Poly-FEVER: A Multilingual Hallucination Detection Benchmark

Anonymous ACL submission

Abstract

We present *Poly-FEVER*, an open-source large-scale multilingual benchmark for hallucination detection through controlled factuality judgment in large language models. *Poly-FEVER* extends FEVER, Climate-FEVER, and SciFact into 77,973 Supported/Refuted claim instances aligned across 11 languages: English, Chinese, Hindi, Arabic, Bengali, Japanese, Korean, Tamil, Thai, Georgian, and Amharic. We augment each claim with topic metadata from a 22-topic LDA model, enabling topic-aware analysis of multilingual factuality behavior. Translations are produced with Google Cloud Translation and validated through manual inspection and GEMBA, with average scores around 90. Using *Poly-FEVER*, we benchmark ChatGPT-3.5, LLaMA-2 and LLaMA-3 series, Phi-3.5, and Mistral under general, language-wise, and classification prompt families, and evaluate self-detection through controlled claim rephrasing. We further analyze resource imbalance using web presence as a proxy for online claim availability, and evaluate retrieval-augmented factuality judgment with DPR over Wikipedia. Results show pronounced cross-lingual disparities under aligned claims, with stronger accuracy in high-resource languages such as English and Chinese and weaker accuracy in lower-resource languages such as Amharic and Tamil. Topic-aware prompting and retrieval provide selective gains, especially in Arabic and Amharic, while their effects vary across languages and can conflict with model priors. *Poly-FEVER* establishes a rigorous resource for responsible, language-adaptive hallucination detection and language-sensitive factuality evaluation.

1 Introduction

Large language models (LLMs) are increasingly deployed in high-stakes domains such as education, healthcare, and law, where factual reliability is critical (Huang et al., 2025). Yet, the corpora used to train mainstream LLMs remain demographically and linguistically imbalanced (Shah et al., 2019; Li et al., 2023). Language is a major axis of demographic and cultural diversity, yet hallucination detection remains disproportionately studied in a small set of high-resource languages, limiting assessment fairness across linguistic communities.

Prior hallucination studies have focused primarily on high-resource languages such as English (Yao et al., 2023), Chinese (Cheng et al., 2023), and German (Sennrich et al., 2023). These studies have advanced hallucination mitigation through prompt and system design, including retrieval-augmented prompting (Niu et al., 2024), feedback loops, and prompt tuning, as well as modeling approaches such as architectural changes, loss design, and supervised fine-tuning (Tonmoy et al., 2024). However, the field lacks a unified multilingual evaluation setting in which the same factual propositions are tested across languages. Without this control, language effects are confounded with topic choice, source distribution, and claim selection, making it difficult to draw systematic conclusions about cross-lingual hallucination behavior.

Figure 1 illustrates this mismatch by contrasting (A) the global distribution of native speakers (Dunn, 2020; Paolillo and Das, 2006), (B) the concentration of hallucination research in a small set of high-resource languages (Joshi et al., 2020; Yu et al., 2022; Ranathunga and De Silva, 2022), and (C) the language coverage of *Poly-FEVER* by resource group. This divergence motivates a controlled multilingual testbed that can compare model behavior while holding factual content constant.

A central conceptual distinction in this work is between fact verification and hallucination detection (Bang et al., 2025). Fact verification determines whether a claim is *supported or refuted* by established knowledge sources and thus often assumes access to external evidence (Murayama, 2021; Zhu et al., 2021). Hallucination detection, by contrast, evaluates whether model outputs contain false or unsupported content. Multilingual eval-

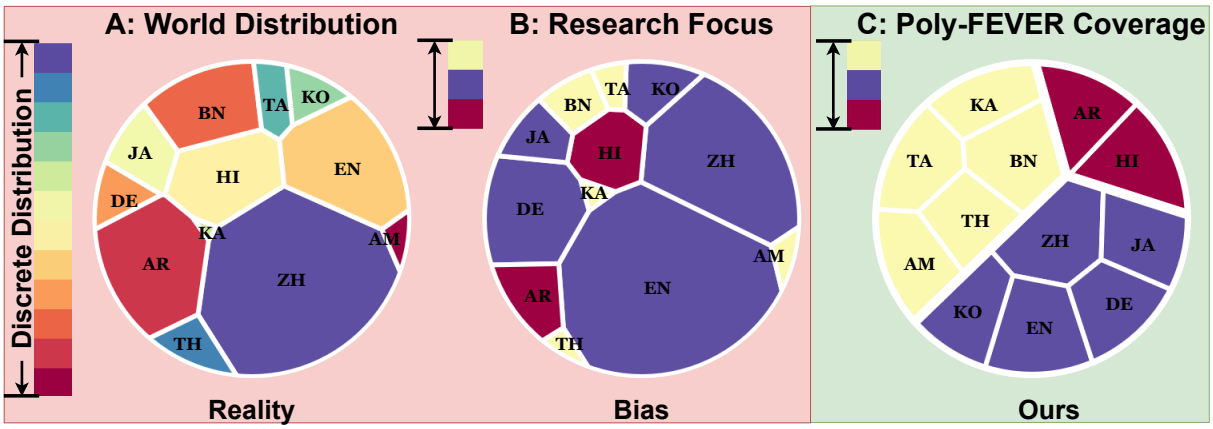

Figure 1: **(A) World distribution:** each polygon corresponds to a language, with area proportional to its share of native speakers, colored uniquely by language. **(B) Research focus:** areas represent how much existing hallucination research currently concentrates on each language, with colors indicating *High-, Medium-, or Low-resource* categories. **(C) *Poly-FEVER* coverage:** a hierarchical treemap where outer regions denote resource groups and inner polygons represent languages, with area proportional to their share in the dataset and colors mapped to resource groups.

uation further complicates this domain for models operating under heterogeneous data distributions, resource availability, and linguistic structures. Therefore, ensuring cross-language consistency demands preserving *logical equivalence* across scripts and morphologies so that labels remain faithful to ground truth. This framing positions hallucination detection as a test of a model's internal world knowledge rather than its evidence-retrieval ability.

We address these gaps by introducing *Poly-FEVER*, a multilingual benchmark for hallucination detection through controlled factuality judgment. *Poly-FEVER* aligns the same factual claims across 11 languages, allowing language to vary while factual content and ground-truth labels remain fixed. This design enables structured comparisons of model behavior across languages, prompts, topics, self-detection settings, web-presence conditions, and retrieval-augmented evidence.

Our contributions are threefold:

1. We introduce *Poly-FEVER*, a controlled multilingual hallucination-detection benchmark with 77,973 Supported/Refuted source claims, each aligned across 11 languages.
2. We provide evaluation suite covering contemporary LLMs, general, language-wise, and classification prompt families, and self-detection through controlled claim rephrasing.
3. We present diagnostic analyses of cross-lingual factuality behavior using topic metadata, web-presence estimates, and retrieval-augmented evidence, exposing failure modes that are difficult to observe in English-only evaluation.

2 Related Work

LLMs may generate hallucinations that contradict established facts or introduce unsupported information, posing challenges for reliable AI applications. Huang et al. (2023) categorize hallucinations into two types: intrinsic and extrinsic. Intrinsic hallucinations occur when a model produces **self-contradictions or logical inconsistencies within its own response**, while extrinsic hallucinations involve **factually incorrect statements that do not align with established knowledge sources**. Evaluating these errors requires either comparison with verified ground truth or analysis of consistency within model outputs. Reliable hallucination detection is therefore important for deploying generative AI systems in high-stakes settings such as healthcare, law, and scientific research.

Existing hallucination-evaluation methods assess factuality through entity correctness, factual consistency, and contradiction detection. Early methods relied on n-gram overlap, such as ROUGE and PARENT-T (Lin, 2004; Wang et al., 2020), but lexical overlap does not reliably capture factual correctness. Entity-based hallucination metrics (Nan et al., 2021) and relation-based fact extraction models (Goodrich et al., 2019) provide more structured alternatives. However, many of these methods were developed and evaluated primarily on English outputs, limiting their ability to support controlled multilingual hallucination analysis.

Fact-checking datasets provide another foundation for factuality evaluation. FEVER (Thorne et al., 2018), Climate-FEVER (Diggelmann et al., 2020), and SciFact (Wadden et al., 2020) pair

claims with evidence-based veracity labels, making them useful for evaluating whether models can distinguish supported from refuted claims. However, these resources were originally designed for fact verification, where evidence access is central, rather than for controlled multilingual hallucination-detection evaluation.

Language coverage remains a major limitation. Many fact-checking resources are English-only, which restricts their ability to evaluate factuality judgment across linguistic contexts. Existing multilingual fact-checking efforts (Gupta and Srikumar, 2021a) broaden language coverage, but they often aggregate language-specific claims rather than aligning the same factual proposition across languages. Without aligned claims, language effects are confounded with topic, source, and claim-selection effects, making cross-lingual hallucination patterns difficult to assess.

A separate line of work studies hallucinations in machine translation, where models may introduce content not supported by the source sentence (Pfeiffer et al., 2023). These benchmarks are important for evaluating translation faithfulness, but they do not directly test whether an LLM can judge the factuality of the same claim across languages.

3 *Poly-FEVER* Benchmark

This section details the construction, structure, and application of the *Poly-FEVER* benchmark. We describe the dataset curation process, present its organizational structure, and analyze the distribution of topics within the data.

3.1 *Poly-FEVER* Overview

Poly-FEVER provides a controlled multilingual benchmark for hallucination detection through factuality judgment in LLMs. It contains 77,973 language-specific claim instances aligned across 11 languages. The benchmark extends three English fact-checking resources: FEVER (Fact Extraction and VERification) (Thorne et al., 2018), Climate-FEVER (Diggelmann et al., 2020), and SciFact (Wadden et al., 2020). These resources provide claims with evidence-based veracity labels, allowing us to construct a multilingual factuality benchmark while preserving the original Supported/Refuted distinction.

FEVER contains claims derived from Wikipedia and labels each claim as *Supported*, *Refuted*, or *NotEnoughInfo*. We retain *Supported* and *Refuted* claims, excluding *NotEnoughInfo* cases so that each retained example has a determinate veracity label. Climate-FEVER extends the FEVER framework to climate-related claims verified against scientific and expert-reviewed sources. We again retain only *Supported* and *Refuted* claims to keep the label space consistent across source datasets. SciFact contains scientific claims from biomedical literature and labels whether cited papers *Support* or *Refute* each claim. Its inclusion broadens *Poly-FEVER* beyond general-domain and climate claims to include scientific factuality judgments. Together, these sources provide claims across general, climate, and scientific domains. After label normalization, each retained claim is paired with a binary veracity label and aligned multilingual versions. This design allows factual content to remain fixed while language varies, enabling controlled analysis of cross-lingual factuality judgments.

- **ID:** A unique identifier assigned to each claim, allowing for consistent referencing and seamless integration with other annotations or future expansions.
- **Claim:** The textual content of the claim, is provided in 11 distinct languages. This multilingual format highlights the dataset's cross-lingual capabilities, enabling evaluations of how models handle claims across diverse linguistic and cultural contexts.
- **Label:** An annotated veracity label (*true* or *false*) indicating whether the claim aligns with established factual evidence. This ensures that each claim is directly tied to a definitive truth value, reflecting its correctness within the dataset.
- **Topic Distribution:** A vector of the top five predicted topics for the claim, derived through thematic analysis (e.g., LDA). This field offers contextual clues about the claim's domain (e.g., sports, history, science) and aids in cross-topic assessments of model performance.

3.2 *Poly-FEVER* Schema

To support reuse with prior work and ease of adoption, *Poly-FEVER* mirrors the FEVER family structure while adding aligned multilingual and topic-aware fields. Each entry in *Poly-FEVER* contains four primary fields as shown in Box 1 The shared ID and label fields are central to the benchmark design: the same factual proposition can be evaluated across languages while preserving a common ground-truth label.

3.3 Topic Distribution

We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to assign topic metadata to claims. Based on coherence scores, we select 22 topics and store the top topic assignments for each claim. These metadata support topic-aware analysis of fac-

tuality judgments across languages. Section 5.5 analyzes the resulting topic effects, and Appendix A.3 provides preprocessing and modeling details.

3.4 Language Selection

We select 11 languages that span high-resource to low-resource, alphabetic to logographic scripts, and diverse grammatical families, enabling systematic study of linguistic bias (). The dataset covers English (en), Mandarin Chinese (zh-CN), Hindi (hi), Arabic (ar), Bengali (bn), Japanese (ja), Korean (ko), Tamil (ta), Thai (th), Georgian (ka), and Amharic (am). The languages are ordered by the number of native speakers. The selection ensures *Poly-FEVER* evaluates LLM performance across linguistic groups with different structures, resources, and usage patterns.

Each language introduces distinct linguistic and resource-related conditions that may affect hallucination-detection behavior in LLMs. Mandarin Chinese, Arabic, and Japanese include *logographic or syllabic writing systems* that differ from English's alphabetic structure. Tamil, Thai, and Georgian introduce *distinct syntactic rules* that affect sentence formation. *Ambiguity, polysemy, and homophones* in Hindi, Korean, and Bengali require models to resolve multiple interpretations of words in context. Amharic and Tamil contain *complex morphological patterns* that affect tokenization and semantic understanding.

Dialects and sociolects create additional challenges in fact verification. Hindi and Arabic contain regional variations that influence word choice and sentence construction. Korean and Japanese incorporate honorifics and context-dependent expressions that require models to infer meaning from cultural and social cues. LLMs which are trained primarily on English and widely spoken languages struggle to adapt to these variations, leading to inconsistencies in hallucination detection.

The selection also considers differences in script and orthography (Rust et al., 2021; Pfeiffer et al., 2021). Non-Latin scripts appear in Chinese, Arabic, Thai, Hindi, Bengali, Tamil, Georgian, and Amharic. These scripts require different tokenization strategies, which affect how models encode and process text. The dataset ensures that evaluations reflect performance across writing systems with unique linguistic properties.

3.5 Multilingual Claim Translation

We construct the multilingual claim set by translating English source claims into 10 target languages. We considered Google Cloud Translation, DeepL, and LLM-based translators. Because translation quality is central to the validity of a multilingual benchmark, we evaluate translations through manual inspection and automatic quality scoring. Our manual audit covers 1% of the dataset, with two annotators independently reviewing each sampled claim. The audit focuses on whether translations preserve the factual meaning of the source claim. Google Translate produced more accurate translations than other methods. DeepL Translate introduced structural errors, such as altering verb tense and misordering words. One example included the mistranslation of "The book was read by him" as "Book by him read," which changed the grammatical structure. These inconsistencies led to the exclusion of DeepL and other APIs.

Concerns about **hallucinations in LLM-based translations** influenced the decision to use Google Translate. LLMs generate fluent text but may introduce fabricated details when translating factual claims. Google Cloud Translation provided greater consistency by following source structures without generating new information.

Lang.	**zh**	**hi**	**ar**	**bn**	**ja**	**ko**	**ta**	**th**	**ka**	**am**
Score	91.3	92.4	90.8	91.8	91.5	**93.0**	90.0	91.1	90.8	88.9

Table 1: Average GEMBA translation-quality scores on a 5% sample of *Poly-FEVER*. Higher scores indicate better meaning preservation.

The translation covered over 80,000 English claims into 10 target languages, with a total cost of $2,644 using Google Cloud Translation. To assess quality, we applied the GPT Estimation Metric Based Assessment (GEMBA) (Kocmi and Federmann, 2023b) to 5% of the data. GEMBA provides reference-free scores of meaning preservation, and results averaged around 90 across languages (Table 1). These scores indicate that *Poly-FEVER* maintains high semantic fidelity while supporting semantic fidelity, providing a reliable basis for cross-lingual evaluation. We provide further analysis of translation quality distributions and filtering trade-offs in Appendix A.4.

4 Hallucination Causes Exploration

We use "causes exploration" in a diagnostic sense: hallucination-detection behavior may vary with model properties, decoding settings, external resource availability, and prompt specification. To characterize these variations, we examine hallucination detection through three diagnostic perspec-

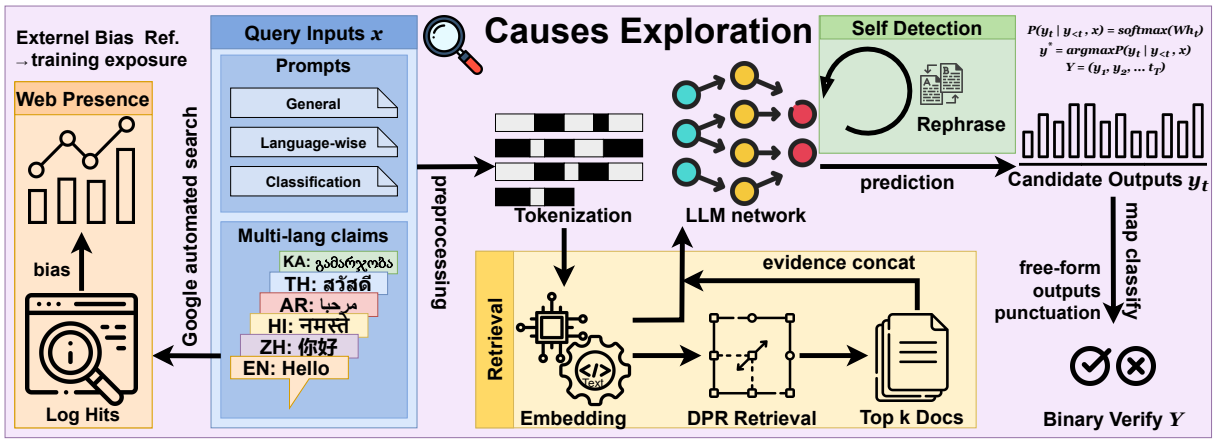

Figure 2: A conceptual overview of factors examined in our hallucination causes exploration, showing how multilingual claims interact with tokenization, embeddings, retrieval, prediction, self-detection, web presence, and prompt type. Multilingual claims and prompts (x) are preprocessed through normalization and tokenization, then passed into the LLM network. **Web presence** measures search-hit counts as an externally observable proxy for online claim availability. **Retrieval** encodes claims into dense embeddings, retrieves top-k Wikipedia passages, and concatenates the retrieved evidence with the claim. The LLM predicts candidate outputs (y_t) using token-level probabilities ($P(y_t \mid y_{<t}, x)$) under the decoding setting used in evaluation. A **self-detection loop** rephrases claims and reclassifies them to assess whether factuality judgments remain consistent under model-generated claim variants.

tives. First, we evaluate how models classify factual claims under different prompts and whether they recognize their own hallucinations. Second, we analyze web presence as a proxy for online claim availability and examine its association with multilingual factuality performance. Third, we assess how retrieval-augmented evidence changes factuality judgments by providing external context. Together, these analyses identify observable conditions under which multilingual hallucination-detection performance changes.

4.1 Model Performance

4.1.1 Effect of Prompt Design

Because our task requires binary factuality labels, prompt design affects both answer accuracy and output compliance. We use three prompt families: *General*, *Language-wise*, and *Classification*. The **General** prompt gives a minimal binary instruction; the **Language-wise** prompt is our default evaluation prompt and asks models to return only `true` or `false`; the **Classification** prompt maps free-form outputs to a binary label. We show the Language-wise prompt in the main text because it provides the default setting for our reported evaluations; full verbatim prompts are provided in Appendix B.1.

Language-wise Prompt for Multilingual Eval.

Assess the truthfulness of the user's claim. Respond with *true* if the claim is true and *false* if the claim is false. Your response should contain only *true* or *false*, without additional characters or punctuation.

4.1.2 LLM-Rephrased Claims as a Proxy for Intrinsic Hallucinations

We define self-detection as a controlled rephrase-and-classify setting. The model first generates a meaning-preserving rephrasing of a labeled claim and then judges the factuality of that generated claim under the original veracity label. This setting tests whether factuality judgments remain stable when the evaluated claim is model-generated rather than directly taken from the benchmark.

Rephrase Prompt for Synthetic Claim Generation

Rephrase the following claim without changing its meaning. Ensure the essence and intent remain unchanged.

Open-ended LLM generations lack predefined ground-truth labels, which makes systematic factuality evaluation difficult. To approximate intrinsic hallucinations under controlled conditions, we rephrased all 77,973 claims through LLM generation while anchoring to their original veracity labels. This yields model-generated claim variants that remain evaluable when the rephrasing preserves factual meaning. A manual audit of 100 English and Chinese rephrased claims checks semantic fidelity and supports the use of this controlled self-detection setting. This setting allows us to compare direct classification with self-detection under model-generated rephrasing.

4.2 Web Search on References Imbalance

We analyze web presence to estimate the availability of each claim on the open web across the 11

languages. Because LLM pretraining data are not directly observable, web presence serves only as an external proxy for online claim availability, not as a direct measure of training exposure.

For consistent data collection, we use standardized query collection with randomized user-agent strings and inter-query intervals to reduce rate-limit artifacts and avoid systematic collection differences across languages (Olteanu et al., 2019; Kulshrestha et al., 2017), so that all languages are sampled comparably, emulating heterogeneous internet access patterns to reduce sampling bias and evade anti-bot detection. We operationalize web presence using Google search hit counts. These counts provide a coarse estimate of online claim availability across languages and allow us to examine whether online availability is associated with factuality performance. Section 5.4 provides the detailed analysis.

4.3 External Knowledge Factuality Judgment

To test how external evidence changes multilingual factuality judgments, we apply Dense Passage Retrieval (DPR) (Karpukhin et al., 2020). DPR uses dense retrieval to identify passages that are semantically related to a query claim, allowing the factuality prompt to include external evidence. In our setting, each claim is paired with top-k retrieved Wikipedia passages before factuality classification. We fix the retrieval corpus to English Wikipedia across all languages to provide a controlled and directly comparable evidence source. Full retrieval pipeline details, including native-language retrieval results and corpus statistics, are provided in Appendix B.2. Section 5.5 analyzes how retrieval-augmented evidence interacts with topic-aware evaluation.

5 *Poly-FEVER* Case Studies

5.1 Evaluation Setup

We report *baseline case studies* under fixed prompts and decoding conditions to illustrate multilingual failure modes and sensitivities. All results are obtained without parameter tuning, and the full hardware and software settings are provided in Appendix C.1. We evaluate ChatGPT 3.5, LLaMA-2 (7B/13B/70B), LLaMA-3.1 8B, LLaMA-3.2 (1B, 3B), Phi-3.5-4B, and Mistral-7B-Instruct as reference baselines using exact-match accuracy over binary truth labels with language-wise prompts and low-temperature decoding unless otherwise stated.

5.2 Case Study A: Temperature Sensitivity

We include a temperature sweep for LLaMA-3.1 8B across $\{0.1, \ldots, 2.0\}$ as a sensitivity check under fixed prompts (Appendix C.2, Table 10). The sweep shows language-specific response patterns: English increases with higher temperatures, Chinese and Amharic decrease, and several languages, including Japanese and Bengali, remain comparatively stable.

5.3 Case Study B: Multilingual Baselines for Fact Verification

We evaluate the same aligned-claim factuality task on ChatGPT 3.5, the LLaMA-2 series, LLaMA-3.1, the LLaMA-3.2 series, Phi-3.5-4B, and Mistral-7B-Instruct. For ChatGPT 3.5, we also evaluate self-detection by asking the model to rephrase claims and then classify the factuality of the rephrased claims under the original labels. Appendix C.3 reports results separately for FEVER, Climate-FEVER, and SciFact. Table 2 shows that English generally achieves higher accuracy than most other languages across models.

Self-detection (Rephrase+Classify) shifts vary by language: for some languages, such as Hindi, it improves accuracy, whereas English remains strongest under direct classification. These deltas should be interpreted as diagnostic evidence of prompt and generation sensitivity rather than as tuned improvements. Appendix C.4 discusses task feasibility and contextualizes below-chance scores as known failure modes (Li et al., 2024).

5.4 Case Study C: Web Presence and Resource Imbalance

We estimate web presence per language using an automated search procedure (details in Section 3.3). This serves as a coarse proxy for resource availability that may relate to model behavior.

Figure 3 compares search-hit count distributions across the 11 languages. English shows a wider interquartile range, indicating greater variability in search counts, while languages such as Thai show narrower distributions. Amharic and Georgian have lower median search counts, suggesting lower online claim availability in our collected search results. Languages like Amharic and Georgian have lower median and mean search counts, indicating less available content or fewer search results for these languages. This disparity could lead to an imbalanced performance in multilingual fact-checking, with better results in languages that have more content available, like English, and worse results in languages with less content.

Figure 4 compares average web-search counts with hallucination-detection accuracy across lan-

Lang.	ChatGPT		LLaMA-2			LLaMA-3.x			Phi	Mistral
	GPT 3.5	Self-Detection	7B	13B	70B	3.1-8B	3.2-1B	3.2-3B	3.5-4B	7B Inst.
en	**65.89%**	**61.88%**	**63.35%**	**64.27%**	**64.56%**	48.43%	**52.76%**	**50.06%**	**63.17%**	**65.20%**
zh-CN	58.25%	53.94%	58.29%	59.88%	38.75%	54.51%	41.08%	43.77%	46.10%	52.11%
hi	52.90%	58.10%	48.59%	54.33%	45.68%	47.94%	43.90%	43.77%	47.62%	55.40%
ar	45.48%	58.80%	50.55%	54.97%	32.97%	49.33%	33.50%	41.59%	38.23%	45.80%
bn	55.65%	58.73%	49.05%	52.30%	33.87%	45.86%	44.54%	38.90%	48.01%	53.90%
ja	55.89%	58.95%	58.24%	59.57%	41.37%	**55.34%**	29.53%	40.31%	34.27%	40.12%
ko	57.29%	60.14%	56.67%	58.74%	46.06%	52.47%	44.93%	49.42%	48.91%	56.60%
ta	55.67%	59.98%	49.54%	50.86%	19.47%	53.18%	30.68%	40.56%	35.30%	41.29%
th	56.82%	52.04%	53.90%	53.46%	48.09%	50.37%	34.53%	40.18%	39.79%	45.55%
ka	57.37%	51.69%	47.39%	53.45%	46.15%	48.71%	34.53%	41.85%	39.28%	44.83%
am	47.53%	47.06%	43.42%	48.64%	26.34%	48.37%	21.82%	39.15%	26.70%	33.41%

Table 2: Comparison of hallucination detection accuracy on the fact-checking task across models: ChatGPT-3.5, ChatGPT-3.5 Self-Detection, LLaMA-2 (7B, 13B, 70B), LLaMA-3.1-8B, LLaMA-3.2 (1B, 3B), Phi-3.5-4B, and Mistral-7B-Instruct. Highest values are bolded and lowest values are underlined.

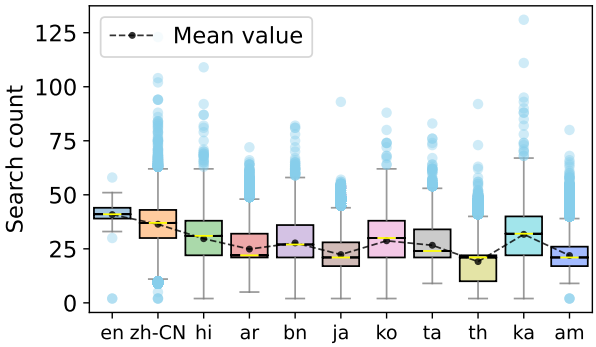

Figure 3: Web search distribution on multilingual claims. The middle 50% of search counts inside each box, mean values for each language are connected.

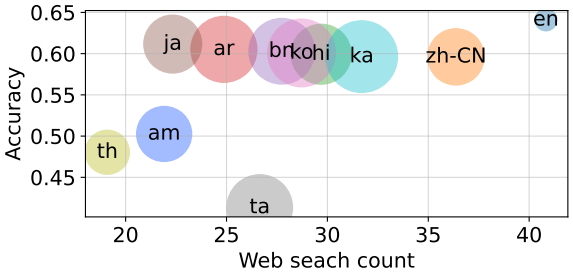

Figure 4: Detection accuracy on web search count. The x-axis reflects average search-hit counts, and the y-axis shows hallucination detection accuracy. Bubble sizes encode the variance-to-mean ratio of search counts.

guages. We observe a moderate association between the two ($\rho \approx 0.49$, $p = 0.13$). Languages with higher search-hit counts, such as English and Chinese, tend to show higher accuracy, whereas languages with lower search-hit counts, such as Amharic and Tamil, show lower overall accuracy. These patterns should be interpreted descriptively: web presence is one observable factor associated with accuracy differences, consistent with prior multilingual findings (Pires et al., 2019; Singh et al., 2023), but linguistic, model-specific, and prompt-compliance factors may also contribute.

5.5 Case Study D: Topic Sensitivity

We analyze topic-level sensitivity using the 22-topic LDA metadata described in Section 3.3.

LLMs generally perform best in English and struggle with Amharic and Thai as shown in Figure 5. Topics such as Politics, Sports, Film/Television, and Warfare History prove challenging across languages due to their subjective nature, where personal biases and interpretations can obscure the distinction between fact and opinion(Mahl et al., 2024; Hrckova et al., 2022). These results motivate releasing topic metadata with the benchmark so that future work can analyze whether multilingual factuality errors concentrate in particular topical regions.

The standard deviations show that topic-level performance is not uniform across languages. Several topics have high average accuracy but still exhibit cross-language variance, with English and Chinese often scoring higher than low-resource languages. This pattern indicates that topic effects and language effects should be evaluated jointly rather than treated as independent factors.

Table 3 reports topic-aware accuracy changes for non-English languages across selected topics. The effect is language-dependent: Thai and Amharic show consistent gains across many topics, while Tamil shows decreases in several topics. These results indicate that topic-aware prompting can change factuality judgments, but the direction and magnitude of the effect vary by language.

5.6 Case Study E: Topic-Aware and Retrieval-Augmented Evaluation

Beyond direct prompting, structured context can change multilingual factuality judgments. We therefore evaluate explicit topic awareness (LDA)

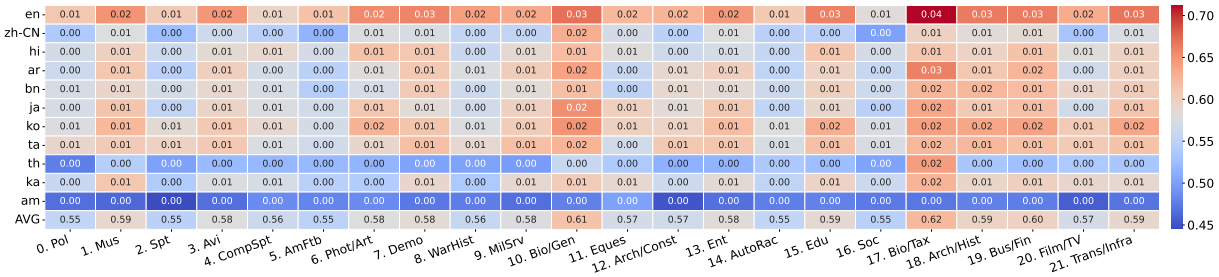

Figure 5: Average accuracy (represented by color) across topics by language, with standard deviation indicated by annotation. The color intensity reflects accuracy levels, where higher accuracy is represented by stronger shades, while lower accuracy is indicated by lighter tones. Standard deviation values illustrate the degree of variability within each topic, highlighting cases where the model exhibits inconsistent fact-checking performance across languages. The bottom row aggregates the mean accuracy and standard deviation for each topic across all languages, providing an overall assessment of topic difficulty and model robustness.

Lang.	Poli	Sport	Comp	Football	WarHist	Equestr	ArchConst	AutoRace	Soccer	FilmTV
zh-CN	2.62	2.5	1.17	4.72	2.78	3.02	1.02	1.28	2.66	2.59
hi	1.97	0.14	-1.33	2.80	4.66	1.87	0.88	0.62	-0.39	1.52
ar	1.57	0.42	0	4.99	0.91	1.47	0.73	1.99	0.29	1.64
bn	0.86	0.28	-0.81	3.31	2.16	3.75	-0.58	-0.66	0.09	-1.13
ja	-0.35	-1.81	-1.38	1.11	1.21	-2.85	0.22	0.43	0.35	0.21
ko	0.4	-2.36	-0.76	1.08	1.87	-0.98	0.22	0.66	-0.83	-0.32
ta	-1.41	-4.31	-0.48	1.05	-1.62	0.33	-0.95	-1.89	-1.98	-3.00
th	6.95	6.05	2.9	4.09	6.65	4.07	6.43	2.27	3.79	3.93
ka	2.28	2.57	-0.6	2.83	4.16	1.39	0	0.76	0.67	0.25
am	7.71	7.44	3.76	5.05	6.98	4.24	5.41	3.60	4.41	7.09

Table 3: Percentage improvement of correct judgment. Topic names are shown in English; claims remain in their original languages. Positive values indicate improved accuracy, while negative values indicate a decrease.

Lang.	Original	RevPrompt	LDA	LDA+RAG
en	**65.89%**	**64.09%**	**63.48%**	**60.5%**
zh-CN	58.25%	59.61%	57.81%	53.74%
hi	52.90%	59.93%	60.03%	54.49%
ar	45.48%	60.52%	59.71%	55.67%
bn	55.65%	60.29%	59.57%	57.35%
ja	55.89%	61.16%	58.44%	55.44%
ko	57.29%	60.08%	60.46%	57.42%
ta	55.67%	41.34%	58.83%	56.89%
th	56.82%	48.04%	57.39%	53.37%
ka	57.37%	59.64%	59.44%	56.99%
am	47.53%	50.26%	54.24%	53.49%

Table 4: Accuracy of hallucination detection with original process, prompts revised, LDA, LDA+RAG

and retrieval-augmented topic-aware prompting (LDA+RAG) for factual claim classification. This analysis examines how topic metadata and retrieved evidence affect accuracy across languages.

Table 4 summarizes hallucination-detection accuracy under different evaluation settings. Topic-aware prompting improves accuracy for several non-English languages, while retrieval-augmented evaluation has mixed effects, consistent with prior multilingual transfer findings (Chang et al., 2023; Quelle et al., 2023; Pires et al., 2019). Arabic and Amharic show the largest gains from LDA (+14.23 and +6.71 percentage points), while English and Chinese show slight decreases. These results suggest that explicit topic information can help in some language settings but is not uniformly beneficial.

Retrieval-augmented evaluation (LDA+RAG) also shows language-dependent effects. Arabic increases by 10.19 percentage points and Amharic by 5.96 percentage points relative to the original setting, while Chinese and Japanese decrease slightly. This mixed pattern shows that retrieved evidence can improve factuality judgment in some languages but may introduce noise or conflict with the model's prior predictions in others.

Some languages, including Tamil and Thai, show small or negative shifts across settings. These results reinforce the need to evaluate multilingual hallucination detection under multiple conditions rather than relying on a single prompt or retrieval configuration. These results reinforce the need to evaluate multilingual hallucination detection under multiple conditions rather than relying on a single prompt or retrieval configuration.

6 Conclusion

We introduce *Poly-FEVER*, a multilingual benchmark with 77,973 claims across 11 languages, designed to evaluate hallucination detection in LLMs. By benchmarking models like ChatGPT and LLaMA, we reveal significant performance disparities, where accuracy in high-resource languages like English correlates with greater web presence, while lower-resource languages like Amharic lag behind. Our analysis shows that topic structuring (LDA) and retrieval-augmentation (RAG) can improve performance for lower-resource languages, but these benefits are not universal and can conflict with the strong internal knowledge of models in high-resource contexts. *Poly-FEVER* enables researchers to disentangle model, prompt, and resource effects, providing a critical tool for developing more robust and equitable language-adaptive mitigation strategies.

7 Limitations

Poly-FEVER is a controlled multilingual benchmark for hallucination detection through factuality judgment, with factual content and veracity labels held fixed across languages. This design enables direct cross-language comparison, but it also scopes the benchmark to translated claims rather than native-authored claims, so it may not fully capture culturally grounded facts, local discourse conventions, or language-specific expressions of factual knowledge. *Poly-FEVER* also focuses on Supported/Refuted claims, which supports clear accuracy-based evaluation but leaves out uncertain, partially supported, and evidence-insufficient cases. Finally, our web-presence and retrieval analyses are diagnostic: Google search-hit counts are a coarse proxy for online claim availability, and DPR results depend on the retrieval corpus and retriever quality. These choices position *Poly-FEVER* as a controlled factuality-judgment resource for studying multilingual hallucination-detection behavior, while future work can extend it to native-authored claims, ambiguous evidence settings, cross-language hallucination transfer, and fact-grounded generation in low-resource languages.

8 Ethical Considerations

The construction of *Poly-FEVER* raises ethical considerations around fairness, inclusivity, data provenance, and manual validation. *Poly-FEVER* aims to reduce linguistic imbalance in hallucination-detection evaluation by covering 11 languages across different resource levels, scripts, and linguistic families. All source claims and veracity labels are derived from publicly available fact-checking datasets, and no private or personally identifiable information is introduced. Our manual work focuses on translation-quality validation rather than creating new factual labels: annotators review whether translated claims preserve the factual meaning of the English source claims. Because translation errors, cultural mismatches, and resource imbalance may still affect results, cross-lingual performance gaps should be interpreted as limitations of current LLMs and evaluation settings, not as properties of the languages or their speakers. We will release the dataset schema, prompts, preprocessing code, evaluation scripts, and translation-quality metadata to support transparent and reproducible research.

References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Yi-Chen Chang, Canasai Kruengkrai, and Junichi Yamagishi. 2023. Xfever: Exploring fact verification across languages. *arXiv preprint arXiv:2310.16278*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and 1 others. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

Ashim Gupta and Vivek Srikumar. 2021a. 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.

Ashim Gupta and Vivek Srikumar. 2021b. X-fact: A new benchmark dataset for multilingual fact checking. *arXiv preprint arXiv:2106.09248*.

Andrea Hrckova, Robert Moro, Ivan Srba, Jakub Simko, and Maria Bielikova. 2022. Autonomation, not automation: Activities and needs of fact-checkers as a basis for designing human-centered ai systems. *arXiv preprint arXiv:2211.12143*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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. *arXiv preprint arXiv:2004.09095*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kocmi and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4. *arXiv preprint arXiv:2310.13988*.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 417–432.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. *arXiv preprint arXiv:2405.19744*.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Daniela Mahl, Jing Zeng, Mike S Schäfer, Fernando Antonio Egert, and Thaiane Oliveira. 2024. "we follow the disinformation": conceptualizing and analyzing fact-checking cultures across countries. *The International Journal of Press/Politics*, page 19401612241270004.

Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. *arXiv preprint arXiv:2111.03299*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13.

John C Paolillo and Anupam Das. 2006. Evaluating language statistics: The ethnologue and beyond. *Contract report for UNESCO Institute for Statistics*.

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmt5: Modular multilingual pre-training solves source language hallucinations. *Preprint*, arXiv:2305.14224.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. Unks everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Dorian Quelle, Calvin Cheng, Alexandre Bovet, and Scott A Hale. 2023. Lost in translation–multilingual misinformation and its evolution. *arXiv preprint arXiv:2310.18089*.

Surangika Ranathunga and Nisansa De Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the nlp world. *arXiv preprint arXiv:2210.08523*.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2):2.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2023. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. *arXiv preprint arXiv:2309.07098*.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Pranaydeep Singh, Orphée De Clercq, and Els Lefever. 2023. Distilling monolingual models from large multilingual transformers. *Electronics*, 12(4):1022.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online. Association for Computational Linguistics.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Xinyan Velocity Yu, Akari Asai, Trina Chatterjee, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. *arXiv preprint arXiv:2211.15649*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Biru Zhu, Xingyao Zhang, Ming Gu, and Yangdong Deng. 2021. Knowledge enhanced fact checking and verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3132–3143.

A *Poly-FEVER* Benchmark Details

To support reproducibility, we provide:

- **Dataset availability:** *Poly-FEVER* is derived entirely from public benchmarks (FEVER, Climate-FEVER, SciFact). No private or user data are included. Upon acceptance, the dataset will be released on HuggingFace Datasets (anonymized link during review).
- **Code availability:** All preprocessing, evaluation scripts, prompts, and experiment logs will be released in a public GitHub repository. This includes random seeds, pinned requirements, and command-line scripts.
- **Evaluation:** All models are tested with fixed random seeds. Deterministic inference is enforced where supported.
- **Cost:** Multilingual translation was conducted using Google Cloud Translation, costing $2,644 for 10 non-English languages.

A.1 Dataset Statistics

Character length varies across languages even for identical claims. The main drivers are script density (e.g., logographic vs. alphabetic/abjad), morphology (agglutinative vs. isolating), and orthographic conventions (spacing and punctuation). As a result, the same semantics tend to compress in Chinese/Japanese and expand in languages with richer inflection or compounding (e.g., Tamil, Amharic), consistent with the ranges in Table 5.

A.2 Example Record

Each entry in *Poly-FEVER* is a *single claim* aligned across 11 languages with a *shared* binary veracity label and lightweight topic metadata. Concretely,

Field	Value
ID	1
Label	TRUE
Topic Distribution (top-5)	16, 21, 0, 1, 2
en	Aramais Yepiskoposyan played for FC Ararat Yerevan, an Armenian football club based in Yerevan during 1986 to 1991.
zh-CN	Aramais Yepiskoposyan 于 1986 年至 1991 年间效力于 FC Ararat Yerevan，这是一家位于埃里温的亚美尼亚足球俱乐部。
hi	अरामैस येपिस्कोपोसियन 1986 से 1991 के दौरान येरेवन में स्थित अर्मेनियाई फुटबॉल क्लब एफसी अरारत येरेवन के लिए खेले।
ar	لعب أراميس يبيسكوبوسيان لصالح نادي أرارات يريفان، وهو نادي كرة قدم أرمني مقره في يريفان خلال الفترة من 1986 إلى 1991.
bn	Aramais Yepiskoposyan FC Ararat Yerevan, 1986 থেকে 1991 সাল পর্যন্ত ইয়েরেভানে অবস্থিত একটি আর্মেনিয়ান ফুটবল ক্লাবের হয়ে খেলেছিলেন।
ja	アラマイス・イェピスコポシアンは、1986年から1991年までエレバンに本拠を置くアルメニアのサッカークラブ、FCアララト・エレバンでプレーした。
ko	Aramais Yepiskoposyan은 1986년부터 1991년까지 예레반을 연고로 하는 아르메니아 축구 클럽인 FC Ararat Yerevan에서 뛰었습니다.
ta	Aramais Yepiskoposyan 1986 முதல் 1991 வரை யெரெவனில் உள்ள ஆர்மீனிய கால்பந்து கிளப்பான FC அரரத் யெரெவனுக்காக விளையாடினார்.
th	Aramais Yepiskoposyan เล่นให้กับ FC Ararat Yerevan สโมสรฟุตบอลอาร์มีเนียที่ตั้งอยู่ในเยเรวานระหว่างปี 1986 ถึง 1991
ka	არამაის იეპისკოპოსიანი თამაშობდა FC Ararat Yerevan-ში, სომხურ საფეხბურთო კლუბში, რომელიც დაფუძნებულია ერევანში 1986-1991 წლებში.
am	አራማይስ ዬፒስኮፖስያን ከ1986 እስከ 1991 በዬሬቫን ለነበረው ለአርሜኒያ እግር ኳስ ክለብ ለኤፍሲ አራራት ዬሬቫን ተጫውቷል።

Figure 6: Example record from *Poly-FEVER*. Claims are aligned across 11 languages with a shared truth label and top-5 LDA topics. Cross-script variation highlights differences in length and tokenization.

a record contains: `ID`, `Claim` (parallel text in 11 languages), `Label` (`true`/`false`), and the top-5 `TopicIDs` from LDA. The fields are designed to be minimal yet sufficient for (i) multilingual fact verification, (ii) hallucination analysis, and (iii) topic-aware diagnostics. All fields are required and stored in UTF-8 NFC. Language tags follow BCP-47; we use `en`, `zh-CN`, `hi`, `ar`, `bn`, `ja`, `ko`, `ta`, `th`, `ka`, `am`. The `Label` is *identical* for all languages of the same record; translations are intended to be logically equivalent (Section 4.3 discusses extrinsic conditioning). `TopicIDs` are the indices of the five most probable LDA topics for the *English* surface form (Section A.3); they provide a shared topical anchor across languages.

Lang.	zh-CN	hi	ar	bn	ja	ko	ta	th	ka	am
Min	6	6	14	12	11	10	16	12	16	2
Max	0.6k	1.8k	1.9k	1.9k	0.8k	0.9k	2.2k	1.7k	2.0k	1.2k

Table 5: Character-length ranges (min and max) of claim strings by language in the *Poly-FEVER* metadata.

A.3 Topic Modeling via LDA

LDA models each claim as a mixture of topics, with each topic represented by a distribution of words. The model assumes that claims share underlying themes and assigns probability distributions to capture these relationships. Each claim consists of weighted topic assignments that indicate its alignment with different themes. The model learns these distributions by adjusting word probabilities to match patterns observed in the dataset. We evaluate topic classifications from 0 to 50 and determine that 22 topics (Table 6) achieve the highest coherence scores on *Poly-FEVER*.

To prepare claims for topic modeling, we apply a preprocessing pipeline that standardizes text for analysis. The pipeline converts all text to lowercase, tokenizes words, removes typos, eliminates stopwords, and applies stemming and lemmatization. This process ensures that LDA captures relevant word distributions without noise from text inconsistencies.

After preprocessing, we construct a Gensim Dictionary (Rehurek and Sojka, 2011) and transform the corpus into a Bag-of-Words (BoW) representation. We apply Term Frequency-Inverse Document Frequency (TF-IDF) weighting to refine feature representation. The LDA model runs for 200 iterations to identify topic distributions and optimize coherence scores. The model assigns probabilities to claims based on topic composition, revealing structures that influence hallucination patterns in multilingual verification.

A.4 Translation Filtering & Quality Distribution

We follow the design of prior multilingual fact-checking benchmarks, such as X-FACT (Gupta and Srikumar, 2021b) and XFEVER (Chang et al., 2023), rather than to differences in which issues are salient or how facts are framed in different communities. We evaluate translation quality using the GEMBA (Kocmi and Federmann, 2023a) framework and filter translations using a GEMBA threshold of ≥ 85. While intuitive, our analysis shows that a single global cutoff would introduce systematic bias and remove substantial high-quality data.

(a) Quality is not uniformly distributed across languages. As shown in Table 7, Amharic has 10.6% of translations below 85, while other languages range from 3.6–6.4%. This reflects language-specific translation difficulty rather than claim-level noise. A universal threshold would disproportionately prune low-resource languages such as Amharic, artificially inflating apparent performance by discarding the hardest cases.

(b) Claim-level filtering causes large collateral loss. Translation quality varies across languages

ID	Topic	ID	Topic	ID	Topic
0	Politics	8	Warfare History	16	Soccer
1	Music	9	Military Services	17	Biology/Taxonomy
2	Sports	10	Biology/Genetics	18	Architecture/History
3	Aviation	11	Equestrian	19	Business/Finance
4	Competitive Sports	12	Architecture/Construction	20	Film/Television
5	American Football	13	Entertainment	21	Transportation/Infrastructure
6	Photography/Arts	14	Automotive Racing		
7	Demographics	15	Education		

Table 6: Number-topic references table for LDA topic distribution on 22 topics. This table presents the topic assignments derived from LDA applied to the *Poly-FEVER* dataset. These topic classifications serve as a basis for evaluating the impact of topic awareness.

Lang.	zh-CN	hi	ar	bn	ja	ko	ta	th	ka	am
% < 85	5.4	3.6	3.7	4.0	4.7	6.2	6.1	6.4	4.3	10.6

Table 7: Percentage of translations with GEMBA score less than 85 by language.

for the same claim. As shown in Table 8, 34.5% of claims exhibit mixed quality (i.e., some languages above threshold and others below), while only 0.6% of claims are low-quality across all languages. Thus, filtering at the claim level would discard many otherwise strong translations.

(c) A global threshold removes a large amount of good data. Applying a GEMBA threshold of 85 would remove 35.0% of claims, including 220,769 high-quality translations (28.3%) that are lost solely because another language in the same claim falls below the threshold.

Category	Count	% of Data
Any Language < 85	27,304	35.0%
All Languages < 85	435	0.6%
Mixed Quality	26,869	34.5%
High-Quality Data Lost	**220,769**	**28.3%**

Table 8: Data loss under GEMBA threshold of 85 showing claims removed due to low-quality translations in any language, all languages, or mixed-quality cases, and the resulting loss of high-quality data.

To avoid encoding a specific quality–coverage trade-off into the benchmark, we retain all translations and release per-claim, per-language GEMBA scores. This enables downstream users to define their own strict or standard subsets depending on application needs.

B Causes Exploration

B.1 Effect of Prompts

We provide verbatim text for the three prompt families (General, Language-wise, and Classification) and their translations for all 11 languages. For completeness, we also include the minor variants used in ablations.

We designed both a language-wise prompt and a classification prompt (each with 11 translated versions) to guide LLMs toward producing either *true* or *false* outputs for fact-checking. In early trials, lengthy prompts filled with additional context often confused the models, leading them to ignore specific instructions and produce irrelevant explanations. To mitigate this issue, we refined the prompts through several iterations, prioritizing clarity and brevity so that the task constraints were explicit without adding unnecessary complexity. Ultimately, we developed a concise general prompt to classify factual claims as *true* or *false* based on truthfulness, reducing irrelevant explanations and improving the accuracy of LLM responses.

General Prompt Baseline Evaluation

Label the following claim with truthfulness into: *true* and *false*. No other characters, no punctuation, all lower size.

Nevertheless, LLMs exhibit overenthusiasm by providing detailed explanations or references, straying from the intended binary classification of each claim as *true* or *false*. To counter this, we restructured the language-wise prompt into a three-sentence directive outlining the overall task definition, the precise answer format, and forbidden responses. We found that performance dropped when the prompt and claim were in different languages, so we translated the standardized prompt into the same 11 languages as the claims. This alignment between prompt language and claim language helps curb irrelevant model outputs and ensures consistency in binary fact-checking responses.

Language-wise Prompt Language-wise Prompt for Multilingual Evaluation

Assess the truthfulness of the user's claim and provide a response. Use *true* to indicate that the claim is true, *false* to indicate that it is false. Your response should only consist of *true* or *false*, without any additional characters or punctuation.

Although LLMs are directed to evaluate claims as *true* or *false* without contextual analysis, their responses are integrated sentence structure. Although LLMs are directed to output only *true* or *false*, their responses often appear within broader sentence structures. To improve classification accuracy, we moved beyond simple pattern matching and designed a classification prompt that captures how LLMs express these labels in context.

Classification Prompt Free-form Output Mapping

Classify the input as *true* or *false* based solely on the indicative words or phrases within it. Assign *true* if the input contains affirming terms such as *Correct*, *TRUE*, *really*, or *the truth*. Assign *false* if it contains negating or contradictory terms such as *Fake*, *False*, or any phrasing that signals correction or contradiction. Respond with only *true* or *false* for the input, without any additional text, characters, or punctuation.

By filtering responses through this classification prompt, we ensure outputs conform to a strict binary format. This approach standardizes hallucination detection across multiple languages, preventing inconsistencies in model interpretation.

B.2 External Knowledge Retrieval

We employed the RAG technique, leveraging its state-of-the-art capabilities to bolster the accuracy and relevance of responses produced by LLMs within our fact-checking framework. This system applies the Dense Passage Retrieval (DPR) mechanism (Karpukhin et al., 2020), which utilizes embeddings for document retrieval. RAG revolutionizes NLP by amalgamating generative models with an external knowledge retrieval component, enabling dynamic access to a vast corpus of information. This external augmentation enhances the model's internal knowledge base with pertinent external data during generation.

For external retrieval, we employ the wiki_dpr dataset, an extensive collection of 21 million Wikipedia passages, each adorned with DPR embeddings. These documents are segmented into 100-word, non-overlapping text blocks, optimizing the dataset for precise analysis and the evaluation of DPR's retrieval efficacy.

Leveraging the DPR-embedded Wikipedia corpus, we construct a Facebook AI Similarity Search (FAISS) based dense index over **21 million passages**, enabling sub-second nearest-neighbor retrieval. Each claim is encoded with DPR's Question Encoder, matched against the index, and augmented with the top-5 semantically aligned passages. This augmentation equips the LLM with contextualized evidence that can mitigate *extrinsic hallucinations* by grounding generations in verifiable text. This process retrieves documents that are semantically aligned with the claim, giving the LLM an input context that combines the claim, task constraints, and supporting evidence—helping ground predictions and reduce reliance on uneven internal knowledge across languages.

We retrieve Wikipedia passages in each claim's native language, and then intentionally retrieve from English Wikipedia for all languages to ensure methodological consistency and a uniform retrieval corpus. By fairness, we refer specifically to equalizing the retrieval domain so that models across languages are grounded in the same evidence source, avoiding advantages caused by richer or more frequently updated corpora in high-resource languages. Results for native vs. English Wikipedia retrieval (LLaMA-3.2-3B) with source size and document passages are shown in Table 9. This experiment confirms that mixed-language retrieval introduces comparability issues. Dense-retrieval models, tokenizer tools, and high-quality cleaned Wikipedia dumps are not uniformly available across languages.

Using language-specific corpora would therefore introduce uncontrolled variation in coverage, article quality, and update frequency. Fixing the retrieval corpus to English ensures a uniform factual grounding source, so that cross-lingual performance differences primarily reflect model reasoning and linguistic generalization, rather than retrieval-domain imbalance (Zhang et al., 2021, 2023). This further underscores a core motivation of *Poly-FEVER* that even with retrieval-based mitigation, most languages remain dependent on English-centric retrieval infrastructure, as robust multilingual retrieval resources are still limited.

Metric	en	zh-CN	hi	ar	bn	ja	ko	ta	th	ka	am
Src_size	74G	5.3G	985M	5.5G	1.4G	12G	2.2G	1.3G	1.9G	835M	31M
Doc_passages	21,015,300	5,855,162	948,985	6,421,656	1,201,155	5,256,853	2,986,027	791,163	950,254	599,116	31,489
Without RAG	50.06	43.77	43.77	41.59	38.90	40.31	49.42	40.56	40.18	41.85	39.15
RAG (native)	51.09	46.72	45.70	44.93	44.29	40.82	45.06	39.67	44.80	44.03	39.28
RAG (English)	51.08	48.69	42.90	52.89	46.92	41.75	50.00	39.35	47.37	46.34	39.12

Table 9: Retrieval corpus statistics and hallucination detection accuracy under different retrieval settings.

C Evaluation

C.1 Evaluation Setup (Extended)

The selection of ChatGPT, LLaMA-2, and LLaMA-3.1 was driven by their extensive language support and significant influence in the AI field. Although LLaMA-2 is primarily designed for English, it includes 27 other languages (Touvron et al., 2023), prompting us to examine its non-English hallucination detection. Initial evaluations (see Table 2: columns 4, 5, 6) revealed its multilingual limitations, as its performance approximated random guessing. Therefore, we extended our evaluation to LLaMA-3.1, which introduces improvements in multilingual capabilities and factual consistency.

We conduct evaluations of ChatGPT 3.5 Turbo, LLaMA-2 (7B, 13B, and 70B), and LLaMA-3.1 8B on a server equipped with a 12-core CPU and dual NVIDIA A5500 GPUs, each with 24GB of memory. For the evaluation of the larger LLaMA-2 70B model, we use a model comprising 2 NUMA nodes. Each of these nodes features a 20-core CPU and 4 NVIDIA V100 GPUs with 32GB of memory, interconnected via NVLink to ensure fast and efficient data transfer. The LLaMA-3.1 8B model is evaluated on the same hardware as LLaMA-2 7B and 13B, ensuring a consistent computational environment for comparison. All models are evaluated using the PyTorch deep learning framework, version 2.0.1.

C.2 Temperature Sensitivity Sweep

To examine temperature sensitivity, we evaluate LLaMA-3.1 8B on the TACC Lonestar6 system using three NVIDIA A100 GPUs with 40GB HBM2 memory. We sweep temperature values from 0.1 to 2.0 while keeping the prompt format fixed. Table 10 shows that temperature effects vary by language. English accuracy increases at higher temperatures, while Chinese and Amharic decrease as temperature increases. Japanese and Bengali remain comparatively stable, whereas Tamil and Korean show larger fluctuations. These results indicate that decoding temperature can affect multilingual factuality judgment differently across languages, so a single temperature setting may not be optimal for all languages.

C.3 Hallucination on Multilingual Fact-checking Task

As shown in Table 11, ChatGPT-3.5 shows relatively stable performance across the three source domains: FEVER, Climate-FEVER, and SciFact. English achieves the highest factuality-judgment accuracy across all three datasets, while Arabic has the lowest accuracy. The gap between the highest and lowest language accuracy is 23.11 percentage points for FEVER, 38.22 percentage points for Climate-FEVER, and 35.06 percentage points for SciFact. These results suggest that cross-lingual variation is largest on Climate-FEVER and smallest on FEVER, indicating that domain and language jointly affect multilingual factuality judgment.

LLaMA-2 (7B, 13B, and 70B) results are illustrated in Table 12, 13, and 14. As the size of the LLaMA2 models increases, a noticeable bias towards different languages in LLMs becomes apparent. In the case of LLaMA-2 7B and 13B, Amharic exhibits the poorest performance in general fields, whereas Bengali demonstrates the weakest performance in categorizing science claims. LLaMA2 70B displays significant variation in performance across different languages, particularly in terms of the lowest accuracy. In general and science topics, Tamil exhibits remarkably low estimation rates, registering only 15.00% and 10.17%, respectively.

C.4 Task Feasibility

Poly-FEVER departs from the traditional FEVER formulation by evaluating hallucination detection through controlled factuality judgment rather than evidence retrieval. The task asks a model to judge whether an aligned claim is supported or refuted under a fixed prompt, enabling cross-language comparison while holding claim content and labels constant. Performance reflects a combination of parametric knowledge, language coverage, instruction following, and calibration.

Feasibility without retrieval. A natural concern is whether factuality judgment becomes infeasible without external evidence. We assess this by

Lang.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
en	47.72	47.71	47.74	47.8	47.78	47.89	48.04	47.97	48.06	48.25	48.39	48.52	48.47	48.78	48.79	48.83	49.06	49.07	**49.26**	49.17
zh-CN	**55.79**	55.66	55.53	55.43	55.31	55.28	54.88	54.88	54.67	54.56	54.52	54.28	54.00	54.02	54.02	53.85	53.52	53.47	53.24	52.95
hi	46.72	46.70	46.73	46.98	46.96	47.22	47.58	47.64	47.65	47.86	48.00	48.23	48.23	48.69	48.68	48.54	48.51	48.43	48.63	**48.84**
ar	48.41	48.38	48.49	48.64	48.66	49.16	49.06	49.18	49.28	49.23	49.62	49.60	49.70	49.62	49.92	49.66	49.97	**50.07**	49.99	49.86
bn	42.71	42.85	43.21	43.52	44.03	44.58	45.28	45.40	45.47	46.27	46.28	46.69	46.99	47.27	47.20	47.04	**47.56**	47.52	47.46	47.53
ja	56.08	**56.10**	55.96	55.99	55.96	55.95	55.61	55.62	55.54	55.40	55.37	55.25	55.23	55.23	55.05	55.16	54.70	54.65	54.41	54.29
ko	52.97	52.86	**52.99**	52.78	52.82	52.82	52.78	52.78	52.73	52.58	52.63	52.22	52.41	52.52	52.18	51.96	52.28	51.97	51.88	51.84
ta	55.87	**55.91**	55.71	55.25	54.94	54.63	53.87	53.72	53.72	53.09	53.00	52.45	52.34	52.06	51.76	51.62	51.43	50.70	50.49	50.35
th	49.61	49.76	49.92	49.86	49.89	50.04	50.22	50.15	50.22	50.48	50.45	50.54	50.77	50.61	50.77	50.81	50.66	**50.93**	50.58	50.46
ka	47.86	47.94	48.19	48.13	48.35	48.40	48.59	48.66	48.64	48.79	48.91	**49.26**	49.09	49.04	49.16	49.26	48.95	48.86	48.65	48.61
am	53.75	**53.43**	52.69	51.80	51.26	50.43	49.76	49.32	49.06	47.73	47.12	47.02	46.52	46.16	46.39	46.35	46.22	46.09	46.20	46.48

Table 10: Performance of LLaMA-3.1 8B across temperature values from 0.1 to 2.0 for 11 languages. The highest accuracy for each language is highlighted in bold, while the lowest is underlined. The results provide insights into how different temperature settings influence multilingual fact verification and hallucination patterns across diverse linguistic contexts.

Lang.	FEVER	Climate-FEVER	SciFact
en	**65.89%**	**74.29%**	**71.57%**
zh-CN	58.25%	55.12%	54.83%
hi	52.90%	41.79%	42.14%
ar	45.48%	36.07%	36.51%
bn	55.65%	41.19%	43.15%
ja	55.89%	55.00%	55.56%
ko	57.29%	50.60%	54.11%
ta	55.67%	55.95%	43.58%
th	56.82%	46.31%	44.44%
ka	57.37%	50.36%	49.93%
am	47.53%	39.52%	37.95%

Table 11: Accuracy of hallucination detection by ChatGPT 3.5 on *Poly-FEVER*.

Lang.	FEVER	Climate-FEVER	SciFact
en	**63.35%**	**77.70%**	**70.71%**
zh-CN	58.29%	58.59%	62.46%
hi	48.59%	41.92%	40.33%
ar	50.55%	46.53%	51.52%
bn	49.05%	46.31%	36.04%
ja	58.24%	61.23%	64.54%
ko	56.67%	52.36%	63.87%
ta	49.54%	46.56%	35.03%
th	53.90%	44.73%	48.49%
ka	47.39%	43.00%	37.31%
am	43.42%	48.15%	45.26%

Table 12: Accuracy of hallucination detection by LLaMA-2 7B.

Lang.	FEVER	Climate-FEVER	SciFact
en	**64.27%**	**72.00%**	**72.44%**
zh-CN	59.88%	62.82%	66.57%
hi	54.33%	56.06%	57.33%
ar	54.97%	56.77%	61.41%
bn	52.30%	51.01%	52.25%
ja	59.57%	66.46%	67.65%
ko	58.74%	62.97%	67.36%
ta	50.86%	50.38%	61.02%
th	53.46%	45.57%	58.79%
ka	53.45%	55.00%	62.69%
am	48.64%	44.44%	53.68%

Table 13: Accuracy of hallucination detection by LLaMA-2 13B.

Lang.	FEVER	Climate-FEVER	SciFact
en	**64.56%**	**78.42%**	**75.32%**
zh-CN	38.75%	49.72%	42.08%
hi	45.68%	45.96%	33.67%
ar	32.97%	30.03%	31.92%
bn	33.87%	19.46%	18.02%
ja	41.37%	51.27%	42.77%
ko	46.06%	47.88%	51.31%
ta	19.47%	19.85%	10.17%
th	48.09%	31.22%	30.65%
ka	46.15%	40.00%	33.03%
am	26.34%	11.11%	22.11%

Table 14: Accuracy of hallucination detection by LLaMA-2 70B.

evaluating recent open and closed LLMs on *Poly-FEVER* under the same fixed prompting protocol. Table 15 shows that stronger recent models perform above simple binary baselines across multiple languages, with several exceeding the historical English FEVER baseline of 50.9%. This indicates that the task is feasible without retrieval and captures meaningful differences in multilingual factuality judgment. These scores reflect a combination of parametric knowledge, language coverage, prompt compliance, and calibration.

Interpretation of low scores. Some models obtain substantially low accuracy in certain languages, such as LLaMA-2-70B on Tamil. These results should be interpreted as multilingual reliability failures rather than as evidence that the task is invalid. In our evaluation, outputs that fail to follow the required binary format are strictly scored as incorrect. Older or less instruction-following models may produce conversational filler, repetitions, or non-binary answers, especially in lower-resource language settings, consistent with instruction-following failures reported in multilingual evaluation (Li et al., 2024). In addition, models may collapse toward a single

label under uncertainty, a behavior related to label bias (Zhao et al., 2021) and sycophancy or agreement bias (Sharma et al., 2023). If the collapsed label is poorly matched to the label distribution of a language subset, accuracy can fall far below a simple binary baseline. These failures are part of what *Poly-FEVER* is designed to expose: multilingual hallucination detection requires not only factual knowledge, but also language coverage, instruction following, and calibrated binary decision making.

Model	en	zh	ar	ja	ko	am
LLaMA-3.2-3B	50.06	43.77	41.59	40.31	49.42	39.15
Phi-3.5-4B	**63.17**	46.10	38.23	34.27	48.91	26.70
Mistral-7B	**65.20**	**52.11**	45.80	40.12	56.60	33.41
Gemma3-27B	**64.18**	**62.90**	49.42	**52.50**	**62.13**	41.21
Qwen-2.5-32B	**58.54**	**54.94**	**49.42**	**53.15**	46.73	**41.21**

Table 15: Accuracy (%) of recent models on Poly-FEVER across selected languages.