International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

THE CONFIDENCE ILLUSION IN MULTILINGUAL LARGE LANGUAGE MODELS:EVALUATING ANSWER RELIABILITY AND UNCERTAINTY ACROSS ENGLISH, HINDI, AND TELUGU

Ms. Qudsiya Basreen [1], Mr. Mohammed Aman Faraz [2], Ms. Etikala Sri Nithya Reddy [3]

[1] Student, Department of Computer Science, Vishwa Vishwani Institute of Systems and Management, Hyderabad, Telangana, India| Email: qudsiyabasreen@gmail.com| Mobile no: 8978314943

[2] BBA Graduate, Department of Business Administration, Vivekananda Degree and PG College, Karimnagar, Telangana, India| Email: mdamanfaraz0001@gmail.com| Mobile no: 9346296042

[3] Student, Department of Computer Science, Vishwa Vishwani Institute of Systems and Management, Hyderabad, Telangana, India| Email: srinithya.etikala@gmail.com| Mobile no: 9440868629

ARTICLE INFO

KEYWORDS:

Large Language Models (LLMs); Multilingual AI; Answer Reliability; Confidence Calibration; Uncertainty Estimation; Hallucination Detection.

ABSTRACT

Large language models (LLMs) are increasingly used to gather information, explain complex concepts, and assist with decisionmaking. However, their answers can appear confident even when the available evidence is insufficient or contradictory. This raises concerns about reliability, calibration, and uncertainty communication, particularly across languages with different linguistic structures and translation challenges.

This study investigates whether ChatGPT, Google Gemini, and Claude demonstrate consistent answer reliability and confidence calibration across English, Hindi, and Telugu under different evidence conditions. The experimental dataset includes 180 language-specific question versions derived from 60 original questions, equally distributed across three languages and three evidence conditions: sufficient, insufficient, and conflicting evidence. Each question version is assessed by all three models, resulting in 540 responses.

The evaluation framework examines answer correctness, confidence scores, unsupported claims, hallucination indicators, uncertainty communication, abstention behaviour, and recognition of conflicting evidence. Quantitative measures include accuracy, the Brier score, Expected Calibration Error (ECE), overconfidence rates, and appropriate abstention rates. Comparative analyses examine model reliability and uncertainty handling across languages and evidence conditions. The study investigates whether high confidence corresponds to correct, evidence-supported answers and whether reliability varies by language and evidence condition. It provides a systematic approach to evaluating multilingual LLM reliability beyond answer accuracy alone. The findings may inform more reliable and transparent evaluation practices, particularly when evidence is incomplete, uncertain, or inconsistent.

INTRODUCTION

Background

Large Language Models (LLMs) are artificial-intelligence systems trained on large amounts of text to understand and generate human-like language. Models such as ChatGPT, Google Gemini, and Claude are increasingly used to answer questions, explain concepts, summarise information, support learning, and assist with everyday tasks. Because these systems can produce fluent and detailed responses, many users rely on them as a convenient source of information. However, the usefulness of an LLM depends not only on its ability to generate text but also on whether its answers are accurate, supported by evidence, and appropriately qualified. Reliable AI-generated information is especially important in education, research, and decision-making, where an incorrect answer may mislead users or contribute to poor decisions (Huang et al., 2025). DOI

Answer Reliability

LLM-generated answers are not always correct. A model may produce a response that sounds reasonable but includes an unsupported claim, misrepresents the supplied evidence, or presents information that cannot be verified from the available source. This problem is commonly described as hallucination. Importantly, an LLM can express an answer with high confidence even when the answer is incorrect or insufficiently supported. Such confident but inaccurate responses create a "confidence illusion," because users may mistake fluent wording, certainty, or a high stated confidence level for reliability. Therefore, users should not judge an AI response solely by how confident it appears; correctness must be assessed against the question and the available evidence (Geng et al., 2024).

aclanthology.org

Confidence and Uncertainty

In this study, confidence refers to the level of certainty expressed by a model about the correctness of its answer. Uncertainty refers to a model's acknowledgment that the available information is incomplete, ambiguous, or insufficient to establish a definitive answer. When evidence is insufficient, an appropriate response should recognise the limitation and avoid presenting an unsupported conclusion as fact. When evidence conflicts, a model should identify the contradiction

rather than select one claim without justification. Confidence should therefore be evaluated alongside answer correctness, because high confidence does not necessarily mean that an answer is accurate (Geng et al., 2024). aclanthology.org

Multilingual Evaluation

Most evaluations of LLMs are conducted primarily in English, but reliable performance in English does not automatically demonstrate equally reliable performance in other languages. Hindi and Telugu are widely used Indian languages with distinct scripts, vocabulary, grammatical structures, and naming conventions. These differences can affect how a question is understood, how evidence is interpreted, and how names, numbers, dates, and uncertainty are expressed in a response. Evaluating models in English, Hindi, and Telugu is therefore necessary to determine whether their reliability is consistent across languages. A model that performs well in English may still make translation, transliteration, or interpretation errors when responding in another language.

Evidence Conditions

This study examines model behaviour under three evidence conditions. In the sufficient-evidence condition, the provided information contains enough detail to answer the question. In the insufficient-evidence condition, the information does not provide enough support to establish a definitive answer. In the conflicting-evidence condition, the available sources contain contradictory claims. An appropriate LLM should answer accurately when evidence is sufficient, acknowledge uncertainty when evidence is insufficient, and recognise contradictions when evidence conflicts. These conditions allow the study to examine not only whether models produce correct answers but also whether they respond appropriately when a definitive answer cannot be justified.

Purpose of the Study

This research evaluates the answer reliability and uncertainty communication of ChatGPT, Google Gemini, and Claude across English, Hindi, and Telugu. Using sufficient, insufficient, and conflicting evidence conditions, the study examines answer correctness, confidence calibration, unsupported claims, hallucinations, uncertainty communication, appropriate abstention, and recognition of conflicting evidence. By comparing model performance across languages and evidence conditions, the study aims to assess whether these LLMs demonstrate consistent

reliability or whether their accuracy and confidence vary by language and evidence availability. The findings will help clarify whether high reported confidence reliably reflects answer quality across multilingual settings.

REVIEW OF LITERATURE

Previous research has examined both the usefulness and the limitations of Large Language Models (LLMs). Ji et al. (2023) define hallucination in natural-language generation as generated content that is nonsensical or unfaithful to the provided source content. Their review shows that hallucinations can occur in tasks such as question answering, summarization, dialogue, and machine translation, and that fluent wording alone does not guarantee factual accuracy. This is relevant to the present study because an LLM may provide a confident response even when its answer or explanation is unsupported by the available evidence.

Research on confidence calibration has also shown that a model's stated confidence does not always correspond to its actual correctness. Kadavath et al. (2022) found that language models can sometimes assess whether their own answers are likely to be correct, particularly in structured multiple-choice and true-or-false settings. However, this ability is imperfect and depends on the task format and evaluation method. Their work suggests that models may possess some internal signal about uncertainty, but this does not guarantee that users will receive a well-calibrated confidence estimate in every response.

Xiong et al. (2024) examined whether LLMs can express uncertainty through verbalized confidence. They found that models often tend to be overconfident when stating their confidence in natural language, although calibration and failure prediction may improve as model capability increases. Their findings remain important for the present study because they show that a high confidence score should not automatically be interpreted as evidence of answer correctness. Instead, confidence must be compared with actual performance across different questions and evidence conditions.

Multilingual evaluation is another important area of LLM research. Hada et al. (2024) highlighted the difficulty of evaluating LLMs across languages and noted that evaluation practices

are often centred on English. Their work emphasizes that performance in English cannot automatically be assumed to transfer to other languages, particularly where scripts, naming conventions, and linguistic structures differ. This supports the need to evaluate models in Hindi and Telugu as well as English, since translation, transliteration, and interpretation errors may affect answer reliability.

Overall, prior research establishes that hallucinations, overconfidence, and uneven multilingual performance are important concerns in LLM evaluation. However, fewer studies jointly examine answer correctness, confidence calibration, unsupported claims, uncertainty communication, and conflict recognition across English, Hindi, and Telugu under sufficient, insufficient, and conflicting evidence conditions. The present study addresses this gap by comparing ChatGPT, Google Gemini, and Claude across these languages and evidence conditions.

RESEARCH GAP

Although previous studies have investigated hallucinations, confidence calibration, and multilingual performance in Large Language Models (LLMs), these areas are often examined separately or under different evaluation settings. Consequently, further investigation is needed to understand how these factors interact when models answer questions across multiple languages and evidence conditions.

An important area for investigation is whether a model's confidence consistently reflects the correctness of its answers across languages. A response may be accurate when sufficient evidence is available, while a model may produce unsupported claims when evidence is incomplete or express unjustified certainty when sources contradict one another. Evaluating these situations together can provide a more comprehensive understanding of model reliability.

Furthermore, although multilingual benchmarks assess LLM capabilities across different languages, additional evaluation is needed to understand how evidence availability, uncertainty communication, and conflicting information affect reliability specifically in English, Hindi, and Telugu. Differences in language structure, translation, transliteration, and interpretation may influence how models process questions and communicate their answers.

Therefore, this study investigates the combined evaluation of answer correctness, confidence calibration, unsupported claims, uncertainty communication, appropriate abstention, and recognition of conflicting evidence across three languages and three evidence conditions. It compares ChatGPT, Google Gemini, and Claude to examine whether their reliability and confidence vary according to language and evidence availability.

OBJECTIVES OF THE STUDY

Main Objective

The main objective of this study is to evaluate the answer reliability, confidence calibration, and uncertainty communication of ChatGPT, Google Gemini, and Claude across English, Hindi, and Telugu under different evidence conditions.

Specific Objectives

1.To evaluate the correctness of responses generated by the selected LLMs across English, Hindi, and Telugu.

2.To examine the relationship between model-reported confidence and actual answer correctness.

3.To assess the occurrence of unsupported claims and potential hallucination indicators under sufficient, insufficient, and conflicting evidence conditions.

4.To evaluate how effectively the models communicate uncertainty and abstain from providing definitive answers when the available evidence is insufficient.

5.To examine whether the models recognise and appropriately respond to contradictions in conflicting evidence.

6.To compare the reliability and confidence calibration of the selected models across the three languages and evidence conditions, based on researcher-verified response labels.

RESEARCH METHODOLOGY

Research Design

This study adopts a comparative evaluation design to examine the reliability of Large Language Models across different languages and evidence conditions. Three AI models—ChatGPT, Google Gemini, and Claude—are evaluated in English, Hindi, and Telugu. The evaluation considers three evidence conditions: sufficient, insufficient, and conflicting evidence. This design enables comparisons of answer correctness, confidence, and uncertainty communication across models, languages, and evidence conditions.

Dataset and Experimental Setup

The study is designed around 180 question instances distributed equally across English, Hindi, and Telugu. Each language contains 60 question instances, comprising 20 questions for each evidence condition. The three selected models are evaluated against these question instances, resulting in a dataset of 540 model responses.

The questions cover multiple subject areas, including general knowledge, science, health, technology, history, and cultural context. The evidence provided with each question establishes the information available to the models. The experimental setup supports a consistent comparison of model responses across the three languages and evidence conditions.

Evaluation Criteria

The study considers several criteria for assessing model reliability:

Answer correctness: Whether the response provides an answer supported by the supplied evidence.

Confidence calibration: Whether reported confidence corresponds to the observed correctness of responses.

Unsupported claims: Whether a response introduces claims that are not supported by the available evidence.

Hallucination indicators: Whether a response contains potentially fabricated or factually unsupported information.

Uncertainty communication: Whether the model appropriately acknowledges limitations in the available evidence.

Appropriate abstention: Whether the model refrains from giving a definitive answer when the evidence does not justify one.

Conflict recognition: Whether the model identifies contradictions in the supplied evidence.

These criteria provide a framework for comparing response quality and the appropriateness of model confidence. Potential hallucination indicators are treated as screening flags and are not automatically considered confirmed hallucinations.

Data Organisation and Analysis

The model responses and associated evaluation records are organised in an Excel research workbook. The workbook serves as the supporting data file for the study and provides a structured record of the responses and evaluation fields.

All 540 responses were manually reviewed by the researcher against the supplied question and evidence. The review assessed answer correctness, unsupported claims, hallucination indicators, uncertainty communication, appropriate abstention, and recognition of conflicting evidence. Final response labels were based on the researcher's manual verification.

The analysis compares model performance across the three languages and evidence conditions. Quantitative measures include answer correctness, confidence calibration, and rates of unsupported claims and hallucination indicators. Qualitative examination is also used to identify patterns in uncertainty communication and responses to conflicting evidence.

The interpretation of results accounts for the evaluation procedure and the limitations of a singleresearcher review. Independent double human review and source-transcript verification were not conducted.

Research Data Availability

The research workbook contains the organised model responses and associated evaluation records used in this study. It is provided as a supplementary file to support transparency and enable readers

to examine the underlying data, subject to applicable access and sharing restrictions. Research workbook: Access the supplementary research dataset

CONCEPTUAL BACKGROUND

Reliability of Large Language Models

Large Language Models (LLMs) generate responses by predicting sequences of text based on input prompts. However, fluent answers are not always factually accurate or supported by evidence. In this study, reliability refers to a model's ability to provide correct, evidence-supported answers and appropriately handle limitations in the available information.

Confidence Calibration

Confidence calibration concerns the relationship between a model's stated confidence and the actual correctness of its answers. Models may express high confidence in incorrect responses or uncertainty about correct ones. This study compares confidence scores reported by ChatGPT, Google Gemini, and Claude with their evaluated correctness across English, Hindi, and Telugu. Confidence scores are treated as model-reported estimates, not guaranteed measures of reliability.

Hallucinations and Unsupported Claims

Hallucinations involve generated information that is false or inadequately grounded in available evidence. Unsupported claims are statements that cannot be justified using the evidence provided for a question. However, an unsupported claim is not necessarily false. Therefore, this study distinguishes potential hallucination indicators and unsupported claims from confirmed factual errors.

Uncertainty Communication and Abstention

Uncertainty communication occurs when a model acknowledges limitations in the available information. Appropriate abstention occurs when it avoids giving a definitive answer because the evidence is insufficient. These behaviours are important for preventing uncertain information from

being presented as established fact. This study examines how the selected models communicate uncertainty and abstain under insufficient-evidence conditions.

Multilingual Reliability and Evidence Conditions

LLM performance may vary across languages because of differences in linguistic structure, terminology, and cultural context. Comparing English, Hindi, and Telugu helps examine whether answer quality and confidence behaviour remain consistent across languages.

The study considers three evidence conditions: sufficient evidence, which supports answering the question; insufficient evidence, which does not justify a definitive answer; and conflicting evidence, which contains contradictory information. Together, these concepts provide a foundation for evaluating correctness, confidence calibration, unsupported claims, uncertainty communication, abstention, and conflict recognition across languages and evidence conditions.

Analysis and Discussion

Overview of the Experimental Dataset

The experimental dataset contains 540 model responses generated by ChatGPT, Google Gemini, and Claude. The responses are distributed equally across English, Hindi, and Telugu, with 180 responses for each language. Each model contributes 180 responses, and each evidence condition—sufficient, insufficient, and conflicting evidence—contains 180 responses. The dataset is balanced across the three languages, three models, and three evidence conditions, with 20 responses in each language–model–condition combination.

This balanced structure enables comparisons across the selected models, languages, and evidence conditions. However, balanced response counts do not by themselves establish the validity of the evaluation labels. The results below should therefore be interpreted in light of the workbook's preliminary evaluation status.

Answer Correctness Across Languages

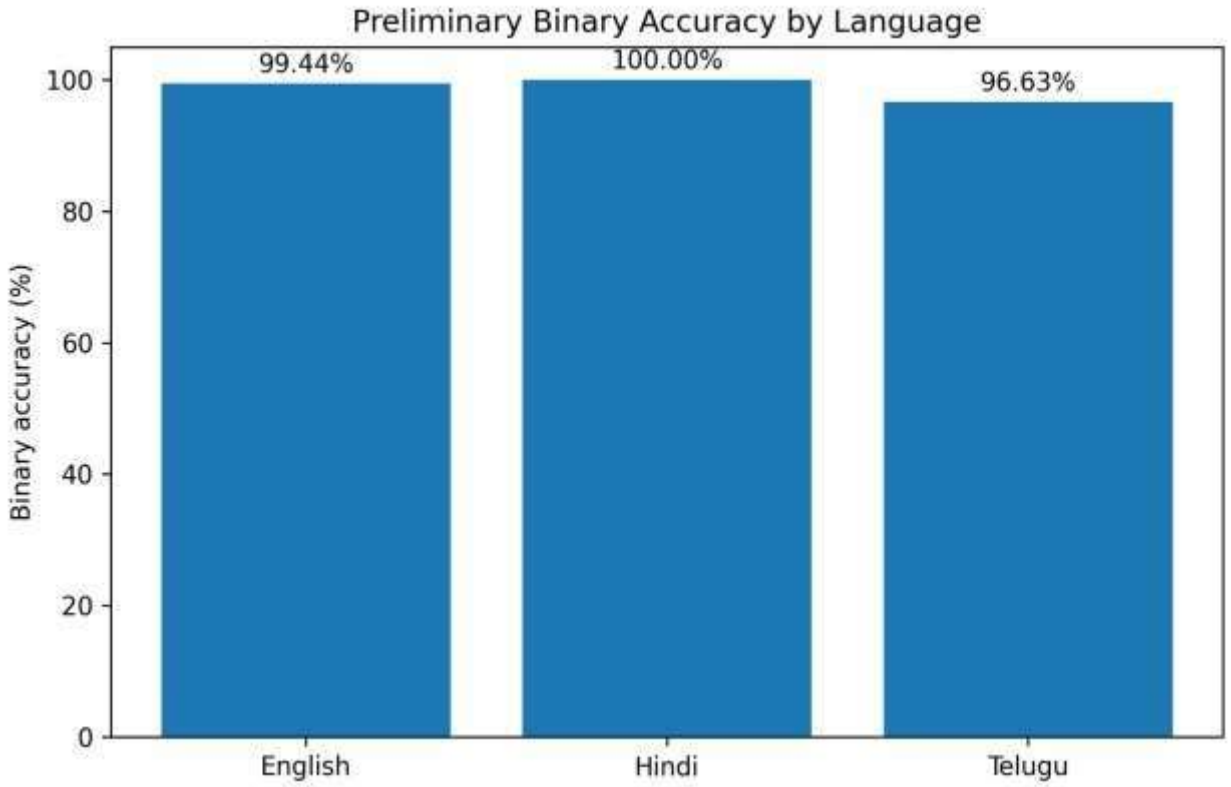
The workbook's proposed evaluation labels indicate differences in answer correctness across the three languages.

Language	Correct	Incorrect	Partially correct	Binary accuracy
English	179	1	0	99.44%
Hindi	180	0	0	100.00%
Telugu	172	6	2	96.63%
Total	531	7	2	98.70%

Source: OVERALL_RESULTS and BY_LANGUAGE sheets. Binary accuracy excludes partially correct responses; the labels remain preliminary.

The preliminary labels show the highest accuracy for Hindi, followed by English and Telugu. Telugu contains six responses labelled incorrect and two labelled partially correct. This pattern suggests that language-related factors deserve closer examination, including translation, transliteration, names, and interpretation of the supplied evidence.

However, the results do not establish that Telugu inherently causes lower model reliability. The observed differences may also reflect the particular questions, response wording, or evaluation decisions in this dataset. Further verification is necessary before drawing stronger conclusions.



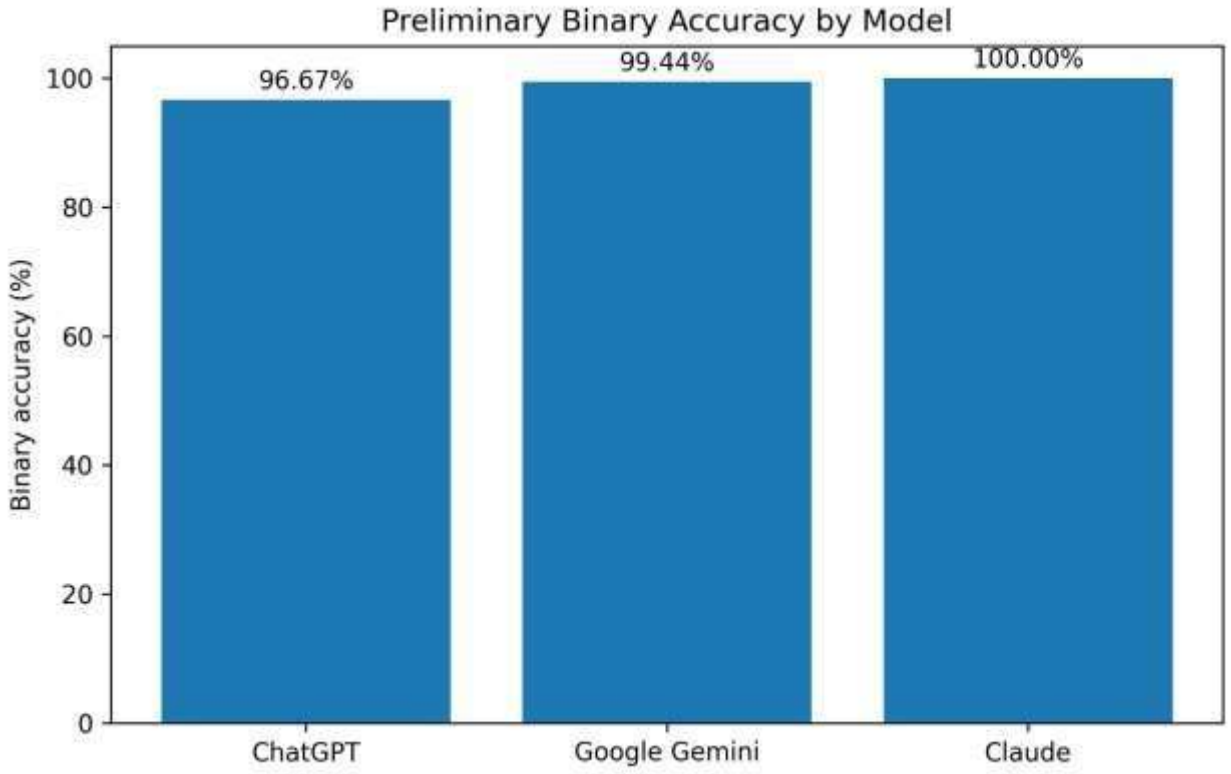
Comparison Across AI Models

The workbook also reports differences among the three models.

Model	Correct	Incorrect	Partially correct	Binary accuracy
ChatGPT	174	6	0	96.67%
Google Gemini	178	1	1	99.44%
Claude	179	0	1	100.00%*

Source: BY_MODEL sheet. *Claude's binary accuracy excludes one partially correct response; its strict correct proportion is 99.44% of all 180 responses.

According to the proposed labels, Claude has no responses classified as incorrect, while Google Gemini has one and ChatGPT has six. Gemini and Claude also have one partially correct response each.



These figures provide a descriptive comparison of the current labels. They should not be interpreted as definitive evidence that one model is generally superior to another. The responses are drawn from a specific dataset, and independent double human review was not conducted. The differences should be revisited after the response labels have been reconciled and verified.

Answer Correctness Under Different Evidence Conditions

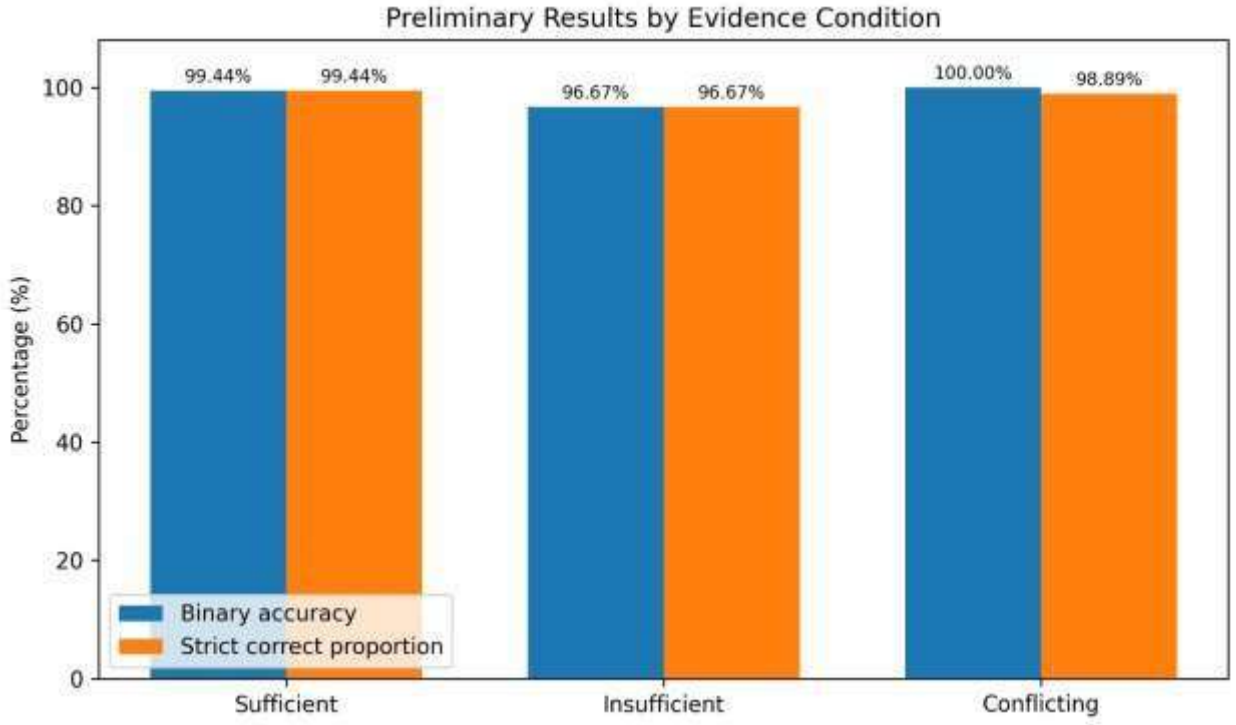
The workbook reports the following results for the three evidence conditions.

Evidence condition	**Correct**	**Incorrect**	**Partially correct**	**Binary accuracy**
Sufficient evidence	179	1	0	99.44%
Insufficient evidence	174	6	0	96.67%
Conflicting evidence	178	0	2	100.00%*

Source: BY_EVIDENCE_CONDITION sheet. The conflicting-evidence binary accuracy excludes the two partially correct responses. The strict correct proportion is 98.89%.

The proposed labels show lower binary accuracy under insufficient evidence than under sufficient evidence. This is consistent with the possibility that questions lacking enough information make reliable answering more difficult. The conflicting-evidence condition has no responses labelled fully incorrect, but two are partially correct.

Importantly, accuracy alone cannot determine whether models handled uncertainty appropriately. A model might reach the expected answer while failing to acknowledge a contradiction, or it might appropriately abstain when the evidence is insufficient. These behaviours need to be assessed separately.



Confidence Calibration and Overconfidence

The workbook reports a mean model-reported confidence of 99.44 out of 100, a preliminary Brier score of 0.01310, and a preliminary Expected Calibration Error (ECE) of 0.00827.

The Brier score and ECE are intended to compare reported confidence with observed correctness. However, their interpretation depends on the validity of the correctness labels and on whether the confidence scores are comparable estimates of answer probability. In this workbook, the calculations use confidence divided by 100 and exclude partially correct responses.

The high mean reported confidence, combined with the proposed incorrect-answer labels, makes high-confidence errors particularly relevant for further investigation. The workbook identifies seven incorrect responses with confidence of at least 90 out of 100. It also identifies nine highconfidence responses when partially correct answers are included in the broader error definition.

These are preliminary findings. They do not establish that confidence is well calibrated simply because the Brier score and ECE are numerically small. The labels, scoring assumptions, and calibration calculations should be checked before these metrics are presented as final validated results.

Unsupported Claims and Hallucination Indicators

The workbook's screening records identify 21 potential unsupported-claim cases, six uncertain cases, and one case not assessed. Its hallucination screen identifies 16 potential cases and 12 uncertain cases.

These counts indicate that evidence support and potential hallucinations warrant separate attention. An unsupported claim is not necessarily false, because a statement may be true but not established by the supplied evidence. Similarly, a potential hallucination flag is a screening result rather than confirmation that a model fabricated information.

For this reason, the reported counts should be described as screening results, not definitive hallucination rates. Claim-level review against the actual question, evidence, and answer is needed before firm conclusions can be drawn.

Uncertainty Communication, Abstention, and Conflict Recognition

The workbook reports that 174 of the 180 responses in the insufficient-evidence condition were coded as appropriate abstentions, giving a preliminary rate of 96.67%. It also records appropriate uncertainty communication in 354 of 360 responses across the insufficient- and conflictingevidence conditions, or 98.33%.

For conflicting evidence, the workbook records conflict recognition in 180 of 180 responses. However, it also flags two responses for proper-name fidelity issues. Therefore, the 100% figure should be interpreted as a preliminary behaviour-coding result, not proof that every response handled every contradiction perfectly.

These measures illustrate why uncertainty communication, abstention, and conflict recognition should be analysed separately from answer correctness. A complete evaluation needs to assess whether the model's behaviour was appropriate for the specific evidence condition, not merely whether a label indicates that it recognised uncertainty or a conflict.

Overall Discussion

Taken together, the workbook's proposed labels suggest high answer correctness across the dataset, with descriptive differences across languages, models, and evidence conditions. The same records also highlight the importance of evaluating confidence, unsupported claims, uncertainty communication, and conflict recognition independently.

The dataset's balanced design provides a useful structure for these comparisons. Nevertheless, the findings must be interpreted cautiously because the workbook contains preliminary AI-assisted evaluation labels, unresolved data-quality flags, and no independent second human evaluation. The recorded mental review is not auditable at the response level, so it cannot by itself establish that every final label was verified.

Accordingly, these results should be treated as preliminary descriptive findings until the labels and metrics are reconciled and checked. Once that work is complete, the paper can report the final figures and discuss whether the observed patterns provide evidence about the relationship between confidence and reliability across English, Hindi, and Telugu.

Major Findings

The analysis of 540 responses generated by ChatGPT, Google Gemini, and Claude across English, Hindi, and Telugu produced the following researcher-verified preliminary findings.

Overall answer correctness

Of the 540 responses, 531 were labelled correct, seven incorrect, and two partially correct. Excluding the partially correct responses from binary evaluation, the preliminary accuracy was 98.70% (531/538).

Differences across languages

English achieved a preliminary binary accuracy of 99.44%, Hindi achieved 100%, and Telugu achieved 96.63%. Telugu had the largest number of incorrect and partially correct responses, suggesting that language-specific reliability deserves further investigation.

Differences across models

ChatGPT recorded 96.67% preliminary binary accuracy, Google Gemini recorded 99.44%, and Claude recorded 100%. These figures are descriptive results from the current labels and should not be interpreted as definitive rankings without further validation.

Influence of evidence conditions

Preliminary accuracy was 99.44% under sufficient evidence and 96.67% under insufficient evidence. Under conflicting evidence, binary accuracy was 100% among fully classifiable responses; however, two responses were partially correct. This highlights the importance of reporting partial responses separately rather than relying on accuracy alone.

Confidence and calibration

The mean reported confidence was 99.44 out of 100. The calculated Brier score was 0.01310, and the expected calibration error (ECE) was 0.00827. Despite these relatively low calibration-error values, seven binary-evaluable responses with confidence of at least 90 were labelled incorrect. High stated confidence therefore should not be treated as a guarantee of correctness.

Unsupported claims and potential hallucinations

The preliminary screening flagged 21 responses for potentially unsupported claims, while six cases remained uncertain and one was not assessed. A separate hallucination screen flagged 16 responses, with 12 cases remaining uncertain. These are screening outcomes, not confirmed hallucination rates.

Uncertainty communication and abstention

Preliminary behaviour coding identified appropriate abstention in 174 of 180 insufficient-evidence cases (96.67%). Appropriate uncertainty communication was recorded in 354 of 360 cases involving insufficient or conflicting evidence (98.33%). Conflict recognition was recorded in all 180 conflicting-evidence cases, although two responses had proper-name fidelity concerns.

Overall finding

Overall, the researcher-verified preliminary results indicate that answer correctness, stated confidence, and handling of limited or conflicting evidence should be evaluated together. These findings provide a descriptive assessment of model performance in this dataset. They should not be interpreted as definitive evidence of differences in reliability between the models, languages, or evidence conditions, particularly because independent double human review and sourcetranscript verification were not conducted.

SUGGESTIONS OR RECOMMENDATIONS

Based on the preliminary findings, the following recommendations are proposed:

Improve multilingual evaluation: Large language models should be tested more extensively in English, Hindi, Telugu, and other languages to identify language-specific reliability differences.

Improve confidence calibration: Models should communicate confidence in ways that better reflect the actual reliability of their answers. High confidence should not automatically be interpreted as evidence of correctness.

Strengthen evidence-based answering: Models should distinguish between information directly supported by the available evidence and claims that remain uncertain or unsupported.

Encourage appropriate abstention: When evidence is insufficient, models should acknowledge limitations, request clarification, or abstain from giving a definitive answer rather than speculate.

Improve handling of conflicting evidence: Models should identify contradictions, explain relevant differences between sources, and avoid presenting disputed information as unquestionably correct.

Strengthen evaluation procedures: Future studies should use clearly defined scoring criteria, documented response-level verification, and independent evaluation where feasible. Potential hallucination indicators and unsupported claims should be checked through claim-level review before being reported as confirmed errors.

Expand future research: Further studies should use larger datasets, additional languages, diverse subject areas, and repeated trials to determine whether the observed patterns remain consistent.

These recommendations may help researchers and developers evaluate and improve the reliability, transparency, and responsible use of multilingual large language models.

CONCLUSION

This study examined the preliminary answer reliability, confidence calibration, and uncertainty communication of ChatGPT, Google Gemini, and Claude across English, Hindi, and Telugu under sufficient, insufficient, and conflicting evidence conditions. The dataset comprised 540 responses generated from 180 language-specific question instances, allowing comparisons across models, languages, and evidence conditions.

The preliminary results showed an overall binary accuracy of 98.70% among 538 responses classified as either correct or incorrect. However, performance varied across languages and models, with Telugu recording lower accuracy than English and Hindi in this dataset. The results also highlighted the importance of evaluating confidence alongside correctness. Although the average reported confidence was high, some highly confident responses were labelled incorrect, demonstrating that confidence alone cannot guarantee answer reliability.

The analysis further examined potential unsupported claims, hallucination indicators, uncertainty communication, abstention, and the recognition of conflicting evidence. These dimensions provide a broader understanding of model behaviour than accuracy alone. In particular, appropriate abstention and transparent communication of uncertainty are important when evidence is incomplete or contradictory.

However, the findings should be interpreted cautiously. The study used researcher manual verification rather than independent double human review, and source-transcript verification was not conducted. In addition, the study is limited to three models, three languages, and the selected question set; therefore, its findings should not be generalised to all multilingual tasks or model versions.

In conclusion, evaluating multilingual large language models requires a combined assessment of answer correctness, confidence calibration, evidence support, and uncertainty handling. Future research using larger datasets, documented human verification, and independent evaluation can help establish more reliable conclusions about how these systems perform across languages and evidence conditions.

REFERENCES

1. Hada, R., Gumma, V., de Wynter, A., Diddee, H., Ahmed, M., Choudhury, M., Bali, K., & Sitaram, S. (2024). Are large language model-based evaluators the solution to scaling up multilingual evaluation? In Findings of the Association for Computational Linguistics: EACL 2024 (pp. 1051–1070). Association for Computational Linguistics. https://aclanthology.org/2024.findings-eacl.71/

2. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

3. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). Language models (mostly) know what they know. arXiv. https://arxiv.org/abs/2207.05221

4. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In Proceedings of the International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=gjeQKFxFpZ

5. Yona, G., Aharoni, R., & Geva, M. (2024). Can large language models faithfully express their intrinsic uncertainty in words? In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 7752–7764). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.443