Linguistic Ambiguity and Hallucination in Large Language Models: A Cross-Linguistic and Cultural Analysis of Output Accuracy

Xiaozong Chen*

School of Chemistry and Chemical Engineering, Lingnan Normal University, Zhanjiang Guangdong, China, 524048
**corresponding author: cxz12365411@gmail.com*

Abstract: *An inaccurate response to an ambiguous prompt can reflect a mistaken interpretation, an unsupported factual assertion or a translation problem. These outcomes require different explanations and evaluation criteria. This paper examines their relationship through a structured literature analysis and a proposed Chinese-English evaluation of GPT and Claude models. It distinguishes ambiguity in meaning and reference from factual support, translation fidelity and the adequacy of available evidence. Existing studies inform a comparative framework, while their different models, languages and tasks limit direct comparisons. The proposed evaluation pairs ambiguous prompts with disambiguated versions and contrasts implicit cultural context with explicit context. Bilingual reviewers validate the materials before generation. Exact model versions, prompting conditions and output records are retained, and blinded human annotation separates supported claims, confirmed contradictions, unsupported additions and unresolved cases. Evaluation reports factual outcomes alongside answer coverage, clarification and abstention, with repeated outputs analysed at the item level. Mitigation involves clarifying consequential uncertainty, supplying relevant context and checking claims against appropriate sources. The paper contains no new model outputs, measured hallucination rates or model rankings. Its contribution is a design for examining how language and context affect interpretation and factual reporting without treating every incorrect or unscorable response as hallucination.*

Keywords: *large language models; linguistic ambiguity; hallucination; cross-linguistic evaluation; cultural context; factual accuracy*

1. Introduction

A question containing the word bank may concern a financial institution or the edge of a river. A response about either meaning can be reasonable when the question supplies no further context. Calling one response a hallucination because it differs from a reference answer confuses interpretation with factual support. A different problem arises when the response invents a bank's founding date or attributes a statement to a source that does not contain it. Evaluation needs to preserve this distinction before comparing models or languages.

Ambiguous questions, multilingual evaluation and cultural knowledge have each motivated specialised benchmarks[1,2,3]. Their tasks and scoring rules are not interchangeable. An answer may fit one reading of a prompt yet fail another, and a translation may resolve an ambiguity that remained open in the source. Cultural expectations also depend on the situation and participants. Chinese and English cannot be treated as proxies for two internally uniform national cultures. A comparison that changes wording, background assumptions and factual difficulty simultaneously cannot isolate a language-related difference.

This paper asks how linguistic ambiguity and cultural context should be considered when evaluating output accuracy in large language models. It uses literature analysis to examine definitions and evidence, then proposes a controlled Chinese-English evaluation involving GPT and Claude. These names identify model families; any future experiment must specify the exact models tested. Twelve primary publications were selected for their relevance to ambiguity, multilingual question answering, factuality and cultural evaluation. Selection was purposive, without a claim to exhaustive coverage. The paper reports no API experiment, sampled outputs or measured differences. Existing research informs the design and its limits. The proposed framework connects matched materials, controlled generation and

evidence-based annotation. It also examines how clarification and verification may change the balance between factual reliability and the amount of information supplied.

2. Conceptual Foundations

2.1 Linguistic ambiguity and interpretation

Ambiguity occurs when an expression permits more than one relevant interpretation. Polysemy involves related senses, such as head referring to a body part or a leader; homonymy involves distinct meanings represented by the same form, as with bank. Syntactic ambiguity concerns alternative structures, while reference ambiguity concerns the entity to which an expression points. Pragmatic uncertainty arises when the intended action or implication depends on context. These distinctions help identify what a prompt manipulation changes. They do not imply that each item has only one possible source of ambiguity.

A response can select a reasonable interpretation without recognising the alternatives. Whether that is acceptable depends on the task: a request for all possible readings has different requirements from an ordinary information question. AmbigQA treats identification of plausible answers and disambiguating question reformulation as explicit tasks[1]. In a paired evaluation, reviewers should record the permissible readings before seeing model responses. The eventual test materials require validation in their actual written languages, because an English gloss or pinyin transcription can alter the relevant ambiguity. The financial and river-edge senses of bank normally require different Mandarin expressions. Choosing either expression when translating "Describe the bank" resolves the English homonymy before generation. Such a pair would confound language with access to disambiguating information. Conversely, an illustrative Chinese utterance glossed as "Already submitted?" can leave both the actor and object unstated. Prior dialogue may resolve those referents, and English also permits ellipsis. This example motivates control of the discourse context; it does not establish that one language is intrinsically more ambiguous. Future stimuli must retain native written forms, including distinctions that romanisation may erase.

2.2 Hallucination, factual error and evidential status

Hallucination has overlapping uses in the literature. This framework reserves its measured category for asserted content that is demonstrably contradicted by the relevant evidence or constitutes an unsupported addition under a task's grounding requirement. Factual error describes a proposition established as false. Some factual errors therefore fall within the operational hallucination category, while an interpretation mismatch or mistranslation requires a separate label. In a source-grounded task, a detail absent from the supplied source can violate grounding even if its wider truth is unknown. In open question answering, omission from one reference source is insufficient to establish either falsehood or fabrication.

Evaluation should code interpretation, factual support and translation fidelity separately. A reasonable alternative reading is recorded as such. An unverifiable assertion remains unresolved unless the task provides grounds for a stronger judgement. FActScore offers a useful precedent for decomposing long responses into factual claims and checking support[4]. Its measure concerns factual precision; completeness requires additional assessment. TruthfulQA also treats truthfulness and informativeness as distinguishable concerns[5]. These distinctions prevent a short refusal from appearing equivalent to an informative, well-supported answer simply because neither contains a confirmed false claim.

2.3 Language, cultural knowledge and contextual assumptions

Language supplies forms for expressing meaning; cultural context supplies knowledge and expectations relevant to interpreting a situation. Their relationship varies across speakers, locations and activities. A Chinese prompt does not identify the user's nationality, and an English prompt does not establish an English-speaking country's social norms. A statement such as "I will think about it" may indicate a request for time or a softened refusal, depending on the stipulated relationship and circumstances. Such possibilities require examination through specified scenarios. Their interpretation cannot be assigned uniformly to an entire language community.

CulturalBench provides evidence about evaluating culturally diverse knowledge, but its largely English question format does not make it a direct Chinese-English language comparison[3]. Cultural knowledge and pragmatic interpretation also require different evidence. A documented institutional custom may have a checkable answer, whereas an unstated interpersonal intention can remain indeterminate. The proposed evaluation separates factual cultural items

from scenarios with several defensible interpretations. Explicit-context versions supply the relevant relationship or setting while avoiding disclosure of the target answer. Bilingual reviewers assess whether the added context reduces the intended uncertainty without revealing the target answer or introducing information beyond the planned contextual manipulation.

3. Evidence and Comparative Framework

3.1 What existing studies establish

The selected studies examine different parts of the problem. Research on selective answering distinguishes uncertainty about knowledge from uncertainty about the question's intended meaning, using PaLM-family models[6]. The prompt-balance study examines word-sense disambiguation with GPT-4o and LLaMA-3.1-70B across five European languages; its sensitivity to sense distributions in examples concerns interpretation performance, not hallucination rates[7]. Mu-SHROOM supplies multilingual hallucination-span annotations from open-weight generators; differences in generators and language-specific material limit direct cross-language comparisons[8].

The measures also differ. AmbigQA evaluates answer sets and disambiguating rewrites, while the prompt-balance study reports micro-F1 for predicted senses. Mu-SHROOM measures hallucination-span overlap and ranking correlation. CulturalBench includes GPT and Claude models, but its main English benchmark scores cultural knowledge through multiple-choice and stricter all-options-correct decisions[1,3,7,8]. The translation study involving GPT and Claude reports Matthews correlation coefficient for their detection decisions[9]. These scores describe performance on distinct tasks. None can be substituted directly for the proportion of a model's factual answers containing hallucinated content.

The distinctions in Table 1 organise potential output risks and the corresponding controls. They are analytical categories; no error frequencies have been measured here. Translation can affect several categories at once: a target language may require information left implicit in the source or permit a different structure. A back-translation study involving fourteen target languages, including Chinese, uses reconstruction behaviour to investigate ambiguity[10]. It does not show that translated prompts preserve equivalent ambiguity. Parallel materials therefore need human review beyond literal similarity. Each item should retain an explanation of the relevant readings and the evidence that would make an answer acceptable.

Table 1 Ambiguity-related risks and evaluation controls

Issue	Potential output risk	Proposed control
Polysemy	Related senses conflated	Specify permissible senses
Homonymy	Wrong meaning selected	Pair with meaning-specific context
Syntax	Alternative structure ignored	Validate readings in each language
Reference	Entity assigned without support	Control antecedents and identities
Pragmatics	Unstated intention treated as fact	Specify situation; allow alternatives
Translation	Meaning or uncertainty altered	Review semantic and ambiguity equivalence
Cultural presupposition	Local assumption generalised	Vary explicit context independently

3.2 A matched framework for Chinese and English

The proposed design uses item families containing Chinese and English prompts, each with an ambiguous and a disambiguated version. A separate contextual contrast varies whether relevant cultural information is implicit or explicit. When a cultural manipulation also resolves a linguistic ambiguity, reviewers record the dependence and interpret the combined contrast accordingly. Materials may include source-grounded tasks and open factual questions, but their grounding requirements and results remain separate.

MKQA illustrates the value of aligning questions and answer representations across languages, although it is not a benchmark of ambiguity or current GPT and Claude systems[2]. Translation-hallucination detection research must also be read carefully: evaluations using GPT or Claude as detectors concern their judgements about translations, not necessarily the factuality of their own generated answers[9]. Detection performance, answer generation and translation quality require distinct records and denominators.

Table 2 Stages of the proposed evaluation

Stage	Procedure	Controls	Records	Evaluation
Material construction	Build and review paired prompts	Meaning; context; answer leakage	Readings; references; revision notes	Bilingual equivalence
Controlled generation	Obtain repeated independent outputs	Version; parameters; tools; length	Exact prompts and responses	Condition compliance
Annotation and analysis	Code claims and response behaviour	Blinding; adjudication; clustering	Labels; evidence; uncertainty	Accuracy, coverage and resolution

4. Potential Mechanisms and Confounding Factors

4.1 Interpretation selection and unsupported completion

An underspecified prompt allows a model to select a reading that the evaluator did not intend. That selection can then shape subsequent details in an otherwise fluent answer. The evaluation must identify where the problem begins: choosing a permissible reading differs from adding unsupported facts within that reading. A disambiguated prompt can test whether clarification changes the response, but a difference does not reveal the model's internal reasoning. It may reflect altered associations with the added words or a change in the information available. Mechanistic claims therefore require more evidence than paired output differences alone.

4.2 Cultural defaults and pragmatic assumptions

A model may complete a scenario using a familiar institutional practice or pattern of politeness when the prompt leaves the setting unspecified. If the response presents that assumed convention as universal, a contextual assumption becomes a factual generalisation. The test therefore needs an explicit description of whose practice is relevant and which answer variation is legitimate. For interpersonal scenarios, inferred intention should remain conditional unless the prompt establishes it. Factual cultural questions can be checked against documented practices; a speaker's unexpressed intention cannot be verified in the same way. Comparing implicit and explicit context tests the response to that information, without treating a country label as a complete cultural description.

4.3 Translation and cross-linguistic asymmetry

Literal translation can introduce or remove ambiguity. A pronoun, omitted subject or syntactic attachment may need a different construction to preserve the range of readings. Pragmatic cues can also shift when a polite expression is transferred between settings. Bilingual validation should examine naturalness and interpretation separately; an idiomatic translation may settle an uncertainty that the source intentionally leaves open. Back-translation can help locate discrepancies but cannot certify equivalence. Items that cannot sustain the intended contrast should be revised or excluded from that comparison, with the reason recorded before model outputs are examined.

4.4 Knowledge coverage and generation conditions

Chinese and English item versions may differ in lexical familiarity, writing conventions or the accessibility of the knowledge they invoke. A cultural reference familiar to one reader population may be unusual for another. These factors can affect performance without establishing a general language advantage. Item review should record the knowledge required and avoid making one version more explanatory. Translated titles and names need particular attention because a familiar canonical form can serve as a retrieval clue. Findings should be reported for the tested materials and model versions, with item-level variability retained in the analysis.

Examples, system instructions and response requirements can favour one sense or discourage clarification. The prompt-balance study makes example composition a relevant concern for disambiguation evaluation[7]. Tool access and retrieval also change the information available to the model, while output limits affect how many claims it can make. A product name alone provides inadequate experimental identification because versions and configurations can differ. Generation conditions must be held comparable where possible and described where they cannot be matched. A difference between two services under different available settings is a difference between those configurations, not an isolated architectural effect.

4.5 Annotation assumptions and uncertainty

Annotators can mistake a culturally unfamiliar answer for an error or favour the intended reading after seeing the answer key. Fluent wording may also conceal an unsupported qualification or attribution. Independent annotation with evidence-based adjudication reduces reliance on these impressions. Uncertainty measures introduce another issue: variation across outputs can reflect several valid readings of the prompt. Semantic entropy concerns variation in meaning and has been investigated for detecting particular confabulations[11]. It cannot make reference evidence unnecessary, and uncertainty under an ambiguous prompt should not itself be labelled hallucination. Repeated agreement can coexist with a stable factual mistake, while several correct interpretations can produce meaning variation. Both cases require checks of the actual claims. An uncertainty score is therefore an auxiliary signal whose interpretation depends on the task and annotation.

5. Proposed Evaluation and Mitigation Framework

5.1 Construct bilingual materials with interpretable contrasts

Each item family should have a documented target question, permissible readings and a reference basis. Reviewers prepare natural Chinese and English versions, then construct disambiguated variants that specify the intended reading without giving the answer. For cultural scenarios, they record the participants and setting and identify which presupposition is varied. An explicit-context version supplies that information while retaining the factual or interpretive task. The protocol should specify which contrasts are primary and which are exploratory before generation. Items with inseparable linguistic and contextual changes can remain useful descriptive cases but cannot estimate independent contributions.

Bilingual reviewers inspect meaning, difficulty and the range of defensible responses. They also check that response-language instructions are consistent and that one version does not reveal information withheld in the other. A second review tests whether the expected answers follow from the stated evidence. Disagreements are resolved through revision or a documented set of permissible answers. Source-grounded items retain the source and the rules for using it; open questions retain independently checkable references. The same material should not be used to develop prompts and report final evaluation results. This separation reduces the risk that repeated adjustments merely fit the selected items. Review also considers the length and specificity of added context. Where feasible, a neutral-context control supplies comparable additional wording without resolving the target ambiguity. This helps examine whether a change is associated with disambiguation or with receiving more information generally. The protocol records items for which such a control is unnatural, instead of forcing all linguistic phenomena into one prompt template.

5.2 Control generation and retain complete provenance

A future implementation must identify each exact model version or snapshot and the date of access. The record includes the complete system and user prompts, sampling settings, tool permissions and requested response language. Baseline conditions should use independent conversations with no carry-over from earlier items. Repeated outputs from each condition are collected according to a predeclared sampling plan. Where a service does not expose an equivalent setting, the limitation is recorded and the interpretation of comparisons narrowed.

Retrieval or browsing should be evaluated as a separate condition with its available sources and returned evidence retained. Baseline models should receive the same task evidence and comparable output instructions. Investigators record actual response length and truncation, since a longer answer has more opportunities to assert unsupported details. Truncated responses remain in the technically valid trial set and are flagged separately. Factual scoring uses only complete, interpretable assertions; incomplete tasks count as inaccurate under the completion criterion. Item order can be randomised, and paired conditions kept in separate conversations to prevent direct hints from the companion prompt. Failed calls, empty returns and interruptions have their own status. They should not disappear from the record or be interpreted as voluntary abstention.

5.3 Annotate interpretation, claims and response behaviour

Annotators should work without model identities or the study's directional hypotheses. Native-language competence alone does not establish expertise in every cultural item, so relevant contextual knowledge must also be available. Reviewers first identify the reading used by a response and whether it is permissible. They then segment factual

assertions while preserving qualifications, attribution and negation. Each claim is coded as supported, contradicted, an unsupported addition established under the grounding requirement, or unresolved. For open questions, a failed search alone cannot justify an unsupported-addition label. A mistranslation receives a separate fidelity label even when it also changes a fact.

Response-level labels record a substantive answer, clarification request, abstention or other behaviour under explicit coding rules. A response can answer part of a question and request clarification about the remainder; this mixed status should remain visible. Two reviewers independently annotate the selected material and document evidence for disagreements. Adjudication produces a final label while retaining the initial judgements, allowing agreement to be reported. Ambiguities in the coding guide are resolved before final scoring. Automated claim decomposition can assist organisation, as in factuality evaluation research, but human review must check whether segmentation changes the proposition being judged[4]. An illustrative source-grounded response might add a date that the source never supplies. That addition is scored against the stated grounding rule, with its real-world truth left unresolved unless independently established. A date directly contradicted by the source receives the contradiction label. These primary categories are mutually exclusive for counting, although a claim can carry a separate translation-fidelity label. The annotation record must identify the evidence and rule supporting the decision.

5.4 Define denominators and account for repeated outputs

Claim-level reporting separates supported claims, contradictions, established unsupported additions and unresolved claims. The operational hallucination proportion is the number of contradictions plus established unsupported additions divided by all assessable factual claims; unresolved claims are excluded from that denominator and their proportion among all factual claims is reported separately. The two hallucination subtypes are also reported separately. A response containing no assessable factual claim has no claim-level proportion. It remains in the response-level account. This measure does not equal one minus answer accuracy, which may also reflect interpretation, relevance or task completion.

Answer coverage is the proportion of technically valid prompt trials containing a substantive answer, including responses that also request clarification. Technically valid means a returned model response, regardless of whether it is correct or completes the task. Clarification and abstention are reported separately, with overlaps disclosed and technical failures outside this denominator. Task accuracy is the proportion of substantive-answer trials meeting the documented interpretation, factual and completion criteria. Partial answers are recorded separately and do not count as fully correct. Results include paired changes between ambiguous and disambiguated conditions. Repeated outputs from one item family are dependent, so uncertainty intervals should resample or model item families. Treating each output as independent would overstate the information obtained. Language and model comparisons require attention to item composition and configuration differences. They support bounded associations, not broad causal claims about languages or cultures. Results should also be reported by task type and language before pooling. Any aggregate uses a declared weighting rule. When models answer different subsets, the comparison reports that selection difference; accuracy on the common answered subset alone can conceal questions that one model systematically avoids.

5.5 Evaluate mitigation together with its costs

A clarification strategy is appropriate when different plausible readings would materially change the answer. Its evaluation should examine whether the request identifies the relevant uncertainty and whether a later response answers the clarified question. Follow-up information comes from predefined, validated replies that supply equivalent clarification to each model. Providing context can also help, but the added information must be checked for answer leakage. Retrieval offers access to evidence only when suitable sources are available and the response uses them correctly. Citation presence alone is insufficient; the cited passage must support the particular claim. These interventions should be compared on the same validated item families.

A verification workflow can separate an initial response, checks of factual propositions and a revised response. Chain-of-Verification provides a research basis for testing this structure without establishing its effectiveness for the proposed Chinese-English comparison[12]. Verification questions should be checked against independent evidence where possible, because repetition of an unsupported assertion cannot validate it. The evaluation records additional turns, response length and changes in coverage alongside factual outcomes. A strategy that reduces confirmed errors by declining most questions has a different practical value from one that preserves useful answers. Selective-answering research makes this coverage-quality relationship an explicit evaluation concern[6].

6. Conclusions and Limitations

6.1 Implications for cross-linguistic accuracy evaluation

Linguistic ambiguity changes what an answer can reasonably mean, while hallucination evaluation concerns the status of asserted content under the task's evidence requirements. A credible comparison must preserve both dimensions. The proposed framework identifies permissible interpretations before generation and evaluates factual support after a response's interpretation has been established. It also separates translation fidelity and unresolved evidence from confirmed factual problems. This prevents a reference-answer mismatch from automatically becoming a hallucination label.

For Chinese-English comparisons, matched wording is insufficient without review of the meanings, contextual assumptions and answer criteria. Cultural scenarios require specified settings, with language and nationality recorded as separate attributes where relevant. Exact model identification and generation records further limit what a comparison can claim. Reporting coverage, clarification and abstention alongside factual outcomes makes the behaviour of a cautious model visible without treating lack of an answer as proof of accuracy. The resulting evidence can support a comparison of defined model configurations on validated materials.

6.2 Limitations and future research

The proposed evaluation has not been implemented. The cited studies differ in languages, tasks and model families, and several concern systems other than GPT or Claude. Their results cannot be pooled into a current Chinese-English hallucination rate. Operational definitions also affect the findings, particularly the treatment of unsupported additions and unresolved assertions. The proposed categories need testing for consistency and usability before a larger study.

A pilot should first establish whether reviewers can validate item equivalence and apply the annotation guide with adequate agreement. Subsequent evaluation can examine paired prompt contrasts and the effects of mitigation while retaining item-level uncertainty. A broader study would need additional languages, cultural settings and knowledge domains before discussing general patterns. Changes in model versions require new records and appropriately bounded comparisons over time. Evidence about observed responses should remain distinct from explanations of model internals: differences associated with ambiguity or context can motivate mechanistic hypotheses, but cannot alone establish why a model generated a particular assertion.

References:

[1] Min, S., Michael, J., Hajishirzi, H., & Zettlemoyer, L. (2020). AmbigQA: Answering ambiguous open-domain questions. Proceedings of EMNLP 2020, 5783–5797. https://doi.org/10.18653/v1/2020.emnlp-main.466

[2] Longpre, S., Lu, Y., & Daiber, J. (2021). MKQA: A linguistically diverse benchmark for multilingual open domain question answering. Transactions of the Association for Computational Linguistics, 9, 1389–1406. https://doi.org/10.1162/tacl_a_00433

[3] Chiu, Y. Y., Jiang, L., Lin, B. Y., et al. (2025). CulturalBench: A robust, diverse, and challenging cultural benchmark by human-AI CulturalTeaming. Proceedings of ACL 2025 (Volume 1: Long Papers), 25663–25701. https://doi.org/10.18653/v1/2025.acl-long.1247

[4] Min, S., Krishna, K., Lyu, X., et al. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. Proceedings of EMNLP 2023, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

[5] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of ACL 2022 (Volume 1: Long Papers), 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

[6] Cole, J. R., Zhang, M. JQ, Gillick, D., Eisenschlos, J. M., Dhingra, B., & Eisenstein, J. (2023). Selectively answering ambiguous questions. Proceedings of EMNLP 2023, 530–543. https://doi.org/10.18653/v1/2023.emnlp-main.35

[7] Sumanathilaka, D., Micallef, N., & Hough, J. (2025). Prompt balance matters: Understanding how imbalanced few-shot learning affects multilingual sense disambiguation in LLMs. Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models, 7–15. https://doi.org/10.26615/978-954-452-105-9-002

[8] Vázquez, R., Mickus, T., Zosa, E., et al. (2025). SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared task on hallucinations and related observable overgeneration mistakes. Proceedings of SemEval-2025, 2472–2497. https://aclanthology.org/2025.semeval-1.322/

[9] Benkirane, K., Gongas, L., Pelles, S., et al. (2024). Machine translation hallucination detection for low and high resource languages using large language models. Findings of ACL: EMNLP 2024, 9647–9665. https://doi.org/10.18653/v1/2024.findings-emnlp.564

[10] Mehrparvar, B., & Pezzelle, S. (2024). Detecting and translating language ambiguity with multilingual LLMs. Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), 310–323. https://doi.org/10.18653/v1/2024.mrl-1.26

[11] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

[12] Dhuliawala, S., Komeili, M., Xu, J., et al. (2024). Chain-of-Verification reduces hallucination in large language models. Findings of ACL 2024, 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212