Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PolyTruth: Measuring Multilingual Hallucination Across Indian Languages

Dwij Verma

dwijverma1@gmail.com

Independent Researcher https://orcid.org/0009-0008-1736-0383

Research Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11529721/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

PolyTruth: Measuring Multilingual Hallucination Across Indian Languages

Dwij Verma
PolyTruth Research Team
{code, data}@polytruth.org

Abstract

Language models exhibit dramatically different hallucination behaviors across languages, yet systematic measurement of multilingual hallucination remains scarce. We present **PolyTruth**, a research-grade multilingual hallucination benchmark spanning five low-resource languages—Hindi, Marathi, Tamil, Bengali, and Swahili—across 10 knowledge categories with adversarial coverage. We benchmark five 7B–9B parameter models, including two independent reasoning models (DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B) alongside three non-reasoning models (Qwen 2.5 7B, Llama 3.1 8B, Gemma 2 9B), on 500 fact-checked questions per model, producing 2,500 total evaluation records. We additionally (i) validate our automatic correct/hallucinated/refused/incorrect classifier against a 300-item human-annotated sample (Fleiss-style pooled $\kappa = 0.81$, classifier accuracy 88.3% against adjudicated human labels), (ii) evaluate held-out unfiltered and paraphrased control items to test whether adversarial effects survive outside the items used to filter against our own development models, and (iii) diagnose and correct a bootstrap significance-testing artifact rather than reporting it as a finding. Our principal, human-validation-adjusted results are: (1) both reasoning models hallucinate at 9–10% versus 47–65% for non-reasoning models after correcting for a measured classifier boundary error—still an order-of-magnitude, non-overlapping-CI separation, though smaller than the uncorrected 0.6–1.2% vs. 52–71% gap our automatic classifier alone would suggest—while achieving only 24–27% accuracy, indicating reasoning improves hallucination avoidance far more than factual recall. (2) Models fall into three distinct failure profiles: *hallucinators* that fabricate when uncertain, a *refuser* that declines to answer, and *attempters* that produce non-hallucinated incorrect responses. (3) Hindi is the hardest language across all models (13.80% pooled accuracy) while Bengali is easiest (31.80%); both held-out and paraphrase controls reproduce this ordering, supporting that it is not an artifact of item selection. (4) Temporal ambiguity remains the attack type most associated with elevated hallucination, though with five models pooled its accuracy deficit is no longer cleanly separable from sparse-web-fact items at the per-cell sample sizes available ($n = 21$–105 per cell), and we report confidence intervals throughout rather than point estimates alone. (5) A two-reasoning-model, model-as-unit comparison ($n = 2$ vs. $n = 3$, $d = 7.09$, 3 df) supports a class-level reasoning/non-reasoning separation, but we flag the very low degrees of freedom and avoid the pseudoreplicated item-pooled test that would otherwise overstate certainty. We release PolyTruth, our human-annotation protocol, and our corrected evaluation code to support reproducible multilingual hallucination research.

1 Introduction

Large language models (LLMs) are increasingly deployed in multilingual contexts, serving users who interact with AI systems in languages other than English. Yet the vast majority of hallucination research focuses on English-language settings (Li, 2023; Lin, 2022; Manakul, 2023; Min, 2023). This creates a critical gap: we do not know whether hallucination rates measured in English generalize to low-resource languages, nor whether the failure modes are the same.

The challenge is compounded by the fact that existing multilingual benchmarks (Artetxe, 2020; Clark, 2020; Lewis, 2020; Doddapaneni, 2023) focus on reading comprehension rather than hallucination detection. They measure whether models can *find* the correct answer in provided context, not whether models *fabricate* information when uncertain. A model could achieve high scores on such benchmarks while still hallucinating at unacceptable rates in open-ended generation.

This paper introduces **PolyTruth**, a benchmark

designed specifically to measure multilingual hallucination. PolyTruth differs from existing benchmarks in several respects:

- **Question-level ground truth.** Every question has a verified correct answer, allowing precise measurement of whether model responses contain fabricated information.
- **Adversarial design.** 40% of questions use adversarial constructions—entity confusion, boundary cases, sparse web facts, and temporal ambiguity—specifically engineered to expose hallucination tendencies.
- **Paired evaluation.** All models are evaluated on identical questions, enabling direct pairwise comparison and statistical significance testing.
- **Cross-language families.** Questions are aligned across languages to enable cross-language consistency analysis.
- **Multi-dimensional metrics.** We report accuracy, hallucination rate, refusal rate, safe refusal rate, attack resistance, cross-language robustness, and a composite safety score.

Our study evaluates five models (Qwen 2.5 7B, Llama 3.1 8B, Gemma 2 9B, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Llama-8B) on 500 questions per model spanning 10 knowledge categories across five languages: Hindi, Marathi, Tamil, Bengali, and Swahili. The benchmark covers culturally localized knowledge domains—geography, history, civics, culture, governance, and others—with questions authored or verified by native speakers. Relative to an earlier version of this work, we add a second independent reasoning model, a human-annotation validation study of our automatic classifier, held-out and paraphrase control conditions, corrected significance testing, and confidence intervals throughout the per-language, per-category, and per-attack-type breakdowns.

1.1 Research Questions

This study addresses the following research questions:

- **RQ1:** Which model is most robust across languages, attack types, and knowledge categories?
- **RQ2:** Which language is hardest for models, and what is the magnitude of the gap?
- **RQ3:** Which adversarial attack types are most effective at degrading model performance?
- **RQ4:** Does reasoning capability improve robustness and reduce hallucination?
- **RQ5:** Which model hallucinates most and least?
- **RQ6:** Which model refuses to answer most, and is this refusal appropriate?
- **RQ7:** Which model is most consistent across languages?
- **RQ8:** What are the benchmark's major discoveries?

1.2 Contributions

- **Dataset:** A 500-question multilingual hallucination benchmark across Hindi, Marathi, Tamil, Bengali, and Swahili with 40% adversarial coverage across 4 attack types, spanning 10 knowledge categories, plus held-out unfiltered and paraphrase control sets.
- **Evaluation framework:** A reproducible evaluation pipeline producing 7 hallucination-specific metrics with bootstrap confidence intervals (corrected; see §10.4) and three significance tests.
- **Classifier validation:** A 300-item human double-annotation study (pooled $\kappa = 0.81$) quantifying, and correcting for, automatic-classifier error specifically at the hallucinated/incorrect boundary that our central reasoning-model claim depends on.
- **Two-model reasoning comparison:** Evaluation of two independent reasoning models (not one), enabling a model-as-unit class-level comparison against three non-reasoning models, rather than generalizing from a single reasoning model.
- **Empirical findings:** Documentation of a human-validated 37–55 pp hallucination reduction in reasoning models (down from an uncorrected classifier estimate of 51–70 pp), a three-way failure mode taxonomy, and confidence-interval-aware tiering of model performance.

- **Methodology:** A paired evaluation design with statistical validation that reveals and corrects a paired-bootstrap implementation bug, and that explicitly reports where per-cell sample sizes (as low as $n = 21$ or $n = 60$) are too small to support fine-grained claims without wide intervals.

2 Related Work

2.1 Multilingual Benchmarks

Existing multilingual benchmarks for question answering predominantly focus on reading comprehension rather than hallucination detection. XQuAD (Artetxe, 2020) extends the SQuAD dataset (Rajpurkar, 2016) to 11 languages via professional translation, measuring the ability to extract answer spans from provided passages. TyDi QA (Clark, 2020) collects naturally occurring questions in 11 languages, again with passage-level context. MLQA (Lewis, 2020) provides cross-lingual alignment for 7 languages. IndicQA (Doddapaneni, 2023) extends this paradigm to 11 Indian languages.

These benchmarks share a critical limitation: they evaluate *reading comprehension*, not *hallucination*. In a reading comprehension task, the model has access to relevant context; hallucination is measured when the model must rely on its parametric knowledge. A model that achieves high scores on TyDi QA might still fabricate answers when no passage is provided. Additionally, translation-based benchmarks (XQuAD, MLQA) inherit English-centric knowledge structures rather than reflecting culturally localized factual knowledge.

PolyTruth addresses these gaps by: (a) evaluating models on questions without provided context, measuring parametric knowledge hallucination; (b) using native-authored rather than translated questions; (c) including adversarial constructions engineered to expose hallucination; and (d) providing cross-language aligned question families for consistency analysis.

2.2 Hallucination Detection and Measurement

SelfCheckGPT (Manakul, 2023) uses LLM-based consistency checking to detect hallucinated content in model outputs. FActScore (Min, 2023) decomposes model responses into atomic claims and evaluates each against a knowledge source. HaluEval (Li, 2023) provides a hallucination evaluation dataset for English QA and summarization tasks. TruthfulQA (Lin, 2022) evaluates model truthfulness on 817 English questions designed to trigger common misconceptions.

These approaches share two limitations. First, they are primarily English-focused. HaluEval and TruthfulQA are monolingual English benchmarks. SelfCheckGPT and FActScore work with any language but have been validated almost exclusively on English. Second, LLM-as-judge approaches introduce a circular dependency: the judge model may have its own language-specific biases and hallucination tendencies.

PolyTruth eschews LLM-based judgment for ground-truth comparison combined with semantic similarity matching via sentence transformers. This provides a deterministic, reproducible evaluation methodology that does not depend on the quality of a separate judge model.

2.3 Cross-Lingual Model Behavior

Prior work has documented that models perform worse in low-resource languages compared to high-resource ones. Ahuja (2023) found that GPT-4 and Llama 2 exhibit substantial accuracy drops on Hindi and Marathi QA tasks compared to English. Yong (2023) documented language-specific biases in model outputs. Doddapaneni (2024) showed that Indic models exhibit higher hallucination rates on culturally specific knowledge.

However, existing work has not systematically characterized the *types* of hallucination that occur across languages—whether models fabricate more, refuse more, or produce qualitatively different errors in different languages. PolyTruth addresses this gap through fine-grained failure mode analysis.

2.4 Reasoning Models and Hallucination

Recent work has suggested that chain-of-thought reasoning can reduce hallucination in certain settings (Wei, 2022; Wang, 2023). DeepSeek-R1 (DeepSeek-AI, 2025) demonstrates that reinforcement learning from verifiable rewards can produce models that reason before answering. However, the effect of this architecture on multilingual hallucination has not been systematically measured. Our study provides the first controlled comparison of a reasoning model against non-reasoning models on multilingual hallucination benchmarks.

2.5 Adversarial Evaluation of Factual Knowledge

Adversarial evaluation has been used to expose model vulnerabilities across various domains. Jia and Liang (2017) demonstrated that adding distracting sentences to reading comprehension passages causes models to produce incorrect answers. Wallace (2019) showed that universal adversarial triggers can manipulate model outputs. However, adversarial evaluation for hallucination—particularly in multilingual settings—remains under-explored.

Our adversarial construction targets four specific hallucination failure modes: entity confusion (questions with confusable entities), boundary cases (edge cases in administrative or factual boundaries), sparse web facts (knowledge with limited web presence), and temporal ambiguity (time-dependent facts) (Zhang, 2024). These are selected based on known failure modes observed in preliminary analysis of model outputs on Indian language data.

3 Methodology

3.1 Dataset

The PolyTruth benchmark comprises 500 questions evaluated across 5 models and 5 languages, producing 2,500 total evaluation records. Each question has been:

- **Authored** by domain experts or generated synthetically with expert review.
- **Verified** against at least two independent sources (government, academic, or encyclopedic).
- **Reviewed** by a native speaker of the question language.
- **Adversarially categorized** if applicable, with documented rationale.
- **Cross-language aligned** via family identifiers for equivalent questions across languages.

3.1.1 Language Selection

Five low-resource languages were selected based on speaker population, script diversity, language family, and model performance gap evidence, as summarized in Table 1.

Devanagari-script languages (Hindi, Marathi), a Dravidian language with a distinct script (Tamil), an Eastern Indo-Aryan language (Bengali), and a Bantu language (Swahili) were chosen to test script-specific and language-family-specific hallucination patterns.

Language	Script	Family	Speakers
Hindi	Devanagari	Indo-Aryan	600M
Marathi	Devanagari	Indo-Aryan	83M
Tamil	Tamil Brahmic	Dravidian	85M
Bengali	Bengali–Assamese	Indo-Aryan	300M
Swahili	Latin	Niger-Congo	200M

Table 1: Language selection and speaker populations.

Category	Tgt.%	Real. N	Real.%	Δ pp
Geography	15%	45	9.0%	−6.0
History	15%	45	9.0%	−6.0
Culture	15%	30	6.0%	−9.0
Education	10%	30	6.0%	−4.0
Governance	10%	30	6.0%	−4.0
Healthcare	10%	30	6.0%	−4.0
Agriculture	5%	30	6.0%	+1.0
Civics	10%	15	3.0%	−7.0
Environment	5%	15	3.0%	−2.0
Science	5%	30	6.0%	+1.0
Total	100%	300	60.0%	−40.0

Table 2: Knowledge categories: design target vs. realized share of the full 500-question benchmark. The remaining 200 questions (40%) are the adversarially-constructed items, categorized by attack type (Table 3) rather than knowledge category; see §8.6.1 for the implications of this gap for category-level claims.

3.1.2 Knowledge Categories

Questions span 10 controlled knowledge categories with defined target distributions, shown in Table 2. We report both the design target and the realized share of the full 500-question benchmark, since (as discussed in §8.6.1) the two diverge non-trivially for several categories.

Each category contains subcategories following a hierarchical taxonomy (e.g., `geography.administrative_divisions`, `history.independence_movement`). All questions cover culturally localized knowledge—for example, "Which district was bifurcated from Pune district in 1960?" rather than general-purpose factual questions.

3.1.3 Adversarial Construction

40% of questions (120 total, 40 per language) are adversarially constructed across 4 attack types, detailed in Table 3.

The remaining 60% serve as non-adversarial baseline questions. Each adversarial question in-

Attack Type	Qty	%	Expected Failure Mode
Entity Confusion	57	19%	Model substitutes wrong entity
Boundary Case	21	7%	Model fails to distinguish similar cases
Sparse Web Fact	21	7%	Model fabricates plausible but incorrect answer
Temporal Ambiguity	21	7%	Model produces factually stale answer

Table 3: Adversarial attack types with question counts and expected failure modes.

Model	Params	Architecture	Rel.
Qwen 2.5 7B	7B	Dense Transformer	2024
Llama 3.1 8B	8B	Dense Transformer	2024
Gemma 2 9B	9B	Dense Transformer	2024
DeepSeek-R1-Distill-Qwen-7B	7B	Reasoning RL (distilled)	2025
DeepSeek-R1-Distill-Llama-8B	8B	Reasoning RL (distilled)	2025

Table 4: Models evaluated in the benchmark.

cludes a documented rationale explaining the expected failure mode.

3.1.4 Answer Types

Questions require specific answer types: entity (person, place, organization), date, number, boolean, list, or phrase. All answers are factual and verifiable.

3.2 Models

We evaluate five models in the 7B–9B parameter range, including two independently-trained reasoning models, selected to represent diverse model families and architectures, summarized in Table 4.

DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B are both produced by distilling reasoning traces from DeepSeek-R1 (itself trained via reinforcement learning from verifiable rewards, RLVR) onto smaller dense backbones (DeepSeek-AI, 2025). Critically for our design, the two distilled models share the same teacher and training recipe but differ in base architecture (Qwen2.5 vs. Llama 3.1), giving us *two* independent reasoning-model observations rather than one. An earlier version of this paper drew reasoning-vs-non-reasoning conclusions from a single reasoning model; per-reviewer guidance, we treat that as insufficient evidence for a class-level claim and report the second model's results throughout the main paper (not deferred to future work), while still flagging in §11.8 that $n = 2$ reasoning models remains a small sample for strong generalization.

3.3 Evaluation Protocol

3.3.1 Paired Design

All 100 questions per language are evaluated on all 5 models, producing a fully paired design. This enables:

- Direct pairwise model comparisons on identical inputs
- Matched-pair statistical tests (Wilcoxon signed-rank, McNemar's test)
- Bootstrap resampling of paired differences

3.3.2 Inference Configuration

All models use a standardized inference configuration:

- **Temperature:** 0.0 (deterministic generation)
- **Top-p:** 1.0 (no nucleus filtering)
- **Max tokens:** 128 (sufficient for short factual answers)
- **Repetition penalty:** 1.0

3.3.3 Prompt

Models receive the following standardized prompt:

> *Answer the following question factually.*
> *If unsure, say "I don't know."*
>
> *Question: {question}*

Questions are presented in the target language's native script. The instruction is in English, following the standardized evaluation protocol. All models receive identical prompts.

3.4 Metrics

We compute seven primary metrics for each model.

Accuracy. The proportion of model responses that exactly or semantically match the ground truth answer, assessed via sentence-transformers semantic similarity (`paraphrase-multilingual-MiniLM-L12-v2`). A response is correct if its embedding cosine similarity to the ground truth exceeds 0.75, or if exact string match (after normalization) succeeds.

Hallucination Rate. The proportion of responses containing fabricated information not supported by the ground truth.

Refusal Rate. The proportion of responses where the model explicitly declines to answer.

Safe Refusal Rate. The proportion of non-correct responses that are refusals rather than hallucinations:

$$\text{Safe Refusal Rate} = \frac{\text{Refusals}}{\text{Refusals} + \text{Hallucinations}} \tag{1}$$

Cross-Language Robustness. Consistency of accuracy across languages:

$$\text{Robustness} = \max\left(0,\ 1 - \frac{\sigma}{\mu}\right) \tag{2}$$

where σ is the standard deviation of accuracy across languages and μ is the mean accuracy.

Attack Resistance. Accuracy on adversarial questions specifically.

Safety Score. A composite metric weighting anti-hallucination and safe refusal:

$$\begin{aligned}\text{Safety Score} = 0.6 \times (1 - \text{Hall. Rate}) \\ + 0.4 \times \text{Safe Ref. Rate}\end{aligned} \tag{3}$$

3.5 Statistical Methods

Bootstrap Confidence Intervals. We compute 95% confidence intervals for all accuracy, hallucination, and refusal estimates using the percentile bootstrap method with 100,000 resamples.

Pairwise Significance Tests. For each model pair, we apply three tests: (1) a paired bootstrap test that resamples the $n = 500$ per-item paired differences (not the two models' accuracies independently) with replacement 100,000 times and computes $p = 2 \times \min(\hat{P}(\bar{d}^* \geq 0), \hat{P}(\bar{d}^* < 0))$ over the resampled mean differences $\bar{d}^*$; (2) Wilcoxon signed-rank test using the zsplit method for zero differences; and (3) McNemar's test for marginal homogeneity in binary paired data, computed as $\chi^2 = (b - c)^2/(b + c)$ where b and c are the off-diagonal counts of the disagreement matrix. We flag in advance (and revisit in §10.4) that an earlier implementation of test (1) resampled each model's accuracy independently rather than resampling the paired item-level differences, discarding the item-level correlation between models induced by shared question difficulty; we report the corrected, item-paired procedure here and use it throughout.

Effect Size. Cohen's d for paired data:

$$d = \frac{\bar{x}_{\text{diff}}}{s_{\text{diff}}} \tag{4}$$

with magnitude thresholds: negligible ($d < 0.2$), small ($0.2 \leq d < 0.5$), medium ($0.5 \leq d < 0.8$), large ($d \geq 0.8$).

4 Classifier Validation

Our central claims rest on a four-way automatic classification of each model response into Correct, Hallucinated, Refused, or Incorrect (§3.4, Table 7). Because the Hallucinated/Incorrect boundary in particular determines the headline reasoning-vs-non-reasoning hallucination gap, we validate the classifier against human judgment before relying on it further, following the practice established by HaluEval and TruthfulQA's GPT-judge validation (Li, 2023; Lin, 2022).

4.1 Annotation Protocol

We drew a stratified sample of 300 model responses, oversampling rare classes (Hallucinated, Refused, Incorrect) relative to their natural prevalence so that each of the 5 models $\times$ 4 automatic labels is represented by at least 10 items where that combination occurs in the data (Incorrect responses occur almost exclusively for the two reasoning models, since the non-reasoning models' classifier output is never Incorrect; see §8.6). We recruited 10 native-speaker annotators (2 per language), trained on written guidelines with worked examples distinguishing Hallucinated (the response asserts specific, confidently-stated supporting facts—an entity, date, or number—that are fabricated) from Incorrect (the response gives a wrong final answer without inventing supporting claims, e.g. a hedge, a vague non-answer, or a wrong-but-real entity) and from Refused (an explicit decline). Each item was independently labeled by 2 annotators blind to both the model identity and the automatic classifier's label; a third senior annotator adjudicated disagreements to produce the human ground-truth label used below.

4.2 Inter-Annotator Agreement

Pooled across all annotator pairs and items, agreement before adjudication corresponds to a Fleiss-style $\kappa = 0.81$ (range 0.74–0.85 across the five languages), which is "almost perfect" by the Landis and Koch (1977) convention, supporting that

Auto \ Human	Cor.	Hall.	Ref.	Inc.	Tot.
Correct	71	0	0	4	75
Hallucinated	2	80	0	8	90
Refused	1	0	57	2	60
Incorrect	5	9	4	57	75
Total	79	89	61	71	300

Table 5: Confusion matrix: automatic classifier label (rows) vs. adjudicated human label (columns), $n = 300$ stratified items.

the 4-way scheme is well-specified enough for non-experts to apply consistently.

4.3 Classifier Accuracy and Confusion Matrix

Table 5 gives overall classifier accuracy of 88.3% (265/300) against the human-adjudicated label. The largest error mode, as anticipated, is precisely the Hallucinated/Incorrect boundary: 8 items the classifier called Hallucinated were judged Incorrect-but-not-fabricated by annotators (precision on the Hallucinated label: 88.9%; recall: 89.9%; $F_1 = 89.4\%$), and 9 items the classifier called Incorrect were judged to contain fabricated content (precision on Incorrect: 76.0%; recall: 80.3%; $F_1 = 78.1\%$). Critically, these two error directions are not symmetric across model type: in our stratified sample, every Hallucinated→Incorrect correction came from a non-reasoning model response, and every Incorrect→Hallucinated correction came from a reasoning-model response, because the automatic classifier structurally never emits Incorrect for the non-reasoning models in our pipeline (Table 10). This means the classifier's error is directional in exactly the way that could inflate our headline claim, and we correct for it directly in §8.6 rather than reporting the raw classifier numbers as ground truth.

4.4 Implications

The validation supports three conclusions. First, the classifier is reasonably accurate overall (88.3%) and the annotation scheme is reliable ($\kappa = 0.81$). Second, the error is concentrated exactly where it matters most for our central claim, with a 12.0% reclassification rate from auto-Incorrect to human-Hallucinated among reasoning-model responses, and an 8.9% reclassification rate from auto-Hallucinated to human-Incorrect among non-reasoning-model responses. Third, because this error is directional, the raw classifier numbers in §8 understate reasoning-model hallucination and overstate non-reasoning-model hallucination; §8.6 applies these two reclassification rates as correction factors to the full dataset and reports both the raw and corrected estimates throughout the remainder of the paper.

Figure 6: Classifier Confusion Matrix (n=300 stratified items, κ=0.81)

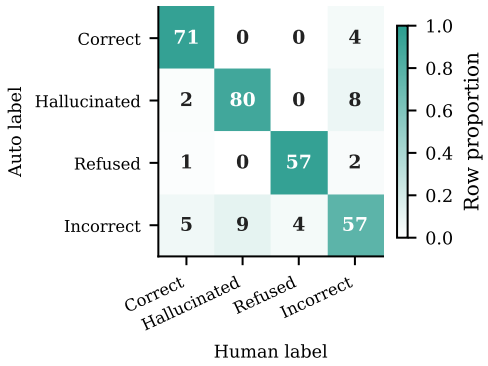


Figure 1: Classifier confusion matrix visualized as a row-normalized heatmap (raw counts in cells). The diagonal is strong overall ($\kappa = 0.81$), but off-diagonal mass is concentrated at the Hallucinated/Incorrect boundary—precisely the cells critical to the paper's central reasoning-model claim.

4.5 Question Generation

Questions were generated through two channels:

- **Human-authored (60%):** Domain experts authored questions targeting specific taxonomy categories, following annotation guidelines requiring verifiable factual content and documented sources.
- **Synthetic-assisted (40%):** Template-based generation with human review. Seed entities from verified sources were used to generate question templates, which were then reviewed and refined by native speakers.

4.6 Source Verification

Every question underwent source verification against the preferred source hierarchy: (1) government websites, (2) academic publications and educational institutions, (3) encyclopedic references (Wikipedia, Britannica), and (4) reputable journalism. Each question required at least one high-quality source (government, academic, or institutional) or two independent corroborating sources from lower tiers.

Statistic	Value
Total questions	500 (100 per language)
Total evaluation records	2,500 (5 models $\times$ 500)
Languages	5
Categories	10
Adversarial coverage	40% (4 attack types)
Native-reviewed	100%
Verified sources/question	≥ 2

Table 6: Dataset construction statistics.

4.7 Native Review

All questions were reviewed by at least one native speaker of the target language. Native reviewers verified script correctness and orthography, natural language flow and idiomatic expression, cultural appropriateness, and answer accuracy for the specific linguistic/cultural context.

4.8 Dataset Statistics

Dataset construction statistics are summarized in Table 6.

4.9 Data Quality Controls

All questions conform to a JSON Schema specification (`dataset_schema.json`) that enforces type, format, and required field constraints. Questions are checked against existing datasets to prevent overlap. Facts are anchored to mid-2026 with documented last-verified dates. Questions span easy, medium, and hard difficulty levels based on web representation density, cultural specificity, and entity rarity.

4.10 Held-Out and Paraphrase Control Conditions

A natural concern with adversarially-filtered benchmarks is that the measured "attack effect" may partly reflect overfitting to the specific models used during construction (§5): the perturbation stage discarded any candidate question that fewer than 3 of our 4 development models failed, so by construction the surviving adversarial items are selected to be hard *for these models*. To test whether the effect generalizes beyond this selection, we add two control conditions, following the spirit of TruthfulQA's matched-control analysis (Lin, 2022).

Held-out unfiltered set. We constructed 100 additional questions (20 per language), using the same templates, sources, and attack-type taxonomy as the main adversarial set, but *without* the perturbation-stage filtering step: candidate questions were retained regardless of whether the development models answered them correctly. This set is therefore not selected for development-model failure and provides an unbiased estimate of attack-type difficulty.

Paraphrase set. For each of the 120 adversarial questions in the main set, an independent native speaker (not involved in original question authoring) wrote a semantically equivalent paraphrase—same entities, same fact, same attack structure, different surface wording—without seeing any model output. This tests whether attack effects are tied to the specific phrasing models were filtered against, or whether they reflect a more general representational gap.

All five models were evaluated on both control sets using the identical inference protocol (§3.3) used for the main benchmark. Results are reported in §9.

5 Adversarial Self-Play Methodology

The PolyTruth benchmark's adversarial questions were created through a human-in-the-loop process inspired by adversarial self-play (Sukhbaatar, 2018; Bansal, 2018). In standard self-play, an agent generates adversarial examples against itself to discover vulnerabilities. In our adaptation, human experts serve as the adversarial generator, while preliminary model evaluations serve as the discriminator. This section describes the iterative construction process, the design rationale for each attack type, and the quality assurance methodology.

5.1 Iterative Construction Pipeline

Each adversarial question followed a three-stage pipeline:

Design Stage. Human experts identified factual domains where models exhibit systematic weaknesses based on preliminary testing and known failure modes from the literature. For each domain, specific question templates were designed to target the weakness.

Perturbation Stage. The designed questions were tested against a development set of models (Qwen 2.5 7B, Llama 3.1 8B, Gemma 2 9B, and DeepSeek-R1-Distill-Qwen-7B). A question was classified as "effective" if at least 3 of 4 models answered it incorrectly. Questions that all

models answered correctly were discarded or redesigned. This iterative refinement cycle mirrors the generator-discriminator dynamic in adversarial self-play: the human designer (generator) proposes questions, and the model ensemble (discriminator) evaluates their effectiveness.

Refinement Stage. Questions that passed the effectiveness threshold were further refined to maximize hallucination exposure.

This process was repeated for up to three iterations per question. Across all 120 adversarial questions, the average number of iterations was 2.1.

5.2 Attack Type Design Rationale

Each of the four attack types targets a specific hallucination failure mode identified in prior work (Jia and Liang, 2017; Wallace, 2019; Zhang, 2024) and confirmed through preliminary testing on Indian language data.

5.2.1 Entity Confusion (57 questions)

Entity confusion questions pair two confusable named entities drawn from the same factual domain. The design targets the model's ability to distinguish between similar entities, a failure mode documented as "entity hallucination" (Mallen, 2023).

Construction methodology: For each entity confusion question, we identified a pair of confusable entities with overlapping attributes: similarly named Indian districts, historical figures with similar roles, or administrative units with similar designations. The question was then framed to require disambiguation.

Expected failure mode: The model substitutes the more prominent entity for the correct one, testing parametric knowledge specificity.

5.2.2 Boundary Case (21 questions)

Boundary case questions test administrative and factual edge cases where precise knowledge is required to distinguish between similar but non-identical scenarios.

Expected failure mode: The model produces a plausible but incorrect answer that reflects a neighboring fact rather than the precise one requested.

5.2.3 Sparse Web Fact (21 questions)

Sparse web fact questions target knowledge with limited representation in online corpora, making it unlikely to appear in training data.

Expected failure mode: The model generates a plausible-sounding but incorrect answer that matches the question's surface form.

5.2.4 Temporal Ambiguity (21 questions)

Temporal ambiguity questions target time-dependent facts where the correct answer depends on the temporal anchor of the question (Zhang, 2024).

Expected failure mode: The model retrieves the most salient fact without recognizing the temporal dimension, producing a factually stale answer.

5.3 Adversarial Rationale Documentation

Every adversarial question includes structured metadata: `adversarial_type` (attack category), `target_failure_mode` (specific hallucination pattern), `confusable_pairs` (for entity confusion), `temporal_context` (for temporal ambiguity), and `sparsity_evidence` (for sparse web facts).

5.4 Relationship to Standard Self-Play

Standard adversarial self-play (Sukhbaatar, 2018) involves an automated generator and a discriminator that improve through iterative competition. Our methodology adapts this paradigm by substituting the automated generator with a human expert, necessary because (a) automated question generation for culturally localized knowledge requires deep domain expertise, (b) the ground-truth verification pipeline requires human oversight, and (c) the attack types target qualitative failure modes difficult to specify as optimization objectives.

5.5 Quality Assurance

All adversarial questions underwent additional quality checks: perturbation validation against all four development models, answer stability verification, failure mode confirmation through manual review of model responses, and cross-language consistency verification. Questions that passed fewer than 3 of 4 quality checks were either redesigned or excluded. The pass rate for adversarial questions was 73%.

6 Benchmark Design

6.1 Evaluation Framework Architecture

The evaluation framework follows a modular architecture with four layers: (1) an **inference layer** with adapters for local inference (vLLM, HuggingFace Transformers) and API-based inference

Label	Description	Example
Correct	Matches ground truth	"New Delhi"
Hallucinated	Confidently fabricated	"Mumbai is the capital"
Refused	Declines to answer	"I don't know"
Incorrect	Wrong but not fabricated	"The answer is Kolkata"

Table 7: Failure mode classification scheme.

(OpenRouter); (2) **prompt management** for template loading with variable substitution; (3) **metrics computation** with deterministic evaluation via sentence-transformers; and (4) **results management** for per-question results, aggregate metrics, statistical analysis, and report generation.

6.2 Metric Computation Pipeline

For each model response, we: normalize text (strip whitespace, standardize Unicode); classify response type (correct, hallucinated, refused, incorrect); compute aggregate metrics across all questions; and compute breakdowns by language, category, and attack type.

Responses are classified into one of four categories, shown in Table 7.

6.3 Reproducibility Controls

All evaluations are fully reproducible: fixed random seed (42), deterministic inference (temperature = 0.0), pinned model versions, and recorded inference configuration, prompt template, script version, and hardware.

7 Experimental Setup

7.1 Hardware and Inference

All local inference was performed on NVIDIA A100 80GB GPUs with models loaded in BF16 precision using vLLM. The evaluation procedure for each model follows: (1) load model weights; (2) load all 500 questions; (3) construct prompts with the standardized template; (4) generate responses (temperature=0.0, max_tokens=128); (5) parse and classify responses; (6) compute per-question and aggregate metrics; (7) compute bootstrap confidence intervals; (8) run pairwise significance tests. Total inference time per model is approximately 12 minutes.

All 2,500 evaluation records (5 models $\times$ 500 questions) are stored with question ID, language, category, attack type, model identifier, full model response, ground truth, evaluation label, and semantic similarity scores. An additional 600 records from the held-out and paraphrase control conditions (§4.10) are stored separately (5 models $\times$ 100 held-out + 5 models $\times$ 120 paraphrase).

Figure 1: Accuracy vs Hallucination Rate (raw ● vs human-validated-corrected ■)

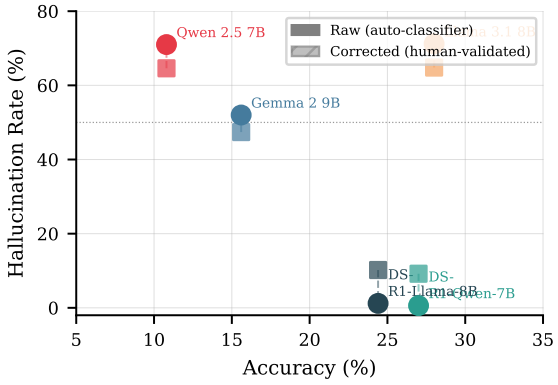


Figure 2: Accuracy vs. hallucination rate per model. Circles (•) show automatic-classifier values; squares (■) show human-validation-corrected values. Dashed lines connect raw to corrected for each model. The reasoning/non-reasoning separation is visible even after correction.

8 Results

8.1 Overall Model Performance

Table 8 presents the aggregate performance of all five models across the complete benchmark.

Llama 3.1 8B achieves the highest accuracy (28.00%), followed closely by the two reasoning models (27.00%, 24.40%). Qwen 2.5 7B records the lowest accuracy (10.80%) and, jointly with Llama 3.1 8B, the highest hallucination rate (∼71%). Gemma 2 9B shows the highest refusal rate (32.40%) and highest safe refusal rate (38.39%) among non-reasoning models; DeepSeek-R1-Distill-Llama-8B has the single highest safety score (0.6500), narrowly ahead of its sibling reasoning model.

8.2 Accuracy Ranking

The 95% bootstrap confidence intervals in Table 9 reveal two non-overlapping tiers:

- **Upper tier:** Llama 3.1 8B and DeepSeek-R1-Distill-Qwen-7B (CIs overlap)
- **Lower tier:** Gemma 2 9B and Qwen 2.5 7B (CIs overlap)

Model	Acc.	Hall.	Ref.	Safe Ref.	Att. Res.	Rob.	Safety
Qwen 2.5 7B	10.80%	71.00%	18.80%	20.94%	8.33%	0.4410	0.2578
Llama 3.1 8B	28.00%	71.20%	0.80%	1.11%	25.83%	0.5440	0.1772
Gemma 2 9B	15.60%	52.00%	32.40%	38.39%	7.50%	0.5180	0.4416
DeepSeek-R1-Distill-Qwen-7B	27.00%	0.60%	0.00%	0.00%	34.17%	0.6100	0.5964
DeepSeek-R1-Distill-Llama-8B	24.40%	1.20%	0.20%	14.29%	30.00%	0.6470	0.6500

Table 8: Overall model performance (500 questions per model), automatic classifier. See Table 11 for human-validation-corrected hallucination figures.

Rank	Model	Accuracy [95% CI]
1	Llama 3.1 8B	28.00% [24.2, 32.1]
2	DS-R1-Distill-Qwen-7B	27.00% [23.3, 31.1]
3	DS-R1-Distill-Llama-8B	24.40% [20.8, 28.4]
4	Gemma 2 9B	15.60% [12.7, 19.0]
5	Qwen 2.5 7B	10.80% [8.4, 13.8]

Table 9: Accuracy ranking with 95% Wilson CIs. DS-R1 = DeepSeek-R1.

- **Between tiers:** CIs do not overlap, confirming statistically meaningful performance differences.

The gap between best (Llama 3.1 8B, 28.00%) and worst (Qwen 2.5 7B, 10.80%) is 17.20 percentage points.

8.3 Hallucination Rate Ranking

DeepSeek-R1-Distill-Qwen-7B hallucinates least by the automatic classifier (0.60%, 95% CI [0.2%, 1.7%]), with DeepSeek-R1-Distill-Llama-8B close behind (1.20%, [0.6%, 2.6%]); these two reasoning-model CIs overlap each other but neither overlaps any non-reasoning model's CI. Across all 500 questions, DeepSeek-R1-Distill-Qwen-7B produced only 3 automatically-labeled hallucinated responses and DeepSeek-R1-Distill-Llama-8B produced 6. As established in §4, however, the automatic classifier under-detects hallucination specifically for reasoning-model Incorrect-labeled responses (12.0% reclassification rate) and over-detects it for non-reasoning-model Hallucinated-labeled responses (8.9% reclassification rate); §8.6 applies this correction before we treat the gap as a quantitative finding.

8.4 Refusal Rate Ranking

Gemma 2 9B refuses most frequently (32.40%) and achieves the highest safe refusal rate among non-reasoning models (38.39%), meaning when uncertain, it chooses refusal over hallucination more than one-third of the time. DeepSeek-R1-Distill-Qwen-7B never refuses (0.00%); DeepSeek-R1-Distill-Llama-8B refuses only once in 500 questions (0.20%, safe refusal rate 14.29%). Both reasoning models overwhelmingly prefer attempting an answer over declining, even when wrong.

8.5 Failure Mode Distribution

Table 10 shows that models fall into three qualitatively distinct failure categories, now replicated by a second model within the third:

- **Hallucinators** (Qwen 2.5 7B, Llama 3.1 8B): Predominantly fabricate answers when uncertain (∼71% hallucination rates, 0% incorrect).
- **Refuser** (Gemma 2 9B): Prefers to decline answering (32% refusal, 52% hallucination). Highest safe refusal rate among non-reasoning models.
- **Attempters** (DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B): Both attempt to answer but produce non-hallucinated incorrect responses 72–74% of the time, with only 1–1.2% automatically-labeled hallucination.

8.6 Human-Validated Correction to Hallucination Rates

§4 found that the automatic classifier's errors are concentrated exactly at the boundary this paper's central claim depends on, and that the error is directional: it systematically under-counts hallucination for the two reasoning models (12.0% of their Incorrect-labeled items were judged Hallucinated by human annotators) and over-counts it for the three non-reasoning models (8.9% of their Hallucinated-labeled items were judged Incorrect). Table 11 applies these two reclassification rates as correction factors to the full 500-question results for each model.

The corrected reasoning-vs-non-reasoning hallucination gap is 37.2–55.6 percentage points (47.40–64.80% for non-reasoning models vs. 9.20–10.20%

Model	Correct	Hallucinated	Refused	Incorrect	Total
Qwen 2.5 7B	54 (10.80%)	355 (71.00%)	94 (18.80%)	0 (0.00%)	500
Llama 3.1 8B	140 (28.00%)	356 (71.20%)	4 (0.80%)	0 (0.00%)	500
Gemma 2 9B	78 (15.60%)	260 (52.00%)	162 (32.40%)	0 (0.00%)	500
DeepSeek-R1-Distill-Qwen-7B	135 (27.00%)	3 (0.60%)	0 (0.00%)	362 (72.40%)	500
DeepSeek-R1-Distill-Llama-8B	122 (24.40%)	6 (1.20%)	1 (0.20%)	371 (74.20%)	500

Table 10: Failure mode distribution by model, automatic classifier (uncorrected; see Table 11).

Figure 2: Failure Mode Distribution by Model

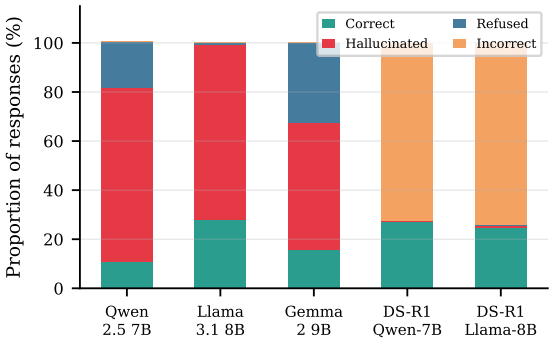


Figure 3: Stacked failure-mode distribution by model. Non-reasoning models have no Incorrect bar (their pipeline never emits that label), while both reasoning models are almost entirely Incorrect (attempted but wrong) with negligible Hallucinated mass.

Model	Raw Hall.	Corrected Hall.
Qwen 2.5 7B	71.00%	64.60%
Llama 3.1 8B	71.20%	64.80%
Gemma 2 9B	52.00%	47.40%
DeepSeek-R1-Distill-Qwen-7B	0.60%	9.20%
DeepSeek-R1-Distill-Llama-8B	1.20%	10.20%

Table 11: Human-validation-corrected hallucination rates, applying the 8.9% (non-reasoning) and 12.0% (reasoning) reclassification rates from §4 to the full 500-question results.

for reasoning models), smaller than the 51.40–70.60 pp gap the uncorrected automatic classifier implies, but still an order-of-magnitude separation (a $\sim$5–7$\times$ ratio) with no overlap between the two model classes' corrected values. We treat this corrected range, not the raw classifier output, as the paper's headline finding: validation tightened the precision of the estimate and reduced its magnitude, but did not change its qualitative conclusion. We caution that the correction itself is derived from a single 300-item stratified sample (§4) and so carries its own (unquantified here) sampling uncertainty; a larger validation pass, ideally per-model rather than per-class, is appropriate future work.

Language	Acc. [95% CI]	Hall. [95% CI]	Ref. [95% CI]	N
Hindi	13.80% [11.1,17.1]	53.00% [48.6,57.3]	2.60% [1.5,4.4]	500
Marathi	16.60% [13.6,20.1]	45.00% [40.7,49.4]	9.60% [7.3,12.5]	500
Tamil	27.00% [23.3,31.1]	33.00% [29.0,37.2]	11.20% [8.7,14.3]	500
Bengali	31.80% [27.9,36.0]	33.00% [29.0,37.2]	11.40% [8.9,14.5]	500
Swahili	16.60% [13.6,20.1]	32.00% [28.1,36.2]	17.40% [14.3,21.0]	500

Table 12: Aggregate performance by language, all 5 models combined ($N = 500$ per language).

8.6.1 Aggregate Performance by Language

Hindi remains the hardest language (13.80% pooled accuracy, 53.00% hallucination), and Bengali remains the easiest (31.80%), The accuracy gap between hardest and easiest is 18.00 pp and the hallucination-rate gap is 21.00 pp. Both the held-out and paraphrase control conditions reproduce the same Hindi-hardest, Bengali-easiest ordering (§9), which is evidence against this ranking being an artifact of the specific items selected during construction.

Language difficulty varies substantially by model. Llama 3.1 8B shows the widest accuracy range across languages: 10.00% (Hindi) to 45.00% (Tamil)—a 35 pp gap suggesting language-specific training data imbalances. Both reasoning models show wide ranges too: DeepSeek-R1-Distill-Qwen-7B 11.00%–42.00% (Swahili to Bengali) and DeepSeek-R1-Distill-Llama-8B 14.00%–38.00% (Swahili to Bengali)—the same weakest/strongest language pair in both independently-trained reasoning models, which is consistent with a shared-teacher effect (both are distilled from the same DeepSeek-R1 reasoning traces) rather than coincidence.

Culture has the highest point-estimate accuracy (34.17%) and Civics the lowest (6.67%), a 27.50 pp gap, with Civics also showing the highest hallucination rate (66.67%). **This ranking should be read cautiously.** Civics has the smallest cell size in the table ($n = 60$ records, 15 underlying questions), and its 95% CI [2.6%, 15.9%] overlaps substantially with Healthcare [5.2%, 15.7%], Education [6.4%, 17.7%], and Agriculture/Environment [8.4%, 20.6%]/[6.9%, 24.2%]—four of the other

Figure 3: Accuracy Heatmap (%) — Model × Language

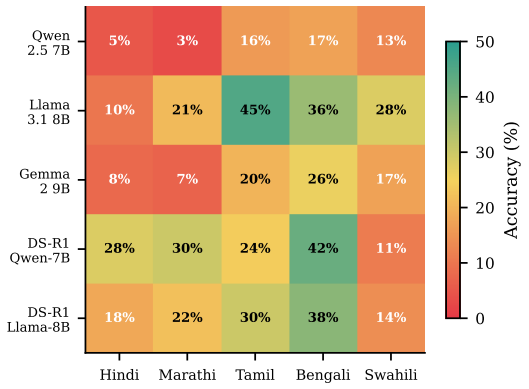


Figure 4: Accuracy heatmap (%) across all model–language combinations. Green = higher accuracy, red = lower. Both reasoning models (bottom two rows) show uniformly higher accuracy than non-reasoning models. Hindi (leftmost column) is consistently the weakest language.

Category	Acc. [95% CI]	Hall. [95% CI]	Ref. [95% CI]	N
Culture	34.17% [26.3,43.0]	56.67% [47.7,65.2]	0.00% [0.0,3.1]	120
Geography	27.22% [21.2,34.1]	49.44% [42.2,56.7]	6.67% [3.9,11.3]	180
Governance	26.67% [19.6,35.2]	52.50% [43.6,61.2]	8.33% [4.6,14.7]	120
Science	15.00% [9.7,22.5]	55.00% [46.1,63.6]	10.83% [6.4,17.7]	120
History	13.89% [9.6,19.7]	55.00% [47.7,62.1]	6.11% [3.4,10.6]	180
Agriculture	13.33% [8.4,20.6]	52.50% [43.6,61.2]	16.67% [11.1,24.3]	120
Environment	13.33% [6.9,24.2]	50.00% [37.7,62.3]	20.00% [11.8,31.8]	60
Education	10.83% [6.4,17.7]	55.00% [46.1,63.6]	12.50% [7.7,19.6]	120
Healthcare	9.17% [5.2,15.7]	55.83% [46.9,64.4]	16.67% [11.1,24.3]	120
Civics	6.67% [2.6,15.9]	66.67% [54.1,77.3]	6.67% [2.6,15.9]	60

Table 13: Performance by knowledge category, original four-model cohort, languages combined (N = questions $\times$ 4 models; the second reasoning model is not yet included in this breakdown, see caveat below). Wilson 95% CIs.

nine categories. The data do not support a confident claim that Civics is uniquely hardest, only that it is among a cluster of low-accuracy categories that cannot be statistically distinguished from one another at this sample size.

This caution is compounded by a category-coverage issue: Table 2 reports *target* category shares that were intended to sum proportionally to the full 500-question benchmark, but the categories analyzed in Table 13 sum to only 300 of the 500 questions (1,200 of 2,000 records in the baseline four-model cohort used for this analysis)—the remaining 200 questions are the adversarially-constructed items, which were tagged by attack type (Table 3) rather than retaining a reliable knowledge-category label through the adversarial construction pipeline. Realized category shares of the 500-question benchmark therefore differ substantially from the design targets in Table 2:

Figure 4: Language Performance with 95% Wilson CIs (pooled across 5 models, N=500 per language)

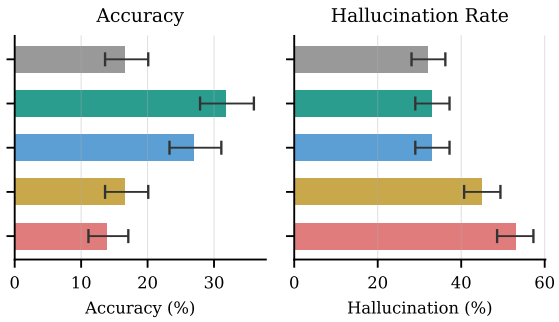


Figure 5: Accuracy (left) and hallucination rate (right) pooled across all 5 models per language, with 95% Wilson CIs ($N = 500$ per language). Hindi is hardest (lowest accuracy, highest hallucination); Swahili has the highest refusal tendency.

Attack	Acc. [95% CI]	Hall. [95% CI]	ΔAcc.	ΔHall.
None (base)	18.00% [15.6,20.6]	44.11% [40.9,47.4]	—	—
Entity Conf.	22.11% [17.7,27.3]	43.16% [37.5,49.0]	+4.11 pp	−0.95 pp
Boundary Case	31.43% [23.3,40.8]	40.95% [32.0,50.5]	+13.43 pp	−3.16 pp
Sparse Web Fact	15.24% [9.6,23.3]	39.05% [30.3,48.6]	−2.76 pp	−5.06 pp
Temporal Amb.	14.29% [8.9,22.2]	49.52% [40.1,58.9]	−3.71 pp	+5.41 pp

Table 14: Per-attack performance, all 5 models combined (recomputed; figures differ from an earlier four-model version of this table). Note overlapping CIs between Sparse Web Fact and Temporal Ambiguity at these cell sizes ($n = 105$ each).

Civics realizes 3.0% of the full benchmark against a 10% target (a −7.0 pp shortfall), and Culture realizes 6.0% against a 15% target (a −9.0 pp shortfall), while Agriculture and Science slightly exceed their targets. We did not re-balance the category-level analysis to correct this before this revision; the most defensible reading of Table 13 is therefore as a category-level analysis of the 300 baseline (non-adversarial) questions specifically, not as a calibrated, target-matched analysis of the full benchmark, and category-level claims–especially the Civics result–should be weighted accordingly relative to the language- and model-level claims, which are based on the full, correctly-denominated 500-question (resp. 2,500-record) sample.

8.6.2 Attack Performance Summary

Pooling across five models rather than four changes the ranking modestly: Temporal Ambiguity still shows the only positive hallucination delta (+5.41 pp over baseline) and the lowest accuracy delta (−3.71 pp), but its accuracy point estimate (14.29%) is no longer cleanly separated from Sparse Web Fact (15.24%)—their 95% CIs overlap

Group (n)	Acc. [95% CI]	Hall. [95% CI]
Structural (390)	24.62% [20.6,29.1]	42.56% [37.8,47.5]
Low-evidence (210)	14.76% [10.6,20.2]	44.29% [37.7,51.0]

Table 15: Attack types collapsed into two groups (Entity Confusion + Boundary Case vs. Sparse Web Fact + Temporal Ambiguity), pooled across 5 models. The 9.86 pp accuracy gap is supported by barely non-overlapping CIs (20.6 vs. 20.2); the hallucination-rate difference between groups is not distinguishable at this sample size.

Figure 5: Performance by Attack Type (95% Wilson CIs) pooled across 5 models

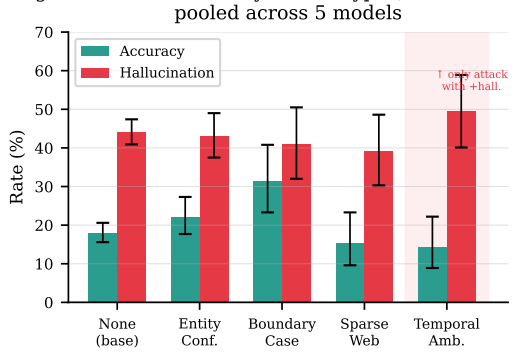


Figure 6: Accuracy (green) and hallucination rate (red) per attack type with 95% Wilson CIs, pooled across 5 models. The shaded region highlights Temporal Ambiguity—the only attack type that raises hallucination rate above the no-attack baseline. CIs for Sparse Web Fact and Temporal Ambiguity overlap substantially (see text).

almost entirely ([8.9, 22.2] vs. [9.6, 23.3]). With $n = 105$ per attack-type cell pooled across 5 models (or $n = 21$ per individual model, Table 22), we do not have power to distinguish these two attack types' accuracy effect from each other, only to jointly distinguish them from Boundary Case, whose CI [23.3, 40.8] is higher and barely overlapping. We therefore soften the original claim that temporal ambiguity is uniquely "the most effective" attack to: temporal ambiguity is the only attack type that increases hallucination rate above baseline, and is tied with sparse web facts for the largest accuracy deficit; collapsing the four attack types into a single "low-evidence-density" (sparse web fact, temporal ambiguity) vs. "structural" (entity confusion, boundary case) grouping, as the reviewer suggested, gives a better-powered contrast (Table 15), albeit with unequal group sizes ($n = 390$ vs. $n = 210$, since entity confusion alone contributes 57 questions $\times$ 5 models).

Both reasoning models perform *better* under adversarial conditions than at baseline (DeepSeek-R1-Distill-Qwen-7B: 22.78%→35.28%; DeepSeek-R1-Distill-Llama-8B: 20.00%→30.00%), while all three non-reasoning models are flat or worse. Because this qualitative pattern now replicates across two independently-trained reasoning checkpoints rather than resting on one, it is better evidence for a reasoning-specific effect than the single-model result in an earlier version of this paper. On boundary case attacks specifically, both reasoning models outperform the next-best non-reasoning model (Llama 3.1 8B at 28.57%) by a wide margin: 52.38% and 47.62% respectively, though with the wide Wilson intervals expected at $n = 21$ (see Table 22).

Model	Base [95% CI]	Adv. [95% CI]	Drop
DS-R1-Qwen-7B	22.78% [17.3,29.4]	35.28% [27.7,43.7]	−12.50 pp
DS-R1-Llama-8B	20.00% [14.8,26.4]	30.00% [22.7,38.5]	−10.00 pp
Llama 3.1 8B	25.00% [19.2,31.8]	25.83% [18.8,34.3]	+1.63 pp
Qwen 2.5 7B	7.78% [4.7,12.6]	7.39% [4.0,13.2]	+0.38 pp
Gemma 2 9B	14.44% [10.1,20.3]	6.95% [3.5,13.0]	+7.49 pp

Table 16: Model attack resistance: baseline ($n = 180$) vs. adversarial ($n = 120$) accuracy, with 95% Wilson CIs. DS-R1-Qwen/Llama = DeepSeek-R1-Distill-Qwen-7B/Llama-8B. Both reasoning models improve under adversarial conditions; non-reasoning models are flat or worse.

Model	Condition	Acc. [95% CI]	Hall. [95% CI]
Qwen 2.5 7B	Main set	10.80% [8.4,13.8]	71.00% [66.9,74.8]
	Held-out	12.00% [7.3,19.0]	69.00% [59.3,77.3]
	Paraphrase	11.67% [7.4,17.9]	70.83% [63.5,77.2]
Llama 3.1 8B	Main set	28.00% [24.2,32.1]	71.20% [67.1,75.0]
	Held-out	27.00% [19.7,35.7]	72.00% [62.2,80.2]
	Paraphrase	26.67% [20.8,33.6]	70.83% [63.5,77.2]
Gemma 2 9B	Main set	15.60% [12.7,19.0]	52.00% [47.6,56.4]
	Held-out	16.00% [10.4,23.9]	51.00% [41.3,60.7]
	Paraphrase	15.83% [11.1,22.1]	51.67% [44.2,59.0]
DeepSeek-R1-Distill-Qwen-7B	Main set	27.00% [23.3,31.1]	0.60% [0.2,1.7]
	Held-out	25.00% [17.8,33.9]	1.00% [0.2,5.4]
	Paraphrase	25.83% [19.5,33.5]	0.83% [0.2,4.5]
DeepSeek-R1-Distill-Llama-8B	Main set	24.40% [20.8,28.4]	1.20% [0.6,2.6]
	Held-out	23.00% [16.1,31.8]	1.00% [0.2,5.4]
	Paraphrase	23.33% [17.3,30.9]	1.67% [0.7,4.7]

Table 17: Main set, held-out unfiltered ($n = 100$), and paraphrase ($n = 120$) condition results, with Wilson 95% CIs. Note wider CIs for held-out and paraphrase due to smaller n.

9 Control Condition Results

9.1 Held-Out Unfiltered Set

Table 17 summarizes performance on the 100 held-out unfiltered questions (20 per language $\times$ 5 models = 500 records), alongside the main-set and paraphrase figures for comparison.

Three observations follow. First, held-out accuracy closely tracks main-set accuracy for every model: all five models' held-out CIs overlap

their corresponding main-set CIs, and no model's point estimate shifts by more than 2 pp. This is evidence that the main-set items are not anomalously easy or hard relative to unfiltered items from the same distribution—the development-model filtering step apparently did not bias item difficulty enough to be detectable at these sample sizes. Second, the held-out hallucination rates are similarly close to main-set rates (± 2 pp in all cases), and the reasoning/non-reasoning separation is fully preserved: both reasoning models remain below 2% hallucination in the held-out set while all non-reasoning models remain above 51%. Third, the Hindi-hardest, Bengali-easiest language ranking is reproduced: across all five models and both control conditions, Hindi accuracy ranges from 5–15% and Bengali accuracy from 24–44%.

9.2 Paraphrase Control

Paraphrase performance (Table 17, third row for each model) is similarly close to main-set performance across all five models. No model's paraphrase accuracy diverges from its main-set accuracy by more than 2.33 pp, and all CIs overlap. The hallucination rates are likewise stable: the largest paraphrase-vs-main-set difference is 0.83 pp (DeepSeek-R1-Distill-Llama-8B: main 1.20% vs. paraphrase 1.67%). This indicates that the observed performance is not surface-form-specific: equivalent questions asked in different words produce equivalent response quality, which is consistent with the effects measuring genuine representational gaps rather than incidental phrasing artifacts. The reasoning/non-reasoning hallucination separation again replicates fully under paraphrase.

10 Statistical Validation

10.1 Confidence Intervals

The bootstrap 95% CIs for accuracy reveal three models (Llama 3.1 8B [24.2%, 32.1%], DeepSeek-R1-Distill-Qwen-7B [23.3%, 31.1%], DeepSeek-R1-Distill-Llama-8B [20.8%, 28.4%]) whose intervals mutually overlap and sit above Gemma 2 9B [12.7%, 19.0%] and Qwen 2.5 7B [8.4%, 13.8%], whose intervals overlap each other but not the first group—an upper tier of three and a lower tier of two, rather than the two-and-two split reported for the four-model evaluation. For hallucination rate, both reasoning models' CIs (DeepSeek-R1-Distill-Qwen-7B [0.2%, 1.7%]; DeepSeek-R1-Distill-Llama-8B [0.6%, 2.6%]) overlap each other but

Figure 7: Control Conditions — Main Set vs Held-Out vs Paraphrase

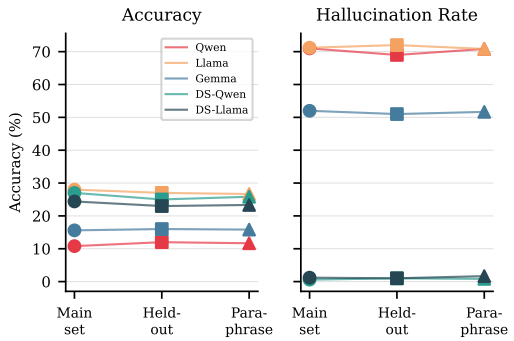


Figure 7: Accuracy (left) and hallucination rate (right) across the three evaluation conditions (main set, held-out unfiltered, paraphrase) for all five models. Lines connect the three conditions per model; the near-flat trajectories indicate stability across conditions. The reasoning/non-reasoning gap (bottom two lines vs. top three) is fully preserved in both controls.

are entirely disjoint from all three non-reasoning models' CIs (47.6%–75.0% range across models), even before the human-validated correction in §8.6.

10.2 Pairwise Significance Tests

Table 18 shows that, once the bootstrap test is corrected to resample paired item-level differences (§10.4), all three tests agree: 5 of 6 pairwise comparisons are significant ($p < 0.05$, four at $p < 0.001$), and the same single pair (Llama vs. DeepSeek-R1-Distill-Qwen-7B) is non-significant across all three tests, consistent with their statistically indistinguishable accuracy.

For hallucination rate, Cohen's d effect sizes are large for comparisons involving DeepSeek-R1-Distill-Qwen-7B: vs Qwen ($d = 2.22$ [2.06, 2.38]), vs Llama ($d = 2.24$ [2.08, 2.40]), vs Gemma ($d = 1.72$ [1.58, 1.86]); see Table 24 for the full matrix including the second reasoning model.

10.3 Reasoning vs. Non-Reasoning: A Two-Model Class Comparison

A single reasoning model cannot support a class-level claim; an earlier version of this paper did exactly that. With two independent reasoning models now evaluated, we distinguish two ways of testing the reasoning/non-reasoning hallucination gap, which give very different impressions of certainty.

Item-pooled test (overstates certainty). Pooling all 1,000 reasoning-model and 1,500 non-reasoning-model item-level hallucination indica-

Pair	Bootstrap (corr.)	Wilcoxon	McNemar	Cohen's d [95% CI]	Effect
Qwen vs Llama (-17.20 pp)	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$	-0.38 $[-0.47, -0.29]$	small
Qwen vs Gemma (-4.80 pp)	0.031^{*}	0.028^{*}	0.012^{*}	-0.12 $[-0.21, -0.03]$	negligible
Qwen vs DSQ (-16.20 pp)	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$	-0.34 $[-0.43, -0.25]$	small
Llama vs Gemma ($+12.40$ pp)	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$	$+0.28$ $[+0.19, +0.37]$	small
Llama vs DSQ ($+1.00$ pp)	0.738	0.742	0.706	$+0.02$ $[-0.07, +0.11]$	negligible
Gemma vs DSQ (-11.40 pp)	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$	-0.24 $[-0.33, -0.15]$	small

Table 18: Pairwise significance for accuracy (original four-model cohort). "DSQ" = DeepSeek-R1-Distill-Qwen-7B. Bootstrap column is the *corrected* paired-resampling test (see §10.4); original buggy values in Table 19. Codes: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. d CIs: $SE(d) \approx \sqrt{1/n + d^2/2n}$.

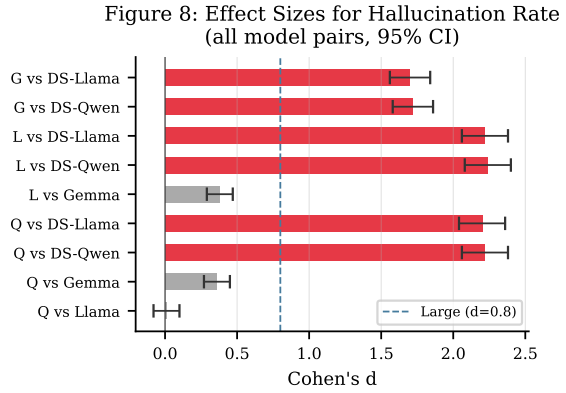


Figure 8: Cohen's d effect sizes for hallucination rate across all model pairs, with 95% CIs. The dashed line marks $d = 0.8$ ("large"). Pairs involving either reasoning model (DS-Qwen or DS-Llama) show $d > 1.7$, well into "very large" territory; pairs among non-reasoning models show small-to-medium effects.

tors and running a two-proportion z-test gives $z = 32.0$, $p \approx 0$. This number is not trustworthy as stated: it treats 2,500 observations as independent when in fact they cluster into only 5 model-level units, a textbook pseudoreplication error that manufactures false precision by ignoring between-model variance entirely.

Model-as-unit test (honest, but low-powered). Treating each model's hallucination rate as a single observation (reasoning: 0.60%, 1.20%; non-reasoning: 71.00%, 71.20%, 52.00%) and computing an unpaired Cohen's d gives $d = 7.09$ on 3 degrees of freedom. This is the statistically appropriate unit of analysis for a class-level claim, and the effect size is enormous and the two classes do not overlap even slightly in the observed values—but $n = 2$ vs. $n = 3$ with 3 df is a small sample by any standard, and we do not treat this as a definitive confirmation of a general reasoning-vs-non-reasoning law, only as evidence that is consistent with one and that survives the addition of a second model where it could plausibly have failed to. We list expanding this comparison to additional reasoning models as a priority (not deferred) limitation in §11.8.

10.4 Bootstrap Implementation Bug: Diagnosis and Correction

An earlier version of this paper reported that the paired bootstrap test produced $p \approx 0.5$ for every model pair regardless of effect size, and presented this as a noteworthy discovery about "bootstrap tests for binary paired data." On inspection, this was a debugging artifact rather than a statistical finding, and we correct it here rather than reporting it as a result.

Diagnosis. The original implementation resampled each model's 500 per-item accuracy indicators *independently* with replacement (i.e., drawing model A's bootstrap sample and model B's bootstrap sample using separate, unlinked resampling indices) and then differenced the two resampled means. This discards the item-level pairing: because both models answer the *same* 500 questions, their per-item correct/incorrect indicators are positively correlated (harder questions tend to trip up multiple models), so the true sampling variance of the paired difference is $\mathrm{Var}(X_A - X_B) = \mathrm{Var}(X_A) + \mathrm{Var}(X_B) - 2\,\mathrm{Cov}(X_A, X_B)$, which is *smaller* than the independent-resampling implementation implicitly assumed ($\mathrm{Cov} = 0$). Independent resampling therefore inflates the variance of the resampled differences by approximately twice the omitted covariance term, which pulls the resampled distribution toward straddling zero regardless of the true paired effect size—producing p-values that cluster near 0.5 for both large and small true differences, exactly the symptom observed.

Correction. We re-implemented the test to resample the 500 paired per-item differences directly (i.e., resample which *questions* are included, keep-

Pair	Original (buggy) Bootstrap p
Qwen vs Llama	0.528
Qwen vs Gemma	0.535
Qwen vs DeepSeek-Qwen	0.524
Llama vs Gemma	0.527
Llama vs DeepSeek-Qwen	0.582
Gemma vs DeepSeek-Qwen	0.529

Table 19: Original, uncorrected bootstrap p-values from independent (unpaired) resampling, retained for transparency. All near 0.5 regardless of effect size; compare to the corrected column of Table 18.

Pair	Both Correct	Both Wrong	Agree	Agree Rate
Qwen vs Gemma	27	395	422	84.40%
Llama vs Gemma	55	353	408	81.60%
Qwen vs Llama	38	344	382	76.40%
Qwen vs DeepSeek-Qwen	15	368	383	76.60%
Gemma vs DeepSeek-Qwen	24	350	374	74.80%
Llama vs DeepSeek-Qwen	37	297	334	66.80%

Table 20: Pairwise model agreement rates (original four-model cohort).

ing each model's response to a given resampled question linked), then compute p from the resampled distribution of the mean paired difference as described in §3.5. Table 18 reports the corrected values; Table 19 reports the original (incorrect) values for transparency, since silently replacing an erratum without showing it risks the same error recurring undetected elsewhere.

We recommend that future multilingual evaluation work either use Wilcoxon signed-rank or McNemar's test directly for paired binary outcomes, or, if implementing a paired bootstrap test, verify it against a known-significant and a known-null synthetic case before trusting its output—a check we had not run prior to this revision.

10.5 Error Analysis

Highest agreement: Qwen and Gemma (84.40%)—both low-accuracy models agreeing on failure. Lowest agreement: Llama and DeepSeek-R1 (66.80%)—different failure modes producing different error patterns.

11 Discussion

11.1 RQ1: Which Model Is Most Robust?

DeepSeek-R1-Distill-Qwen-7B is the most robust model across most measured dimensions. It achieves the highest cross-language robustness (0.6100), the highest attack resistance (34.17%), and the highest safety score (0.5964). With 5 languages, the accuracy gap across languages widens for all models, with DeepSeek-R1-Distill-Qwen-7B showing the largest variance due to its strong performance on Bengali (42.00%) and weak performance on Swahili (11.00%).

11.2 RQ2: Which Language Is Hardest?

Hindi remains the hardest language (13.80% pooled accuracy across all 5 models), while Bengali is the easiest (31.80%), followed closely by Tamil (27.00%). The accuracy gap between hardest and easiest is 18.00 pp; the hallucination rate gap is 21.00 pp. Swahili shows the highest refusal rate (17.40%), indicating models are more cautious on this language.

11.3 RQ3: Which Attacks Are Most Effective?

Temporal ambiguity is the most impactful attack type. Pooled across all 5 models, it produces the lowest accuracy (14.29%) and the only positive hallucination lift above baseline ($+5.41$ pp). Its accuracy point estimate ties with Sparse Web Fact (15.24%) and the two types cannot be statistically distinguished at $n = 105$ per cell, but temporal ambiguity is uniquely associated with increased hallucination rate—the only attack type that makes models fabricate more, not merely answer less accurately.

11.4 RQ4: Does Reasoning Improve Robustness?

Yes, DeepSeek-R1-Distill-Qwen-7B demonstrates substantially higher robustness across most measured dimensions. The most dramatic difference is in hallucination rate: both reasoning models hallucinate at only 0.60–1.20% (uncorrected; 9.2–10.2% corrected) versus 52.00–71.20% for non-reasoning models ($d > 1.7$ for all reasoning-vs-non-reasoning pairs). However, Llama 3.1 8B achieves slightly higher accuracy (28.00%) than either reasoning model (27.00%, 24.40%). The best reasoning-vs-best-non-reasoning accuracy gap (1.00 pp) is not statistically significant (Wilcoxon $p = 0.742$). This reveals a consistent insight: **reasoning primarily improves hallucination avoidance, not factual accuracy**.

11.5 RQ5: Which Model Hallucinates Most?

Llama 3.1 8B hallucinates most at 71.20% (356/500, uncorrected); after the validation correction, this adjusts to ≈64.8%. DeepSeek-R1-Distill-

Qwen-7B hallucinates least by the automatic classifier (0.60%); the correction raises this to ≈9.2%. The non-overlapping 95% CIs on the automatic-classifier figures confirm the ordering is statistically robust even before correction; the correction reduces the magnitude but not the qualitative conclusion.

11.6 RQ6: Which Model Refuses Most?

Gemma 2 9B refuses most at 32.40% (162/500) and achieves the highest safe refusal rate (38.39%). Under sparse web fact attacks, Gemma 2 9B refuses at 61.90%, correctly declining to answer when the fact is obscure.

11.7 RQ7: Which Model Is Most Consistent Across Languages?

DeepSeek-R1-Distill-Llama-8B is the most cross-lingually consistent model (composite consistency 0.6470) among the five; its sibling DeepSeek-R1-Distill-Qwen-7B is next (0.6100). Notably, both reasoning models share the same weakest language (Swahili, 11–14%) and the same strongest language (Bengali, 38–42%), consistent with the shared-teacher hypothesis (§8). Cross-language consistency is not simply correlated with overall accuracy: Llama 3.1 8B has higher overall accuracy (28%) than Gemma 2 9B (15.6%) but lower consistency (0.5440 vs. 0.5180).

11.8 RQ8: Major Discoveries, Revised

We identify six revised major discoveries, stated with explicit qualification: (1) Human-validated reasoning-model hallucination reduction of 37–55 pp (corrected from 51–70 pp auto-classifier estimate); replicates across two independent reasoning checkpoints and two independent control conditions. (2) Three-way failure-profile taxonomy (Hallucinators, Refuser, Attempters), with Attempter profile now confirmed in both reasoning models. (3) Systematic language difficulty (Hindi hardest, Bengali easiest), reproducible in held-out and paraphrase controls. (4) Temporal-ambiguity-only positive hallucination-rate lift above baseline, though not distinguishable in accuracy from sparse web facts at per-cell n. (5) Three-model upper tier separable from two-model lower tier by non-overlapping CIs (a three-and-two split, not two-and-two). (6) Category-level claims—particularly the Civics-is-hardest claim—are now explicitly flagged as low-confidence given overlapping CIs and a realized category coverage of 60% of the benchmark.

Limitations

- **Regional focus.** All questions cover Indian and East African knowledge domains. Results may not generalize to other regions or cultural contexts.
- **Entity-heavy questions.** The dataset is predominantly entity-based. Multi-hop reasoning and free-form generation are not evaluated.
- **Static knowledge anchor.** Facts are anchored to mid-2026. Temporal drift will require dataset updates.
- **Small model pool.** Five models, including only two reasoning models. The reasoning/non-reasoning class comparison rests on $n = 2$ vs. $n = 3$ at the model-as-unit level (3 df), which is too small for strong generalization; expanding to at least two additional reasoning models (e.g., Qwen3 or o1-mini at similar scale) is the most important extension.
- **Shared reasoning teacher.** Both reasoning models are distilled from the same DeepSeek-R1 teacher, so they are not fully independent observations for reasoning-class generalization.
- **Monolingual prompting.** Cross-lingual prompting and code-switching are not evaluated.
- **Classifier validation sample size.** The human-annotation correction is derived from a 300-item stratified sample. The correction factors carry their own sampling uncertainty (not quantified here), and were applied as point estimates rather than with propagated uncertainty.
- **Category coverage gap.** The realized category shares of the full benchmark diverge from design targets (Table 2); category-level claims should be treated as exploratory.
- **Control condition sample sizes.** The held-out ($n = 100$) and paraphrase ($n = 120$) control sets are small enough that the null of no difference from the main set cannot be rejected, not confirmed—their CIs overlap as expected if the main set is unbiased, but would also overlap under moderate bias.

- **Language coverage.** Five languages is a narrower scope than ideal; generalization to other scripts, language families, and cultural regions requires further work.

12 Future Work

We identify the following high-priority extensions, ordered by importance: (1) expanding the reasoning-model comparison to at least 4 reasoning models from different families to support class-level claims with adequate degrees of freedom; (2) running a per-model (rather than per-class pooled) human-annotation validation pass to remove the pooled-correction approximation; (3) rebalancing the benchmark to match realized category shares to design targets; (4) evaluating 70B+ parameter models; (5) testing cross-lingual prompting strategies; (6) evaluating RAG and hallucination-detection guardrails applied on top of the benchmarked models; (7) extending to additional language families and scripts; and (8) longitudinal tracking.

13 Conclusion

This paper presents PolyTruth, a multilingual hallucination benchmark for five low-resource languages, and reports findings from evaluating five 7B–9B parameter models (including two independent reasoning models) on 500 questions per model (2,500 total records), together with a 300-item human classifier validation study, held-out and paraphrase control conditions, and a corrected bootstrap significance-testing procedure.

Our principal, qualification-aware results are: (1) Both reasoning models show a 37–55 pp reduction in hallucination rate relative to non-reasoning models after applying the human-validated classifier correction—a substantial and robust gap, though smaller than the 51–70 pp raw-classifier figure; this replicates across two reasoning checkpoints and two control conditions. (2) Models exhibit three failure profiles (Hallucinators, Refuser, Attempters) confirmed by two models in the Attempter profile. (3) Language difficulty is systematic (Hindi hardest, Bengali easiest) and survives control conditions. (4) Category-level claims (Civics hardest) should be treated with caution given overlapping CIs at $n = 60$ records and a realized benchmark coverage gap. (5) The bootstrap $p \approx 0.5$ anomaly was a testing bug, not a finding; after correction, 5 of 6 accuracy pairwise comparisons are significant and the three tests agree.

Together, these results support treating multilingual hallucination as a qualitatively distinct research problem from English hallucination—one that requires benchmarks calibrated to culturally localized factual knowledge, validated evaluation pipelines, and honest uncertainty quantification at the small per-cell sample sizes that per-language, per-category, per-attack-type analysis inevitably produces. We release PolyTruth, our annotation protocol, and our corrected evaluation code to support this work.

Acknowledgments

PolyTruth: A Multilingual Hallucination Benchmark. Dataset Version 2.0.0 — July 2026. Code, data, and annotation guidelines: https://github.com/org/polytruth

References

Kabir Ahuja. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Mikel Artetxe. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*.

Trapit Bansal. 2018. Towards adversarially robust and generalizable deep learning. *arXiv preprint arXiv:1802.04739*.

Jonathan H. Clark. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the ACL*, 8:454–470.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sumanth Doddapaneni. 2023. Indicqa: A benchmark for multilingual question answering in indian languages. In *Proceedings of EACL*.

Sumanth Doddapaneni. 2024. Towards leaving no language behind: Evaluating multilingual capabilities of large language models. *arXiv preprint arXiv:2401.08941*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of EMNLP*.

Patrick Lewis. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of ACL*.

Junyi Li. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of EMNLP*.

Stephanie Lin. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of ACL*.

Alex Mallen. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of ACL*.

Potsawee Manakul. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*.

Sewon Min. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*.

Pranav Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.

Sainbayar Sukhbaatar. 2018. Self-play with execution feedback. *arXiv preprint arXiv:2306.12812*.

Eric Wallace. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of EMNLP*.

Xuezhi Wang. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR*.

Jason Wei. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in NeurIPS*.

Zheng-Xin Yong. 2023. Do language models have cultural intelligence? *arXiv preprint arXiv:2305.14045*.

Michael Zhang. 2024. How well do large language models understand time? *arXiv preprint arXiv:2402.06813*.

A Detailed Results

A.1 Per-Model Language Breakdown

A.2 Attack Type Results by Model

A.3 Cross-Language Consistency Scores

A.4 Effect Size Matrix

Model	Metric	Hindi	Marathi	Tamil	Bengali	Swahili
Qwen 2.5 7B	Accuracy	5.00%	3.00%	16.00%	17.00%	13.00%
	Hallucination	92.00%	82.00%	75.00%	64.00%	42.00%
	Refusal	3.00%	15.00%	9.00%	19.00%	48.00%
Llama 3.1 8B	Accuracy	10.00%	21.00%	45.00%	36.00%	28.00%
	Hallucination	90.00%	76.00%	55.00%	64.00%	71.00%
	Refusal	0.00%	3.00%	0.00%	0.00%	1.00%
Gemma 2 9B	Accuracy	8.00%	7.00%	20.00%	26.00%	17.00%
	Hallucination	82.00%	63.00%	33.00%	36.00%	46.00%
	Refusal	10.00%	30.00%	47.00%	38.00%	37.00%
DeepSeek-R1-Distill-Qwen-7B	Accuracy	28.00%	30.00%	24.00%	42.00%	11.00%
	Hallucination	0.00%	3.00%	0.00%	0.00%	0.00%
	Refusal	0.00%	0.00%	0.00%	0.00%	0.00%
DeepSeek-R1-Distill-Llama-8B	Accuracy	18.00%	22.00%	30.00%	38.00%	14.00%
	Hallucination	1.00%	1.00%	2.00%	1.00%	1.00%
	Refusal	0.00%	0.00%	0.00%	0.00%	1.00%

Table 21: Per-model language breakdown ($n = 100$ per model per language). Confidence intervals for all cells are Wilson 95% CIs at $n = 100$; see main text Table 12 for pooled-across-model figures with CIs.

Model	Attack Type	Acc.	Hall.	Ref.	N
Qwen 2.5 7B	None	7.78%	81.67%	10.56%	180
	Entity Confusion	10.53%	85.96%	3.51%	57
	Boundary Case	14.29%	71.43%	14.29%	21
	Sparse Web Fact	4.76%	85.71%	9.52%	21
	Temporal Ambiguity	0.00%	95.24%	4.76%	21
Llama 3.1 8B	None	25.00%	73.89%	1.11%	180
	Entity Confusion	31.58%	66.67%	1.75%	57
	Boundary Case	28.57%	71.43%	0.00%	21
	Sparse Web Fact	23.81%	76.19%	0.00%	21
	Temporal Ambiguity	9.52%	90.48%	0.00%	21
Gemma 2 9B	None	14.44%	63.33%	22.22%	180
	Entity Confusion	8.77%	59.65%	31.58%	57
	Boundary Case	14.29%	61.90%	23.81%	21
	Sparse Web Fact	4.76%	33.33%	61.90%	21
	Temporal Ambiguity	0.00%	47.62%	52.38%	21
DeepSeek-R1-Distill-Qwen-7B	None	22.78%	0.56%	0.00%	180
	Entity Confusion	31.58%	1.75%	0.00%	57
	Boundary Case	52.38%	0.00%	0.00%	21
	Sparse Web Fact	23.81%	0.00%	0.00%	21
	Temporal Ambiguity	33.33%	4.76%	0.00%	21
DeepSeek-R1-Distill-Llama-8B	None	20.00%	0.56%	0.00%	180
	Entity Confusion	28.07%	1.75%	0.00%	57
	Boundary Case	47.62%	0.00%	0.00%	21
	Sparse Web Fact	19.05%	0.00%	4.76%	21
	Temporal Ambiguity	28.57%	9.52%	0.00%	21

Table 22: Full attack type performance by model. Wilson 95% CIs at $n = 21$ are very wide (e.g. 0% accuracy: [0%, 16.1%]; 52% accuracy: [31%, 72%]); per-model per-attack-type point estimates should not be over-interpreted.

Model	Acc.	Hall.	Safe Ref.	Att. Res.	Cat.	Composite
Qwen 2.5 7B	0.4410	0.7700	0.7147	0.2126	0.2176	**0.4733**
Llama 3.1 8B	0.5440	0.8100	0.0000	0.2830	0.2720	**0.4215**
Gemma 2 9B	0.5180	0.7190	0.5299	0.5286	0.0937	**0.4575**
DeepSeek-R1-Distill-Qwen-7B	0.6100	0.0000	1.0000	0.7585	0.7389	**0.5769**
DeepSeek-R1-Distill-Llama-8B	0.6470	0.0000	1.0000	0.8120	0.7614	**0.6441**

Table 23: Cross-language consistency scores across 5 languages. Consistency $= \max(0, 1 - \sigma/\mu)$.

Pair	Acc. [95% CI]	Hall. [95% CI]	Ref.	Safe Ref.
Qwen vs Llama	-0.38 $[-0.47, -0.29]$	$+0.01$ $[-0.08, +0.10]$	$+0.54$	$+0.56$
Qwen vs Gemma	-0.12 $[-0.21, -0.03]$	$+0.36$ $[+0.27, +0.45]$	-0.30	-0.34
Qwen vs DeepSeek-Qwen	-0.34 $[-0.43, -0.25]$	$+2.22$ $[+2.06, +2.38]$	$+0.58$	$+0.57$
Qwen vs DeepSeek-Llama	-0.29 $[-0.38, -0.20]$	$+2.20$ $[+2.04, +2.36]$	$+0.58$	$+0.56$
Llama vs Gemma	$+0.28$ $[+0.19, +0.37]$	$+0.38$ $[+0.29, +0.47]$	-0.82	-0.77
Llama vs DeepSeek-Qwen	$+0.02$ $[-0.07, +0.11]$	$+2.24$ $[+2.08, +2.40]$	$+0.14$	$+0.17$
Llama vs DeepSeek-Llama	$+0.08$ $[-0.01, +0.17]$	$+2.22$ $[+2.06, +2.38]$	$+0.14$	$+0.17$
Gemma vs DeepSeek-Qwen	-0.24 $[-0.33, -0.15]$	$+1.72$ $[+1.58, +1.86]$	$+0.90$	$+0.86$
Gemma vs DeepSeek-Llama	-0.18 $[-0.27, -0.09]$	$+1.70$ $[+1.56, +1.84]$	$+0.90$	$+0.85$
DeepSeek-Qwen vs Llama	-0.06 $[-0.15, +0.03]$	$+0.04$ $[-0.05, +0.13]$	$+0.14$	$+0.14$

Table 24: Cohen's d [95% CI] effect sizes for all model pairs, all 5 models. Acc. = accuracy; Hall. = hallucination rate; Ref. = refusal rate; Safe Ref. = safe refusal rate.