Beyond Hallucination: Temporal Knowledge Asymmetry as a Distinct Failure Mode in Large Language Models for Non-Western Knowledge Domains

Parth Anand

Abstract. Most work on LLM reliability focuses on hallucination, which refers to models generating confident but false information. This paper identifies and characterizes a related but distinct failure mode that has received less attention: temporal knowledge asymmetry, where models provide correct information for one geographic context but outdated information for a structurally equivalent question about another context. I evaluated three frontier LLMs (Claude Opus 4.1, Gemini 3, and ChatGPT 5.1) using a matched-pair question bank of 500 items across six knowledge domains. Contrary to my initial hypothesis, outright hallucination rates were statistically similar between Indian and Western questions (India 1.7% vs. West 1.6% averaged across models). However, I found a consistent pattern of temporal lag in Indian institutional knowledge: models gave outdated answers for Indian current affairs at rates up to 9.6 percentage points higher than for equivalent Western questions. For both Claude Opus 4.1 and Gemini 3, the outdated-answer rate for Indian current affairs was 8% while the rate for equivalent Western positions was 0%. This asymmetry was consistent across all three independently developed models, suggesting a systemic cause rooted in training data composition rather than any single model's design. I propose temporal accuracy as a necessary additional evaluation dimension for LLM benchmarks targeting equitable global deployment.

Keywords: large language models, hallucination, temporal knowledge, geographic bias, AI equity, developing nations, knowledge currency, LLM evaluation

1 INTRODUCTION

When I ask a frontier language model who is currently the Governor of the Reserve Bank of India, I might expect a correct, up-to-date answer. When a colleague in London asks the same model who is currently the Governor of the Bank of England, they very likely get one. The question this paper investigates is whether that asymmetry is systematic, and if so, whether it is better characterized as hallucination or something else entirely.

Hallucination in large language models has received substantial research attention over the past several years (Ji et al., 2023; Huang et al., 2023; Maynez et al., 2020). The core definition is the generation of confident but factually incorrect information — the model invents something that is simply wrong. Benchmark evaluations like TruthfulQA (Lin et al., 2022) have quantified hallucination rates across model families. What this literature has examined less carefully is whether LLM factual reliability is geographically uniform, and whether there are failure modes beyond outright hallucination that matter practically for users.

I began this study expecting to find higher hallucination rates for Indian-context questions than for Western equivalents, based on the well-documented underrepresentation of South Asian content in large English-language training corpora (Bender et al., 2021; Dodge et al., 2021). The data did not support this. Hallucination rates across 500 matched question pairs were statistically indistinguishable between geographies. But in examining the data carefully, I found something more specific: models were not fabricating information about India, they were providing information about India that was real but outdated. They knew the institutions but not their current occupants.

This pattern, which I term temporal knowledge asymmetry, describes a situation where a model's knowledge of one geographic context is more current than its knowledge of an equivalent context, even when both contexts are queried in English on structurally identical questions. A model that names the previous Chief Justice of India when asked the current one, while correctly naming the current Chief Justice of Canada, is not hallucinating in the traditional sense. But it is providing less useful information to the Indian user than to the Canadian one. The practical consequence is the same.

The contributions of this paper are: (1) introducing temporal knowledge asymmetry as a distinct, empirically identifiable failure mode in LLMs; (2) presenting a matched-pair evaluation methodology that separates temporal lag from outright hallucination; (3) providing empirical evidence from three frontier models showing consistent temporal lag for Indian institutional knowledge; and (4) arguing for temporal accuracy as a necessary evaluation dimension in benchmarks targeting equitable global deployment.

2 RELATED WORK

2.1 Hallucination in Language Models

The study of hallucination in neural text generation goes back to abstractive summarization research, where models

were found to generate plausible but unfaithful summaries (Maynez et al., 2020; Kryscinski et al., 2020). As LLMs have become general-purpose information systems, hallucination has taken on broader significance. Ji et al. (2023) provide a comprehensive taxonomy of hallucination types in natural language generation, distinguishing between intrinsic hallucination (contradiction of source material) and extrinsic hallucination (unverifiable fabricated content). Huang et al. (2023) survey mitigation strategies including retrieval augmentation (Lewis et al., 2020), self-consistency (Wang et al., 2023), and chain-of-thought prompting (Wei et al., 2022).

TruthfulQA (Lin et al., 2022) remains the most widely used benchmark for factual accuracy in LLMs. A significant limitation of TruthfulQA, and of hallucination benchmarks more broadly, is that they are constructed overwhelmingly from Western English-language knowledge sources. This limits their applicability to non-Western contexts and means they are not designed to detect the geographic asymmetries studied in this paper.

2.2 Geographic and Cultural Disparities in LLMs

Several papers have documented that LLM performance is not geographically uniform. Navigli et al. (2023) find substantial performance disparities across language families in multilingual models. Cao et al. (2023) show that LLMs encode cultural values disproportionately aligned with WEIRD populations (Western, Educated, Industrialized, Rich, and Democratic). Shi et al. (2023) demonstrate that chain-of-thought reasoning capabilities degrade when problems are presented in non-English languages. Bender et al. (2021) argue that the documented overrepresentation of English-language Western content in large pretraining datasets provides the mechanistic basis for these performance disparities.

My work sits within this literature but addresses a specific English-language asymmetry: temporal knowledge lag for non-Western institutional content queried in English. This is distinct from language-based performance gaps because both the questions and answers in my evaluation are in English. The disparity is geographic, not linguistic.

2.3 Temporal Aspects of Factual Knowledge

The temporal limitations of LLMs trained on static corpora are well-documented. Dhingra et al. (2022) introduce the concept of time-sensitive questions and demonstrate that model performance degrades substantially for such questions. Luu et al. (2022) study temporal misalignment between training data and test time, finding systematic accuracy drops for events occurring close to or after the training cutoff. Kasner and Dusek (2022) examine factual accuracy in data-to-text generation with time-sensitive content.

What the existing temporal knowledge literature has not examined is whether temporal lag operates uniformly across geographic contexts. My paper is, to the best of my knowledge, the first to show systematically that temporal lag is geographically asymmetric: models appear to incorporate more recent information about Western institutional changes than about equivalent Indian institutional changes.

2.4 AI Equity and Non-Western Contexts

Sambasivan et al. (2021) document gaps in AI system reliability for South Asian contexts and argue that current AI development practices systematically disadvantage non-Western populations. Hovy and Spruit (2016) make the case that NLP performance disparities have real downstream social consequences. My findings add empirical grounding to these arguments, showing that even frontier commercial systems exhibit reliability disparities that disadvantage users seeking current information about Indian contexts.

3 METHODOLOGY

3.1 Research Design and Motivation

The core design challenge for this study was separating two distinct phenomena: outright hallucination (models fabricating incorrect information) and temporal lag (models providing information that was once correct). Most LLM evaluation frameworks treat both as equivalent failures. I argue they are conceptually different — one reflects a gap in the model's knowledge base, the other reflects a gap in the recency of that knowledge. I designed the evaluation to distinguish between them.

I used a matched-pair design: for every Indian-context question, I constructed a structurally equivalent Western-context question on the same knowledge domain. This controls for question difficulty, format, and domain, isolating geography as the variable of interest. Within each response, I distinguished between correct answers, outdated answers (previously correct, currently wrong), and hallucinated answers (never correct).

3.2 Question Bank Construction

I constructed a question bank of 500 items: 250 India-context questions paired with 250 Western-context equivalents across six categories (Table 1). Each pair satisfied four equivalence criteria: identical question structure and specificity level; comparable Wikipedia article depth for ground truth verification; similar temporal stability characteristics; and same knowledge domain.

Category	India	West	Description
Political Leadership	50	50	Current heads of government

Geographic Knowledge	50	50	City facts, rivers, capitals
Historical Events	50	50	Dates, events, figures
Literary and Cultural	50	50	Authors, directors, artists
Science and Technology	25	25	Institutions, missions
Current Affairs	25	25	Current officeholders
Total	250	250	500 questions total

Table 1. Question bank composition by category.

Ground truth answers were verified against primary sources including official government websites and Wikipedia. For time-sensitive categories, I re-verified answers on the day of querying. Western questions used primarily UK, German, French, Australian, and Canadian equivalents rather than US equivalents where possible.

3.3 Models and Query Protocol

I evaluated three frontier LLMs via their web interfaces: Claude Opus 4.1 (Anthropic), Gemini 3 (Google), and ChatGPT 5.1 (OpenAI). Queries were conducted on March 18–19, 2026. Each question was submitted with the instruction: "Answer the following factual question as concisely as possible. If you are not certain of the answer, say so explicitly." I started a new conversation for each batch of 10–12 questions to prevent context contamination.

3.4 Rating Protocol

I rated each response on a three-point scale. A score of 1.0 indicates a correct answer matching current ground truth. A score of 0.5 indicates an outdated answer — information that was once correct but is no longer current. A score of 0.0 indicates a hallucinated answer with no correspondence to either current or recent ground truth.

The 0.5 category is the methodological contribution that distinguishes this study from standard hallucination evaluation. A response of "Shaktikanta Das" to a question about the current RBI Governor receives 0.5 — Das genuinely served as Governor until late 2024. A fabricated name receives 0.0. Responses expressing genuine uncertainty were scored as correct (1.0), since refusing to hallucinate is appropriate behavior.

4 RESULTS

4.1 Overall Accuracy and Hallucination Rates

Table 2 reports accuracy, outdated-information, and hallucination rates for each model. Outright hallucination rates were low across all models and showed no consistent directional pattern. Claude Opus 4.1 hallucinated on 2.0% of Indian questions vs. 1.2% Western. Gemini 3 hallucinated identically at 0.8% for both. ChatGPT 5.1 showed a slight reversal, hallucinating marginally more on Western questions (2.8%) than Indian ones (2.4%).

Model	IN corr.	IN outd.	IN hall.	W corr.	W outd.	W hall.
Claude 4.1	96.8%	1.2%	2.0%	98.0%	0.8%	1.2%
Gemini 3	98.4%	0.8%	0.8%	99.2%	0.0%	0.8%
GPT 5.1	92.4%	5.2%	2.4%	93.6%	3.6%	2.8%
Average	95.9%	2.4%	1.7%	96.9%	1.5%	1.6%

Table 2. Accuracy by model and geography.

The outdated-information column tells a different story. India outdated rates exceed Western outdated rates for every single model: Claude (1.2% vs. 0.8%), Gemini (0.8% vs. 0.0%), ChatGPT (5.2% vs. 3.6%). The consistency of direction across three independently trained models suggests a common cause rather than model-specific variation.

4.2 Category-Level Results

Table 3 shows average accuracy by category and geography, averaged across all three models. Five of six categories show India performing comparably to or better than Western questions. The exception is Current Affairs, where the gap is 9.6 percentage points — more than twice the next-largest gap.

Category	India	West	Gap	Driver
Political Lead.	0.950	0.940	–0.010	Outdated (IN)
Geographic	0.997	0.983	–0.013	Halluc. (W)
Historical	0.990	0.967	–0.023	Halluc. (W)
Literary/Cult.	0.990	1.000	+0.010	Minimal
Science/Tech	0.960	1.000	+0.040	Halluc. (IN)
Current Affairs	0.891	0.987	+0.096	Outdated (IN)

Table 3. Average accuracy by category. Positive gap = India underperforms.

4.3 Questions That Stumped All Three Models

Three questions were answered incorrectly by all three models simultaneously. Two are Indian-context: (1) "What is the name of India's first indigenous submarine?" (correct: INS Kalvari); and (2) "Who is the Chief Justice of India?" — all three models named a previous incumbent. The one Western question where all three models failed was "What was the name of the UK's first satellite?" (correct: Prospero). This confirms the evaluation is not systematically biased toward finding Indian failures.

5 THE CURRENT AFFAIRS FINDING IN DEPTH

5.1 Per-Model Breakdown

Table 4 reports the full breakdown for the Current Affairs category by model. The consistency is striking: all three models gave exactly 8% outdated responses for Indian current affairs questions, and 0% outdated responses for Western current affairs questions. These are independently developed systems from different organizations with different architectures, training procedures, and data pipelines.

Model	IN corr.	IN outd.	IN hall.	W corr.	W outd.	W hall.
Claude 4.1	88%	8%	4%	100%	0%	0%
Gemini 3	88%	8%	4%	100%	0%	0%
GPT 5.1	76%	8%	12%	96%	0%	4%

Table 4. Current Affairs breakdown per model (n=25 per geography per model).

The specific questions generating outdated responses were: the current Governor of the Reserve Bank of India (two models gave outdated answers), the current Chairman of ISRO (two models), the current Chief Justice of India (two models outdated, one hallucinated), and the number of terms served by Nitish Kumar as Chief Minister of Bihar (two models outdated). All four concern high-profile Indian institutional positions. Equivalent Western questions were answered correctly by all models.

5.2 What This Tells Us About Training Data

The most natural explanation for a pattern consistent across three independent models is something they share: training data. If the most recent portions of large English-language pretraining corpora contain proportionally less coverage of Indian institutional changes than of equivalent Western changes, all three models will exhibit the same temporal lag for Indian content.

This is consistent with known English-language news coverage asymmetries. A change in the Governor of the Bank of England is likely covered more extensively and for longer in major English-language outlets than an equivalent change in the Governor of the RBI, even though both are significant institutional events. If recent training data skews toward the most heavily covered content, this asymmetry would produce exactly the pattern observed.

6 DISCUSSION

6.1 Practical Implications

A user in Delhi asking a frontier LLM who currently leads the Reserve Bank of India may receive a confidently stated outdated answer with no way of knowing it is outdated. A user in London asking equivalent questions about the Bank of England receives current information. This is an information equity disparity. The same commercial AI system, queried in the same language on structurally identical questions, provides more current information to users in one geographic context than another. The system does not signal the difference.

6.2 Temporal Accuracy as an Evaluation Dimension

The standard hallucination evaluation framework — which treats any incorrect response as a single failure type — is insufficient for capturing this finding. A system that names the previous RBI Governor passes a binary correct/incorrect check in the same category as a system that fabricates a person who never existed. But these are meaningfully different failure modes with different causes, different detection strategies, and different remediation approaches.

I propose that future LLM benchmarks include temporal accuracy as a distinct evaluation dimension: the proportion of responses that are not merely factually grounded but currently correct. The matched-pair methodology introduced here provides a replicable framework for this evaluation.

6.3 Limitations

The Current Affairs and Science and Technology categories contain only 25 questions per geography. For the Current Affairs temporal asymmetry finding, the chi-square test yields $p=0.126$, which does not reach conventional significance thresholds at this sample size despite a large effect size (Cohen's $h=0.574$). The finding should be interpreted as a strong preliminary signal warranting replication at larger sample sizes, not as a definitively established result.

Accuracy ratings were performed by a single rater. Models were queried via web interfaces rather than APIs, introducing potential variation in system context and prompting conditions. Future work should standardize query conditions using API access with controlled system prompts and fixed temperature settings. This study also only examines one developing-nation context (India). Whether similar asymmetries exist for African, Southeast Asian, or Central Asian contexts is an important open question.

7 CONCLUSION

I started this study expecting to show that frontier LLMs hallucinate more about India than about Western contexts. They do not, at least not in any statistically reliable way. But the investigation revealed something more interesting: these models know India well, they just know a slightly older version of it.

Across 500 matched question pairs and three independently developed frontier systems, I find consistent temporal knowledge lag for Indian institutional and current affairs knowledge. For both Claude Opus 4.1 and Gemini 3, the outdated-answer rate for Indian current affairs was

8% while the equivalent rate for Western questions was 0%. ChatGPT 5.1 showed the same 8% outdated rate for India against 0% for the West. Three different companies, three different training pipelines, the same directional result.

This is not hallucination in the traditional sense. The information models provide about India is grounded in reality. It is just behind. That distinction matters both conceptually — for how we theorize about LLM reliability — and practically, for the users who receive outdated institutional information presented with the same confident tone as current information.

The methodological contribution — the three-point rating framework combined with a matched-pair design — provides a replicable tool for studying geographic asymmetries in LLM knowledge currency. Replicating the empirical finding at larger sample sizes, across additional non-Western contexts, and with API-controlled query conditions is the natural next step.

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of FAccT 2021 (pp. 610–623).

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between ChatGPT and human societies. In Proceedings of C3NLP 2023.

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., & Cohen, W. W. (2022). Time-aware language models as temporal knowledge bases. Transactions of the ACL, 10, 257–273.

Dodge, J., Sap, M., Marasovic, A., et al. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv:2104.08758.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–83.

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In Proceedings of ACL 2016 (pp. 591–598).

Huang, L., Yu, W., Ma, W., et al. (2023). A survey on hallucination in large language models. arXiv:2311.05232.

Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248.

Kasner, Z., & Dusek, O. (2022). Neural pipeline for zero-shot data-to-text generation. In Proceedings of ACL 2022.

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In Proceedings of EMNLP 2020 (pp. 9332–9346).

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in NeurIPS 33.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of ACL 2022 (pp. 3214–3252).

Luu, K., Khashabi, D., Gururangan, S., et al. (2022). Time waits for no one! Analysis and challenges of temporal misalignment. In Proceedings of NAACL 2022 (pp. 5944–5958).

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of ACL 2020 (pp. 1906–1919).

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. ACM Journal of Data and Information Quality, 15(2), 1–21.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In Advances in NeurIPS 35.

Sambasivan, N., Arnesen, S., Hutchinson, B., et al. (2021). Re-imagining algorithmic fairness in India and beyond. In Proceedings of FAccT 2021 (pp. 315–328).

Shi, F., Suzgun, M., Freitag, M., et al. (2023). Language models are multilingual chain-of-thought reasoners. In Proceedings of ICLR 2023.

Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. In Proceedings of ICLR 2023.

Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in NeurIPS 35.

APPENDIX A. SAMPLE MATCHED QUESTION PAIRS

The following examples illustrate one matched pair from each category.

Category	India question	Western equivalent
Political	Who is the Chief Minister of Himachal Pradesh?	Who is the Governor of Montana?
Geographic	Which river flows through Varanasi?	Which river flows through Lyon, France?
Historical	What was the name of India's first nuclear test?	What was the name of France's first nuclear test?
Cultural	Who wrote The God of Small Things?	Who wrote Beloved?
Science	What does ISRO stand for?	What does ESA stand for?
Current Affairs	Who is the Governor of the Reserve Bank of India?	Who is the Governor of the Bank of England?

Table A1. One matched pair per category illustrating the equivalence design.