running approximately four times higher than their high-resource peers. Across all models, refusal rates are lower for high-resource languages than

for lower-resource ones.

These gaps reveal a structural inequality in automated safety. The sharp degradation in performance for lower-resource languages implies that LLMs provide uneven protection across linguistic communities. This exacerbates digital inequalities by either denying service or misclassifying content in underserved languages, effectively creating a two-tiered safety architecture on the Web.

5.2 RQ2: Epistemic Framing

Claim framing had a consistent and substantial effect on model accuracy across all five models (Figure 3). Fact-Framed claims were consistently harder to verify than Opinion-Framed ones: accuracy dropped 10–20 percentage points for every model. GPT-4o fell from 78.8% on Opinion-Framed claims to 64.8% on Fact-Framed ones; GPT-3.5 Turbo from 77.1% to 58.2%. This gap is not explained by differential refusal behaviour: refusal rates were consistent across framing types, indicating that models do not apply greater caution to factual claims; they simply misclassify them more often.

This pattern is consistent with models relying on surface-level linguistic heuristics: subjective framing carries identifiable rhetorical markers (hedging, emotive language, explicit attributions of belief) that models can parse independently of factual content, while objective assertions require grounding that models cannot supply from parametric knowledge alone (Tai et al., 2025).

5.3 RQ3: Temporal Generalisation

We partitioned claims by each model's training cutoff to compare performance on known (Pre-Cutoff) versus unseen (Post-Cutoff) events. Contrary to the expectation that models perform better on memorised content, accuracy was maintained or improved on post-cutoff claims for closed-source models: GPT-4o improved from 72.6% (pre-cutoff) to 80.5% (post-cutoff); GPT-3.5 Turbo from 66.1% to 79.8%.

To rule out the possibility that post-cutoff claims simply reference older events, we compiled a separate validation set of 3,033 claims published between August 2024 and February 2025, collected outside FactSpan using the same pipeline and annotated for date-relevance (details in Appendix C). Even restricted to *Date-Relevant* claims explicitly referencing post-cutoff events, GPT-4o achieved 83.5% accuracy and GPT-3.5 Turbo 74.6% (Table 3). Refusal rates rose for these claims (GPT-4o refusal: 51.6%), indicating that models recognise their knowledge limits, yet the claims they do classify, they classify well.

Table 3: Accuracy on a separate validation set of 3,033 claims (August 2024 to February 2025), collected outside FactSpan and split by date-relevance. *Date-Relevant*: claim explicitly references an event occurring after the model's training cutoff. The 70B validation run used `llama-3.3-70b-versatile`, as the 3.1 endpoint used in the main experiment was no longer available on Groq at the time of collection.

Model	Date-Relevant	Other
GPT-4o	83.5%	81.7%
GPT-3.5 Turbo	74.6%	75.1%
Llama 3.3 70B (Groq API)	62.5%	65.1%
Llama 3.1 8B (Local)	57.6%	56.1%
Llama 3.1 8B (Groq API)	61.1%	63.5%

6 Discussion

Our results shed light on both the promise and the current shortcomings of deploying LLMs in fact-checking workflows. Especially in low-resource language settings, we observe markedly high refusal rates alongside comparatively low accuracy. This means that for the most under-resourced languages in FactSpan, the best available model effectively abstains from classifying the majority of claims. This reveals a dangerous paradox: LLMs appear least reliable precisely where they are most needed, in the linguistic communities most vulnerable to unchecked misinformation and least served by existing fact-checking infrastructure. If deployed uncritically at scale, these systems risk creating a two-tiered information environment: robust protection for dominant languages, and effective denial of service for linguistically underserved communities, effectively embedding linguistic privilege into the infrastructure that governs online speech.

At the same time, elevated refusal rates should not be interpreted purely as failure. Rather, they suggest that models can often recognise when a claim falls outside their knowledge boundaries. In this sense, refusals may reflect a degree of epistemic awareness. However, this strength is sharply limited. When models do provide an answer, they frequently rely on shallow heuristics—such as lexical patterns, stylistic cues, or general plausibility—rather than engaging in genuine fact verification. Consistent with Tai et al. (2025), this behaviour indicates that model judgments are often

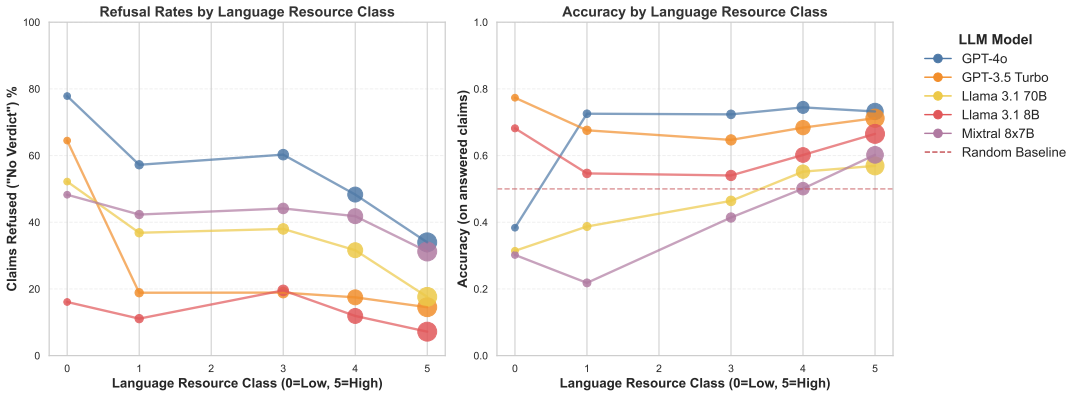

Figure 2: Performance across language resource classes (Class 0 = lowest resource, Class 5 = highest). *Left*: refusal rate (% No Verdict responses). *Right*: accuracy on answered claims. Marker size is proportional to claim volume. The red dashed line at 0.5 indicates chance-level performance.

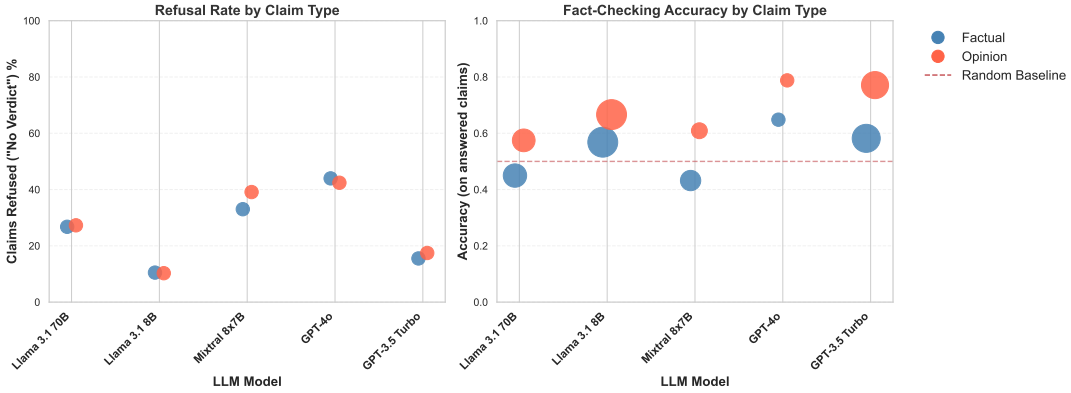

Figure 3: Performance on Fact-Framed vs. Opinion-Framed claims. *Left*: refusal rate (% No Verdict). *Right*: accuracy on answered claims. Points are offset horizontally to prevent overlap; circle area is proportional to claim volume. The red dashed line at 0.5 indicates chance-level performance.

driven by surface-level linguistic signals rather than by retrieval of evidence or grounded reasoning processes.

Further evidence for this limitation comes from our analysis of epistemic framing. Across all five models, we observe a consistent 10–20 percentage point accuracy gap between fact-framed and opinion-framed claims. This gap is not explained by increased caution: models do not refuse more often on factual claims, but instead misclassify them more frequently. This finding identifies epistemic framing as a systematic and previously underexplored source of difficulty in fact-checking evaluation. Importantly, the effect is amplified in low-resource languages, where limited training data and weaker evaluation benchmarks further constrain model performance.

Our temporal analysis reinforces this interpretation. Comparisons between claims published before and after model cutoff dates (Figure 1), as well as between *Date-Relevant* and other claims (Table 3), show little evidence that models rely on up-to-date or memorised factual knowledge. Instead, performance remains relatively stable across these conditions, suggesting that model predictions are largely insensitive to temporal novelty. Notably, refusal rates rise for recent claims while accuracy on answered claims remains stable — pointing to the same surface-level pattern observed in the framing analysis: models appear to engage selectively, but when they do engage, they rely on stylistic cues rather than grounded factual knowledge. This pattern is consistent with the hypothesis that LLMs (or at least those LLMs that we tested) depend primarily on stylistic and linguistic cues rather than on temporally grounded factual information.

Taken together, these findings indicate that LLMs should not be treated as autonomous fact-checkers, particularly in multilingual and low-resource settings. A more appropriate role is as

components within hybrid systems—for example, supporting claim detection, uncertainty signalling, or prioritisation—while delegating final verification to retrieval-based methods or human experts. Future work should therefore prioritise tighter integration with external knowledge sources, improved multilingual training coverage, and the development of methods that move beyond plausibility judgments toward evidence-based verification.

Finally, our findings carry implications for the design of fact-checking benchmarks. They underscore the importance of multilingual evaluation and highlight the need to account for variation in epistemic framing and linguistic style. At the same time, our results caution against assuming that temporal recency alone increases benchmark difficulty. If model performance is largely driven by stylistic features that remain stable over time, then datasets composed of recent claims are not necessarily more challenging. Instead, benchmark construction should explicitly target dimensions that disrupt heuristic-based reasoning and require genuine factual grounding.

7 Conclusion

This work introduced FactSpan, a novel multilingual fact-checking benchmark dataset, and provided a systematic evaluation of LLMs for fact-checking across languages, epistemic framing conditions, and temporal settings, yielding three main findings. First, we show that performance degrades sharply in low-resource languages, where high refusal rates and low accuracy limit practical usefulness. Second, we identify epistemic framing as a key factor, with a consistent accuracy gap between fact-framed and opinion-framed claims, demonstrating that models are sensitive to how claims are expressed. Third, our temporal analyses suggest that model predictions are largely insensitive to knowledge recency, pointing to a reliance on surface-level cues rather than grounded factual verification. Together, these findings offer a clearer understanding of the limitations of current LLMs and provide guidance for developing more robust, multilingual, and evidence-based fact-checking systems. Concretely, we cannot assume that a model performing well on English political opinions will safely moderate factual claims in Serbian. The dataset, annotations, and evaluation pipeline are released as a living benchmark, designed to be re-evaluated as models evolve.[2]

Limitations

Framing annotation and validation. Our epistemic framing labels are model-produced, and GPT-3.5 Turbo appears in three roles: as part of the verifiability filtering pipeline, as the epistemic framing annotator, and as one of the five evaluated models. Although these tasks are distinct in nature and the annotation step is supported by human validation, some degree of circularity cannot be entirely ruled out. That validation itself rests on a relatively small sample of 20 claims, so the reported κ values should be interpreted as indicative rather than conclusive, and expanding this analysis is a natural next step. Relatedly, our binary fact/opinion operationalisation may be coarser than the construct warrants; a graded framing score would likely support more reliable annotation and is a natural refinement.

Cross-model comparability. The two Llama 3.1 configurations differ in instruction tuning: the 8B model was run from the base checkpoint (`Meta-Llama-3.1-8B`), while the 70B was served via Groq's `llama-3.1-70b-versatile` endpoint, which hosts the Llama 3.1 70B Instruct checkpoint. Comparisons between the two open-weights models therefore conflate parameter count with tuning regime.

Temporal alignment. Publication date serves only as an approximate proxy for event knowledge; while our separately collected date-relevance validation set helps address this, some imprecision remains. The pre- and post-cutoff partition is also model-specific, since cutoffs differ across models, which is one reason we treat cross-model accuracy differences as indicative rather than direct (Section 4).

Scope of evaluation. We evaluate models using parametric knowledge alone. This reflects a common deployment pattern in high-volume moderation pipelines, where per-claim retrieval is costly, but it is a lower bound rather than a ceiling on achievable performance. Retrieval-augmented and reasoning-oriented systems would plausibly narrow the gaps we report, and establishing whether they do, in particular whether retrieval in a high-resource language can compensate for

[2] https://zenodo.org/records/15084388 and https://github.com/lorraine-dev/FactSpan

a low-resource query, is a natural extension that FactSpan is designed to support.

Benchmark coverage. FactSpan's coverage of 30 languages reflects the current IFCN ecosystem, which itself is shaped by existing fact-checking infrastructures. As a result, some underrepresented languages are not included, meaning that real-world performance differences in the most low-resource settings may not yet be fully captured.

Ethics Statement

This work introduces a multilingual benchmark for evaluating LLM-based fact-checking systems. Our results reveal substantial disparities in reliability across languages and claim types, particularly for low-resource languages and factually framed claims. Such inconsistencies raise concerns that automated moderation systems may unevenly protect users across linguistic communities and produce confident misclassifications in high-stakes settings.

The **FactSpan** benchmark contains publicly available claims collected from multilingual fact-checking and online information sources. While the dataset does not include personally sensitive information, some claims may involve culturally or politically sensitive topics. We therefore release the dataset and evaluation pipeline for research purposes only and encourage responsible use.

Finally, we emphasise that benchmark performance should not be interpreted as evidence that current LLMs are sufficiently reliable for fully automated fact-checking or moderation. Instead, our work aims to support more transparent, context-sensitive, and uncertainty-aware evaluation of multilingual moderation systems.

Data Availability

All data used in this study are publicly available. The FactSpan dataset is archived on Zenodo at `https://zenodo.org/records/15084388` (DOI: 10.5281/zenodo.15084388). The codebase for updating and processing claims is available at `https://github.com/lorraine-dev/FactSpan`.

Acknowledgments

This work was partially supported by the German Research Foundation (DFG, project no. 504226141).

References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

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.

Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. 2025. M4FC: a Multimodal, Multilingual, Multicultural, Multitask Real-World Fact-Checking Dataset. *arXiv preprint arXiv:2510.23508*.

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.

Salim Hafid, Sebastian Schellhammer, Yavuz Selim Kartal, Thomas Papastergiou, Stefan Dietze, Sandra Bringay, and Konstantin Todorov. 2025. An in-depth analysis of the linguistic characteristics of science claims on the web and their impact on fact-checking. *ACM Transactions on the Web*, 19(3).

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Using ChatGPT to Fight Misinformation: ChatGPT Nails 72% of 12,000 Verified Claims. *PsyArXiv preprint*.

Sean S Huang, Qingyuan Song, Kimberly J Beiting, Maria C Duggan, Kristin Hines, Harvey Murff, Vania Leung, James Powers, TS Harvey, Bradley Malin, and Zhijun Yin. 2024. Fact Check: assessing the response of ChatGPT to Alzheimer's disease myths. *Journal of the American Medical Directors Association*, 25(10):105178.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.

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.

Jana Lasser, Segun T. Aroyehun, Fabio Carrella, Almog Simchon, David Garcia, and Stephan Lewandowsky. 2023. From alternative conceptions of honesty to alternative facts in communications by US politicians. *Nature Human Behaviour*, 7(12):2140–2151.

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. 2025. SemEval-2025 Task 7: Multilingual and Crosslingual Fact-Checked Claim Retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Washington, D.C. Association for Computational Linguistics.

Ihsan Ayyub Qazi, Zohaib Khan, Abdullah Ghani, Agha Ali Raza, Zafar Ayyub Qazi, Wassay Sajjad, Ayesha Ali, Asher Javaid, Muhammad Abdullah Sohail, and Abdul Hameed Azeemi. 2026. Large language models show Dunning-Kruger-like effects in multilingual fact-checking. *Scientific Reports*, 16(1):7594.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697.

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

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. 2024. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Yuehong Cassandra Tai, Khushi Navin Patni, Nicholas Hemauer, Bruce Desmarais, and Yu-Ru Lin. 2025. GenAI vs. Human Fact-Checkers: Accurate Ratings, Flawed Rationales. In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, page 516–521, New York, NY, USA. Association for Computing Machinery.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Dandan Wang, Stephanie Jean Tsang, and Yadong Zhou. 2026. Performance unfairness of large language models in cross-language fact-checking. *Information Processing & Management*, 63(4):104616.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

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.

A Prompts used in the paper

To ensure the claims in our dataset were suitable for text-based fact-checking, we implemented a verifiability check, as described by Pelrine et al. (2023). This process aimed to identify claims that

contained sufficient textual information for independent assessment, excluding those that were truncated or required non-textual context (e.g., images, videos).

The following prompt was used to assess the verifiability of each claim:

> Determine if each claim is fact-checkable. Label each as 'Verifiable' or 'Unverifiable'. A claim is unverifiable if it is truncated or requires non-textual context (like an image or video). Use your pre-training knowledge to understand non-English claims based on linguistic and cultural context. Here are examples to guide your judgment:
>
> Input: 1. "A recent study found that eating two apples a day can drastically reduce your risk of heart disease." - This claim mentions a study with specific outcomes, making it verifiable.
>
> 2. "Watch this video to see the shocking truth about climate change." - This claim requires viewing a video, which is non-textual, making it unverifiable.
>
> Output: '1. Verifiable', '2. Unverifiable'

To keep costs manageable, we used a two-pass approach to assess whether each claim contained sufficient textual information to be fact-checked independently.

First pass (GPT-3.5 Turbo). We ran all claims through GPT-3.5 Turbo, which labelled each as *Verifiable* or *Unverifiable*. This model was effective but conservative: it correctly identified most unverifiable claims, yet also rejected a significant number of claims that were genuinely verifiable.

Second pass (GPT-4). To recover those incorrectly rejected claims, we ran everything GPT-3.5 Turbo had labelled *Unverifiable* through GPT-4 for a second opinion. GPT-4 reinstated many of these as *Verifiable*, producing a cleaner final set.

Why two passes? Running GPT-4 over the full dataset would have been prohibitively expensive. By using GPT-3.5 Turbo as a first filter and applying GPT-4 only to the rejected subset, we kept costs low without losing genuine claims in the process.

For explicitly identifying claims that had associated media, we used the following prompt:

> For each of the following claims, determine if the claim explicitly depends on a specific piece of media (e.g., video, image, audio clip) to convey its full meaning. A claim should be considered 'media-associated' if it directly references or describes media content, such as "in the video," "as shown in the image," or "according to the audio clip." The claim should be marked as 'yes' only if understanding the claim relies on or is significantly enhanced by this media. If the claim makes sense on its own and does not require any media to be understood, respond with 'no.'

For annotating the style of the claim, i.e., whether it was presented as a fact or expressed as an opinion, the following prompt was used.

> Determine the nature of each claim below, classifying it as either 'Factual' or 'Opinion', based on how it is framed in the text. Use these definitions: Factual: A claim asserted as a fact. Opinion: A claim that expresses personal belief or subjective interpretation. Please respond with the classification in English, even if the claim is in a non-English language. Analyze the content and context of each claim to determine its type.

As detailed in Section 3, we validated this classification using a blind inter-annotator agreement study with three expert annotators. While human agreement reflected the inherent subjectivity of the task (Fleiss' Kappa = 0.38), the model showed strong alignment with the per-claim majority label across the three annotators (Cohen's Kappa = 0.74), confirming its reliability as a proxy for the majority human view. For example, factual claims such as *"More than one-quarter of America's young adults are too fat to serve in the U.S. military"* and *"Alfamart Menyum-bangkan 6000 Kupon Untuk Membantu Melawan COVID-19"* (Translation: "Alfamart Announces Develop 6000 Coupons For Help Fighting COVID-19") were correctly labelled as factual. In contrast, opinion claims like *"Bernie Sanders' health care plan would 'empower Republican governors to take away Medicaid, to take away health insurance for low-income and middle-income working Americans.'"* and *"Se ele [João Doria] tivesse cumprido [a promessa de não deixar a prefeitura] não estaria com 60% de rejeição na cidade de São Paulo"* (Translation: "If he [João Doria] had kept his promise not to leave the mayor's

office, he wouldn't have 60% disapproval in São Paulo") were appropriately classified as opinions.

These examples highlight the model's ability to generalise the classification task across both linguistic and cultural contexts, providing a reliable foundation for downstream analysis.

A.1 Fact-Checking Prompt (All Models)

We used a standardised base prompt to enable consistent fact-checking across models. This prompt was minimally adapted for certain architectures to accommodate interface-specific formatting or output behaviours. The objective was to elicit deterministic, label-only responses using the labels `true`, `false`, or `no verdict`.

Base Prompt (Reference)

```
Fact check the following statements. Use labels 'true' or 'false'. If a verdict cannot be reached, use 'no verdict'. The output should only contain the numbered labels corresponding to each statement.
```

Model-specific Prompt Variants

- **GPT-3.5 Turbo and GPT-4 (OpenAI API)**
  System Prompt: Default
  User Prompt: Used the base prompt without any changes.

- **Llama 3.1 8B (local server) and Llama 3.1 70B (Groq API)**
  System Prompt:

  ```
  You are a multilingual language model with pretraining knowledge about the world.
  ```

  User Prompt:

  ```
  Fact check the following statements. Use labels 'true' or 'false'. If a verdict cannot be reached, use 'no verdict'. Output should only contain the assigned labels.
  ```

- **Mixtral 8x7B (Groq API)**
  System Prompt:

  ```
  You are a multilingual language model with pretraining knowledge about the world.
  ```

  User Prompt:

  ```
  Fact check the following statements. For each statement, provide the assigned label in the format '1. True', '2. False', etc. Use 'true' or 'false' as labels. If a verdict cannot be reached, use 'no verdict'. The output should only contain the numbered labels in English corresponding to each statement.
  ```

Prompt adaptations were developed through iterative manual refinement on a small set of test claims. Beginning with a simple base prompt, we introduced minor modifications until each model produced consistent and reliable outputs aligned with the expected labelling format.

B Methodological Details

B.1 Model Versions and Deployment

We evaluated five LLMs (plus one additional deployment used only for the validation set) using a mix of API access and local deployment to simulate diverse resource scenarios. The specific versions and access methods were:

- **GPT-4o:** `gpt-4o-2024-05-13` (via OpenAI API).
- **GPT-3.5 Turbo:** `gpt-3.5-turbo-0125` (via OpenAI API).
- **Llama 3.1 8B:** `meta-llama/Meta-Llama-3.1-8B` (downloaded from HuggingFace, deployed locally on NVIDIA GPUs).
- **Llama 3.1 70B:** `llama-3.1-70b-versatile` (via Groq API). Groq's `-versatile` endpoint serves Meta's instruction-tuned Llama 3.1 70B Instruct checkpoint.
- **Mixtral 8x7B:** `mixtral-8x7b-32768` (via Groq API).
- **Llama 3.1 8B (Groq):** `llama-3.1-8b-instant` (via Groq API). Used only for the date-relevance validation set, as an instruction-tuned comparison to the local base checkpoint.

B.2 Language Resource Class Assignment

To characterise resource availability per language, we use the taxonomy of Joshi et al. (2020), which sorts languages into six classes (0–5) by digital resource richness. We mapped languages using the `lang2tax.txt` file from the accompanying repository.[3] Most languages mapped directly. Three required manual resolution: Norwegian is listed separately as Bokmål and Nynorsk, both in Class 1, so we assigned Class 1; Filipino was assigned the class of Tagalog; and Chinese was assigned the class of Mandarin.

[3] https://microsoft.github.io/linguisticdiversity/

B.3 Computational Performance

Processing throughput varied significantly, illustrating the trade-offs between local control and API scalability. The locally deployed Llama 3.1 8B was the slowest, processing approximately 5 claims every 12 seconds. In contrast, GPT-4o was the fastest, processing batches of approximately 30 claims every 2 seconds. All experiments used a temperature setting near zero to ensure reproducibility.

C Temporal Alignment and Validation

C.1 Cutoff Dates and Classification

To analyse the impact of parametric knowledge, we classified each claim as *Pre-Cutoff* or *Post-Cutoff* based on the model's official knowledge cutoff date at the time of experimentation:

- **GPT-3.5 Turbo**: September 2021
- **GPT-4o**: October 2023
- **Llama 3.1**: December 2023.
- **Mixtral 8x7B**: As the training cutoff for Mixtral is not publicly disclosed, this model was excluded from the temporal (Pre vs. Post-Cutoff) analysis.

C.2 Validation of "Date-Relevance"

A potential confounder in temporal analysis is that a claim published post-cutoff might refer to "timeless" knowledge (e.g., geographic facts) rather than novel events. To verify that our observations reflect genuine generalisation to new information, we compiled a dedicated validation set of 3,033 claims spanning August 2024 to February 2025. This additional dataset was collected using the same methodology described in Section 3, but specifically includes the period following the most recent model cutoff (Llama 3.1, Dec 2023) to ensure no overlap with training data.

We annotated these claims using GPT-4o to distinguish **Date-Relevant** claims (explicitly referencing events post-cutoff, such as recent elections) from **Timeless/General** claims. The prompt used was:

> *"For each of the following claims, determine if the claim is about a specific event that took place after [Model Cutoff Date]. A claim should be considered 'yes' if it explicitly refers to an event that occurred after this date, making it impossible to fact-check without post-[Date] knowledge..."*

Results. Even when restricted to strictly **Date-Relevant** claims (where the model theoretically lacks knowledge), accuracy remained high. For instance, GPT-4o achieved **83.5%** accuracy on claims about events occurring after its training cutoff. This confirms our hypothesis in Section 5.3: models are not relying on memorised ground truth, but are tracking linguistic style in a way that happens to correlate with veracity.