Confidence as a Tie-Breaker: Reassessing Multilingual Hedging Bias in LLM-as-a-Judge Evaluation

Rajashik Datta and Sanjan Baitalik
Institute of Engineering & Management, Kolkata, West Bengal, India
{rajashik.datta2022, sanjan.baitalik2022}@iem.edu.in

Abstract

LLM judges are often used to score generated answers, but their decisions may be affected by surface style rather than semantic correctness. We introduce PolyJudge-Uncertain, a controlled benchmark for studying multilingual hedging effects in LLM-as-a-judge evaluation. The benchmark contains 5,120 short factual QA instances across English, Hindi, Hinglish, and Bengali, balancing assertive versus hedged style and correct versus incorrect answers. A small pilot suggested a large pointwise penalty against hedged answers. After repairing multilingual templates and adding quality-control checks, this pointwise effect largely disappears: final pointwise accuracy is 99.8%, with no meaningful assertive–hedged gap. The robust remaining effect is pairwise: when two answers are equally correct and differ only in style, the judge prefers the assertive answer in 1,276 of 1,280 cases. We interpret this as a protocol- and task-specific assertiveness preference, not as a universal bias against hedging. Our findings highlight benchmark auditing as a central requirement for multilingual judge-bias research.

1 Introduction

Large language models are now routinely used as evaluators for generation quality, instruction following, preference ranking, and dataset annotation because they are inexpensive, flexible, and often correlate reasonably well with human judgments on many tasks (Liu et al., 2023; Zheng et al., 2023; Li et al., 2025a; Bavaresco et al., 2025). At the same time, recent work has shown that LLM judges are not neutral scoring devices. They can exhibit position bias, self-preference bias, authority bias, verbosity bias, formatting bias, and prompt sensitivity, which means that evaluation outcomes may reflect stylistic or contextual artifacts rather than the semantic quality of the candidate answer itself (Chen et al., 2024; Li et al., 2024b; Shi et al., 2025; Xu et al., 2024; Chen et al., 2025; Jeong et al., 2025).

A particularly important case concerns epistemic markers, such as *I think*, *maybe*, *probably*, and analogous expressions in other languages. These markers are often desirable from the perspective of honesty alignment and calibrated uncertainty communication, because they soften unsupported claims and can better signal uncertainty to the user (Geng et al., 2024; Huang et al., 2024; Liu et al., 2025b,a). However, the recent EMBER benchmark showed that LLM judges can penalize epistemically marked answers even when the underlying content remains correct, raising the concern that evaluation pipelines may inadvertently reward overconfident style and punish honest uncertainty expression (Lee et al., 2025).

This issue becomes even more consequential outside English. Multilingual LLM-as-a-judge is already known to be less stable than English-only evaluation, and fairness work has repeatedly shown that non-standard language varieties and dialectal forms can trigger brittle model behavior even when semantics are preserved (Hada et al., 2024; Fu and Liu, 2025; Lin et al., 2025; Pan et al., 2025; Hida et al., 2025). Yet the interaction between hedging, multilinguality, and code-mixing remains underexplored. Real multilingual interaction motivates this question because users routinely hedge, transliterate, and mix registers. At the same time, the present study does not attempt to simulate the full pragmatic complexity of natural conversation. Instead, it uses controlled minimal pairs to isolate one evaluation variable: whether epistemic style changes a judge's decision when semantic content is fixed.

In this paper, we introduce PolyJudge-Uncertain, a controlled multilingual benchmark spanning English, Hindi, Hinglish, and Bengali. Our starting hypothesis followed prior work: correct but hedged answers might receive worse judgments

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 393–402

than equally correct assertive answers. However, our experiments reveal a more nuanced picture. In a small pilot, we indeed observe a large pointwise hedging penalty. After carefully repairing multilingual templates and filtering unnatural constructions, that pointwise effect collapses. What remains is a much sharper and more stable phenomenon: in pairwise comparison, the judge overwhelmingly prefers assertive phrasing over hedged phrasing even when the two answers are equally correct.

This change in empirical story is itself the main contribution. Rather than simply extending English hedging-bias findings to new languages, we show that multilingual judge-bias claims are highly sensitive to benchmark construction quality. Our revised benchmark supports a more defensible conclusion: once template artifacts are controlled, pointwise multilingual judging can be nearly robust to hedging, but pairwise LLM judging still treats assertiveness as a quality signal and uses it as a tie-breaker among semantically equivalent answers.

Our contributions are threefold.

1. We introduce PolyJudge-Uncertain, a compact but carefully controlled multilingual benchmark for studying uncertainty expression in LLM-as-a-judge.

2. We provide an empirical re-assessment of hedging bias under multilingual template repair, showing that a strong pilot effect was largely a construction artifact.

3. We identify a robust remaining failure mode: near-universal pairwise preference for assertive wording among equally correct answers.

2 Related Work

Given our focus on multilingual LLM-judge behavior under hedging, we review three strands of prior work: LLM-as-a-judge and meta-evaluation, known biases in LLM evaluation, and uncertainty expression in multilingual settings.

2.1 LLM-as-a-judge and meta-evaluation

LLM-based evaluation has moved rapidly from a promising idea to a central experimental tool. G-Eval showed that prompted GPT-4 can align well with human judgments on several NLG tasks, helping establish LLM-as-a-judge as a viable alternative to classical metrics (Liu et al., 2023). MT-Bench and Chatbot Arena further popularized pairwise LLM evaluation for open-ended generation and instruction following (Zheng et al., 2023). Subsequent work broadened the framing from isolated evaluators to a general research area, synthesizing evaluator types, prompting strategies, and remaining challenges in robustness and fairness (Li et al., 2024a, 2025a; Bavaresco et al., 2025).

At the same time, meta-evaluation work has increasingly questioned when LLM judges can replace humans and when they cannot. Calderon et al. (2025) argue that replacement should be justified statistically rather than assumed, while Wang et al. (2025) show that richer judgment distributions can improve inference. This paper fits into that meta-evaluation tradition: our concern is not whether the judged answers are good in absolute terms, but whether the *judge itself* behaves robustly when surface style varies and semantics are held fixed.

2.2 Biases in LLM evaluation

A growing body of work has demonstrated that LLM judges are susceptible to diverse biases. Chen et al. (2024) report authority, beauty, misinformation oversight, and gender-related biases in both humans and LLMs, with LLM judges showing substantial vulnerability. Position bias has emerged as another well-studied failure mode, both in direct pairwise ranking and in more general ranking settings (Li et al., 2024b; Shi et al., 2025). Other studies have documented self-preference, auxiliary-context bias, and sensitivity to formatting and verbosity (Xu et al., 2024; Chen et al., 2025; Wu et al., 2025; Li et al., 2025b).

Especially relevant for our work, Jeong et al. (2025) argue that pairwise comparison itself can amplify biased preferences because it encourages judges to rely on salient superficial differences. Our results closely align with that observation, but identify a new variant of the phenomenon centered on epistemic stance: once two answers are semantically tied, assertiveness becomes a near-deterministic winning cue.

2.3 Uncertainty expression and multilingual robustness

Uncertainty estimation and calibrated confidence have become major themes in LLM reliability research (Geng et al., 2024; Huang et al., 2024). Beyond numeric confidence, recent work increasingly studies *natural language uncertainty expression*, where models communicate confidence via epis-

temic markers rather than scalar scores (Liu et al., 2025a,b). This line intersects with longstanding linguistic work on hedging and epistemic stance, where markers of possibility or belief reduce commitment to a proposition rather than changing its propositional content (Lakoff, 1973; Hyland, 2010; Babrow et al., 1998).

The closest prior work to ours is EMBER, which explicitly examines whether epistemic markers distort LLM-based evaluation. Lee et al. (2025) find that LLM judges show a negative bias toward epistemic markers, especially markers of uncertainty, in both single and pairwise evaluation settings. Our paper complements EMBER rather than competing with it. EMBER establishes the phenomenon; we show that in multilingual settings the pointwise effect is much more sensitive to benchmark construction than a first pilot suggests.

Finally, our work relates to multilingual and dialect fairness studies. Prior work has shown that multilingual evaluators require calibration against native-speaker judgments and that low-resource or non-Latin-script settings are especially brittle (Hada et al., 2024; Fu and Liu, 2025; Liu et al., 2024a,b). Separate work on dialect robustness shows that semantic equivalence does not guarantee equal model behavior under non-standard language variation (Lin et al., 2025; Pan et al., 2025; Yoo et al., 2025). We extend this conversation to epistemic stance and code-mixed evaluation.

3 Materials and Methods

This section defines the scope of the benchmark, the controlled generation procedure, the quality-control checks, and the evaluation protocols. We emphasize scope because the benchmark is intended as a diagnostic minimal-pair test, not as a direct model of open-ended multilingual conversation.

3.1 Scope and interpretation

PolyJudge-Uncertain studies a deliberately narrow setting: reference-guided evaluation of short factual QA answers. This design gives strong control over semantic equivalence, but it also means that the task naturally favors concise answers close to a canonical reference format. We therefore interpret pairwise assertive preference as a protocol- and task-specific answer-format prior rather than as a universal bias against hedging in all communicative contexts. The benchmark is useful precisely because it separates semantic correctness from epistemic style under controlled conditions; it should be complemented by more naturalistic dialogue and long-form generation studies in future work.

3.2 Benchmark design

PolyJudge-Uncertain is built from 320 semantic seed items covering short factual question answering. Each seed contains a question, a canonical reference answer, and a deliberately incorrect alternative answer. We expand every seed into four language varieties: English, Hindi, Hinglish, and Bengali. For each language variety, we generate two discourse styles, assertive and hedged, and two correctness states, correct and incorrect. This yields a balanced $320 \times 4 \times 2 \times 2 = 5{,}120$-item benchmark.

The final benchmark uses canonical single-sentence templates designed to minimize unintended semantic or pragmatic drift. The key design decision is that style should vary while proposition remains fixed. For example, a correct assertive answer and a correct hedged answer should denote the same underlying fact; only the degree of epistemic commitment should differ. This enables a controlled test of whether judges respond to style rather than meaning. Figure 1 gives a concrete multilingual minimal-pair example and summarizes the benchmark construction pipeline.

3.3 Realization and quality control

The generation procedure formerly summarized as REALIZE consists of three deterministic steps. First, the seed question, reference answer, and incorrect alternative are mapped to language-specific fields. Second, the candidate answer is inserted into a language-variety-specific style template. Assertive realizations use direct answer forms, while hedged realizations prepend natural epistemic markers such as "I think," "probably," or corresponding Hindi, Hinglish, and Bengali expressions. Third, correctness is controlled only by substituting the candidate answer: the question, reference answer, language variety, and style template remain fixed within a matched group.

The procedure formerly summarized as QUALITYCHECK flags likely construction artifacts before evaluation. The checks verify that required fields are present, language and style labels are consistent, duplicate realizations are not introduced, candidate answers do not accidentally leak the reference in the incorrect condition, and length remains within

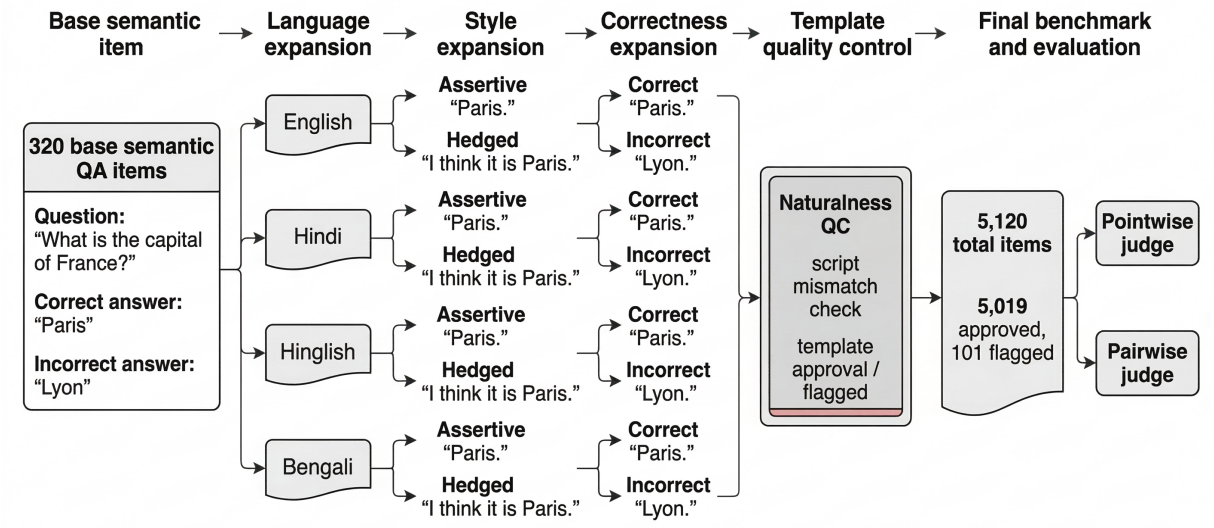

Figure 1: Concrete example and pipeline for PolyJudge-Uncertain. The same semantic seed is realized across English, Hindi, Hinglish, and Bengali while varying only answer style and correctness.

the expected range for short-answer QA. For Hindi and Bengali, the script check allows internationally shared strings such as digits, chemical formulae, and proper nouns, but flags unexpected Latin-script realizations in otherwise native-script templates. Flagged instances are retained in the main benchmark for transparency and analyzed separately in ablations.

3.4 Pilot and benchmark repair

Before constructing the final benchmark, we ran a small pilot with 80 items derived from five seed questions. That pilot showed a large pointwise penalty against hedged answers. Manual error inspection revealed that at least part of the effect was due to multilingual template issues, especially unnatural or ambiguous Hindi formulations that could plausibly be misread by the judge. We therefore revised the benchmark generation templates to use simpler, more natural constructions and introduced the quality-control pass described above.

The revised benchmark has 5,019 approved items and 101 flagged items. No items were rejected outright, but flagged items are tracked so we can test whether the main conclusions depend on including them. The final benchmark balance and quality-control split are summarized in Table 1 and Figure 3.

3.5 Judge setup

We evaluate a single reference-guided judge, GPT-4o-mini, using a base prompt in both pointwise and pairwise settings. In pointwise evaluation, the judge receives a question, a reference answer, and one candidate answer, then predicts whether the candidate is semantically correct. In pairwise evaluation, the judge receives a question, a reference answer, and two candidate answers, then chooses the more correct one or returns a tie. Pairwise answer order is randomized with deterministic seeds and mapped back to semantic categories before metric computation, reducing the risk that the main pairwise effect is merely an A/B position artifact. Figure 2 illustrates the two evaluation settings.

The experimental decision to keep the judge fixed is deliberate. Our goal in this paper is not to compare judge families but to isolate how discourse style interacts with evaluation protocol under tightly controlled semantics. This choice limits generality, and we therefore frame the finding as a focused case study rather than a universal claim about all LLM judges.

Category	Count	Share
Total items	5,120	100.0%
Unique seed questions	320	—
English	1,280	25.0%
Hindi	1,280	25.0%
Hinglish	1,280	25.0%
Bengali	1,280	25.0%
Assertive	2,560	50.0%
Hedged	2,560	50.0%
Gold-correct	2,560	50.0%
Gold-incorrect	2,560	50.0%
Approved by QC	5,019	98.0%
Flagged by QC	101	2.0%

Table 1: Composition of the final PolyJudge-Uncertain benchmark.

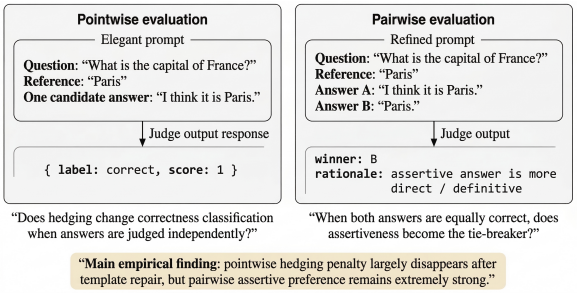

Figure 2: Evaluation protocols. Pointwise evaluation tests whether an individual candidate is judged correct against a reference, while pairwise evaluation tests which of two matched candidates is preferred. The both-correct pairwise setting isolates style preference under semantic equivalence.

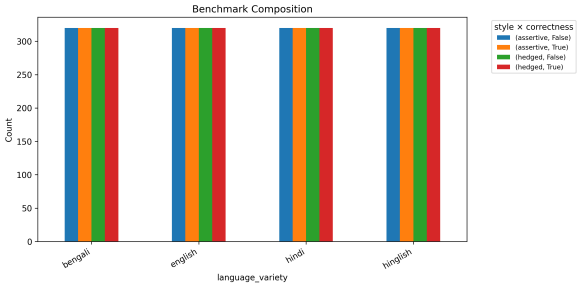

Figure 3: Benchmark composition of the final 5,120-item dataset. This plot reports the empirical distribution after multilingual expansion and quality-control flagging.

3.6 Evaluation metrics

For pointwise evaluation, we report overall accuracy and gold-correct acceptance rate. The latter is especially important because the motivating question is whether correct hedged answers are penalized more often than equally correct assertive answers. For pairwise evaluation, we report winner rates in three settings: both answers correct, both answers incorrect, and hedged-correct versus assertive-incorrect. The first tests style preference under semantic equivalence; the third tests whether style can override correctness.

4 Algorithm

Algorithm 1 summarizes the benchmark construction and evaluation pipeline. The first stage expands semantic seeds into multilingual assertive and hedged variants while preserving proposition identity. The second stage filters or flags unnatural realizations. The final stage runs pointwise and pairwise evaluation over matched item groups.

Algorithm 1 PolyJudge-Uncertain construction and evaluation

Require: Seed set S, language set L, style set $T = \{\text{assertive}, \text{hedged}\}$, correctness labels $C = \{\text{correct}, \text{incorrect}\}$
Ensure: Benchmark B, pointwise results R_p, pairwise results R_{pw}

$B \leftarrow \emptyset$
for all $s \in S$ **do**
  for all $\ell \in L$ **do**
    for all $t \in T$ **do**
      for all $c \in C$ **do**
        $x \leftarrow \textsc{Realize}(s, \ell, t, c)$
        $q \leftarrow \textsc{QualityCheck}(x)$
        Append (x, q) to B
      end for
    end for
  end for
end for
for all $b \in B$ **do**
  $R_p \leftarrow R_p \cup \textsc{PointwiseJudge}(b)$
end for
for all seed-language groups in B **do**
  Create paired comparisons under matched semantics
  $R_{pw} \leftarrow R_{pw} \cup \textsc{PairwiseJudge}(\cdot)$
end for
return B, R_p, R_{pw}

5 Experiments and Results

5.1 Pointwise evaluation

Contrary to the pilot, the repaired benchmark shows no meaningful pointwise hedging penalty. Overall pointwise accuracy on the final benchmark is 99.80% (5,110/5,120), and the style split is effectively identical: 99.80% for assertive items and 99.80% for hedged items. Among gold-correct answers, the judge accepts 99.84% of assertive responses and 99.92% of hedged responses as correct.

Table 2 breaks this result down by language variety, and Figure 4 visualizes the same pattern. The rates are nearly flat. English and Hinglish reach 100% acceptance for both styles; Hindi and Bengali show only isolated deviations. In other words, once template naturalness is repaired, there is no longer evidence that hedged-but-correct answers are systematically downgraded in pointwise multilingual evaluation.

The residual errors are concentrated in just three seed questions. Seven pointwise errors come from a synonym confound in which the benchmark marks *Luna* as incorrect for *Moon*, two come from a likely Hindi answer-quality issue in a Thailand-currency item, and one is an isolated Bengali hallucination. These errors are informative because they are not hedging-driven; they arise from semantic-

Language variety	Assertive correct	Hedged correct
English	100.00	100.00
Hindi	99.69	99.69
Hinglish	100.00	100.00
Bengali	99.69	100.00
Macro average	99.84	99.92

Table 2: Gold-correct acceptance rates (%) in pointwise evaluation on the final benchmark. The pointwise hedging penalty largely disappears after template repair.

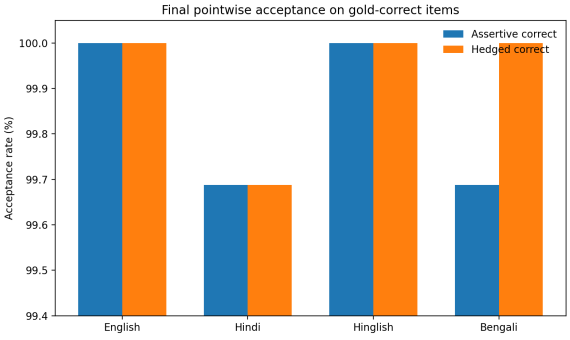

Figure 4: Final pointwise acceptance rates for gold-correct answers. The assertive and hedged bars are nearly identical across language varieties.

equivalence ambiguity and benchmark labeling.

5.2 Pairwise evaluation

The central stable effect appears in pairwise evaluation. When both candidate answers are correct and differ only in style, the judge prefers the assertive answer in 1,276 of 1,280 cases (99.69%), ties four cases, and never prefers the hedged answer. This pattern holds almost perfectly across language varieties: 100% assertive preference in English, Hinglish, and Bengali, and 98.75% in Hindi.

This means that assertiveness remains a near-universal tie-breaker even after the pointwise hedging penalty disappears. The judge can correctly recognize both answers as factually valid in isolation, yet still prefers the more assertive rendering when forced to rank them comparatively.

Table 3 should be read as three diagnostic comparisons. The both-correct row isolates pure style preference under semantic equivalence. The both-incorrect row tests whether the judge mostly recognizes that neither candidate is preferable. The hedged-correct versus assertive-incorrect row is the strongest test of whether assertiveness can override factual correctness; here, an assertive win is a genuine preference reversal. Catastrophic correctness reversal is nearly absent: the judge selects the assertive but wrong answer only 2 times out of 1,280 comparisons (0.16%). Thus, assertiveness does *not* typically override correctness. Instead, it acts as a strong secondary preference cue when correctness is tied.

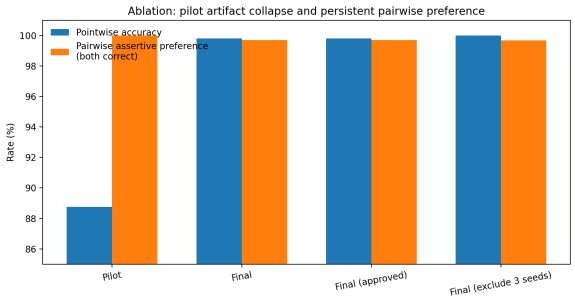

Figure 5: Ablation summary showing the collapse of the pointwise artifact after template repair and the persistence of pairwise assertive preference.

6 Ablation Studies

The ablations explain why the scientific conclusion changed between the pilot and the final benchmark. The pilot contained only 80 items and showed a much stronger pointwise effect: overall accuracy was 88.75%, gold-correct acceptance was 95.0% for assertive answers and only 60.0% for hedged answers, and pairwise evaluation always preferred the assertive answer when both answers were correct. This is the pattern that initially motivated the project.

However, once multilingual templates were naturalized and quality-controlled at scale, the pointwise effect collapsed while the pairwise effect remained. Table 4 shows three robustness checks, and Figure 5 provides a compact visual summary of the same transition. First, restricting analysis to the 5,019 approved items leaves the final conclusions unchanged. Second, excluding the three problematic seed questions yields 100% pointwise accuracy, confirming that the remaining pointwise errors are construction artifacts rather than hedging bias. Third, even after excluding those seeds, the pairwise assertive preference remains at 99.68% in the both-correct condition. The persistent failure mode is therefore comparative, not pointwise.

7 Discussion and Future Work

Our results revise the interpretation of uncertainty bias in LLM-as-a-judge. The strongest reading of the pilot was that hedged truth is directly penalized even in pointwise correctness evaluation. The

Pairwise setting	Assertive wins	Hedged wins	Tie
Both correct ($n = 1{,}280$)	1,276	0	4
Both incorrect ($n = 1{,}280$)	116	30	1,134
Hedged-correct vs. assertive-incorrect ($n = 1{,}280$)	2	1,277	1

Table 3: Pairwise outcomes on the final benchmark. In the third row, an assertive win corresponds to a preference reversal.

Setting	Size	Pointwise acc.	Assertive pref. (both correct)
Pilot benchmark	80	88.75	100.00
Final benchmark	5,120	99.80	99.69
Final, approved only	5,019	99.80	99.69
Final, excluding 3 error seeds	5,072	100.00	99.68

Table 4: Ablation summary. "Assertive pref." is the rate at which the judge prefers the assertive answer when both candidate answers are correct.

revised benchmark does not support that claim. Instead, it supports a narrower but robust conclusion: assertiveness functions as a tie-breaker in pairwise judging even when semantic correctness is equal. This distinction matters in practice. Many modern evaluation pipelines rely on pairwise comparison because it is convenient, intuitive, and often more stable than scalar scoring. Yet our findings suggest that pairwise evaluation may also be uniquely vulnerable to discourse-style cues.

This interpretation also addresses an important alternative explanation. Because the benchmark uses short factual QA, assertive answers may better match the canonical answer format expected by the task. We therefore do not claim that the result proves a general communicative bias against hedging. The more precise conclusion is that, in reference-guided pairwise evaluation of short factual answers, an LLM judge treats assertive formatting as a strong preference cue when correctness is otherwise tied. This is still practically relevant because many automatic evaluation settings use similarly concise references and pairwise comparisons.

The paper also carries a methodological warning. Multilingual robustness research is especially vulnerable to template artifacts because translation, transliteration, punctuation, and script conventions can all create unintended semantic or pragmatic shifts. A benchmark that appears to reveal fairness or robustness failures may partly be revealing unnatural prompt construction. This does not make multilingual judge-bias research less important; it makes careful benchmark auditing more essential.

Future work should extend this study in three directions. First, the benchmark should be expanded to additional languages, scripts, and code-mixed varieties, especially varieties where hedging is expressed through discourse particles rather than clause-level templates; future versions should also include naturally occurring dialogue or user-query contexts rather than only templated short answers. Second, the final benchmark should be evaluated with multiple judge models and multiple prompt variants, including explicit debiasing instructions and checklist-style judging (Liu et al., 2024a,b; Li et al., 2025b). Third, a human evaluation subset should be completed so that the study can quantify not only judge bias but also judge–human divergence under hedging. That comparison is especially important given prior evidence that humans may be substantially less punitive toward epistemic markers than LLM judges (Lee et al., 2025; Bavaresco et al., 2025; Calderon et al., 2025).

8 Conclusion

We introduced PolyJudge-Uncertain, a controlled multilingual benchmark for studying how epistemic stance affects LLM-as-a-judge. The original motivating hypothesis was that hedged answers would be penalized across languages. Our final experiments refine that picture. After template repair and quality control, pointwise hedging bias largely vanishes. The robust remaining phenomenon is pairwise: when two answers are equally correct, the judge almost always prefers the assertive one. This means that the strongest stable failure mode in our setting is not misjudging hedged truth as false, but over-valuing assertiveness as a signal of answer quality.

More broadly, the paper shows that benchmark

construction can qualitatively change the scientific conclusion. For multilingual meta-evaluation, that lesson is as important as the final metric values themselves.

Limitations

This study has several limitations. First, the final benchmark uses a single judge model, GPT-4o-mini, and a single base judging prompt. As a result, the findings establish a strong case study rather than a universal characterization of LLM-as-a-judge. Second, although the benchmark is multilingual and includes a code-mixed variety, it covers only four language varieties and focuses on short factual QA. The results may differ in longer-form generation, subjective evaluation, or dialogue settings.

Third, the benchmark intentionally prioritizes canonical, minimal templates to maximize control. This is appropriate for causal-style analysis, but it also means that the study does not capture the full variability of naturally occurring multilingual hedging. Fourth, the final rerun does not yet include completed human validation, explicit debias prompts, or judge-side calibration experiments, so the paper is diagnostic rather than intervention-oriented. Fifth, some benchmark seeds still reveal how difficult it is to draw a clean line between error and synonymy in multilingual evaluation, as illustrated by the *Moon*/*Luna* case.

Despite these limitations, we believe the paper makes a useful contribution because it identifies a stable effect, explains why an earlier stronger claim was misleading, and offers a benchmark methodology that can be extended in future work.

References

Austin S Babrow, Chris R Kasch, and Leigh A Ford. 1998. The many meanings of uncertainty in illness: Toward a systematic accounting. *Health communication*, 10(1):1–23.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in llm judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2025. Social bias evaluation for large language models requires prompt variations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14507–14530.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 13441–13460.

Ken Hyland. 2010. Metadiscourse: Mapping interactions in academic writing. *Nordic Journal of English Studies*, 9(S2):125–143.

Hawon Jeong, ChaeHun Park, Jimin Hong, Hojoon Lee, and Jaegul Choo. 2025. The comparative trap: Pairwise comparisons amplifies biased preferences of llm evaluators. In *Proceedings of the 8th BlackboxNLP*

Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 79–108.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8962–8984.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2025b. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16537–16552.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024a. Leveraging large language models for nlg evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024b. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J Wooldridge, Janet Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342.

Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim GJ Rudner, and Arman Cohan. 2025a. Metafaith: Faithful natural language uncertainty expression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29600–29644.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. 2025b. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–221.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024a. Calibrating llm-based evaluator. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pages 2638–2656.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024b. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7641–7660.

Eileen Pan, Anna Seo Gyeong Choi, Maartje Ter Hoeve, Skyler Seto, and Allison Koenecke. 2025. Analyzing dialectical biases in LLMs for knowledge and reasoning benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20882–20893, Suzhou, China. Association for Computational Linguistics.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

Victor Wang, Michael JQ Zhang, and Eunsol Choi. 2025. Improving LLM-as-a-judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23173–23199, Suzhou, China. Association for Computational Linguistics.

Ling-I Wu, Weijie Wu, Minyu Chen, Jianxin Xue, and Guoqiang Li. 2025. Co-eval: Augmenting llm-based evaluation with machine metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25765–25787.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2025. Code-switching red-teaming: LLM evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13392–13413, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.