

Cross-Lingual Hallucination Patterns in Indic Languages:

A Typology-Aware Research Agenda for Dravidian and Indo-Aryan Comparison

Authors

Lead Researcher: Pranay M Mahendrakar

Mail: pranaymahendrakar2001@gmail.com

Contact: +91 6361723454

Company: SONYTECH

Date: April 2026

Abstract

Multilingual hallucination in large language models has begun to receive systematic attention. Recent work has measured aggregate hallucination rates across 19 to 30 languages (FACTSCORE-multilingual, MFAVA, Poly-FEVER), introduced multilingual benchmarks (HalluVerseM3, BHRAM-IL for some Indo-Aryan languages), and demonstrated that LLMs hallucinate more in low-resource languages than in English. What this growing literature does not yet provide is a typology-aware comparison of how hallucination patterns — not just rates — differ across closely-related languages. The Dravidian languages (Tamil, Telugu, Kannada, Malayalam) and the Indo-Aryan languages (Hindi, Marathi, Bengali, Gujarati) of India offer a natural experimental setting: substantial variation in training data quantity, sharply different morphological typology (Dravidian agglutination versus Indo-Aryan fusion, presence or absence of grammatical gender), shared script families in some cases, and pervasive code-mixing with English in real use. This paper makes three claims. First, the question "how often does the model hallucinate in language X" is the wrong primary question; the more informative question is "which categories of hallucination dominate in language X, and why." Second, four hypotheses about Indic-language hallucination patterns can be distinguished empirically and have meaningful policy and research implications. Third, the experimental work needed to test these hypotheses is small enough that a single research team in India could execute it within a year. We propose a measurement framework, a concrete experimental protocol, and a case for why this question is particularly tractable in the Indian research context.

Keywords

Hallucination, multilingual LLMs, Indic languages, Dravidian languages, code-mixing, factuality, low-resource NLP, AI4Bharat

1. Introduction

Large language models hallucinate. They produce confident, fluent, and factually wrong outputs. The phenomenon has been measured extensively in English (Ji et al., 2023) and increasingly in other languages. Recent work has documented that hallucination rates rise sharply for low-resource languages: Mehri et al. (2024) extended the FACTSCORE methodology to 19 languages and found systematic gaps; Vamvas et al. (2025) estimated rates across 30 languages; Singh et al. (2024) demonstrated significant performance gaps between English and 29 Indic languages on the IndicGenBench benchmark; the Poly-FEVER benchmark (2025) extended fact verification to 11 languages with 77,973 claims.

This is real progress, and it sets the stage for a question the existing literature does not yet answer. Knowing that a model hallucinates more in Tamil than in English is useful but

coarse. The more informative question — and the one with stronger research and policy implications — is which categories of hallucination dominate in Tamil, and whether those categories are the same as in Telugu, Kannada, Hindi, or English. Aggregate rate measurements treat all hallucinations as equivalent. They are not.

The Indic language landscape is unusually well-suited to address this question. India has over 1.4 billion people, twenty-two scheduled languages, and at least four major language families (Indo-Aryan, Dravidian, Tibeto-Burman, Austroasiatic). The four Dravidian languages — Tamil, Telugu, Kannada, Malayalam — together have over 250 million native speakers and share a typological profile (agglutinative morphology, subject-object-verb word order, no grammatical gender for inanimate nouns) that contrasts sharply with Hindi, Marathi, Bengali, and Gujarati on the Indo-Aryan side. Training data quantities vary by an order of magnitude across these languages. Code-mixing with English

is pervasive in real use. This is, in other words, a near-ideal natural experimental setting.

This paper makes three claims. First, the field's primary question for multilingual hallucination should shift from rates to categories: not "how much does the model hallucinate in language X" but "which kinds of hallucinations dominate, and what does that tell us." Second, four specific hypotheses about Indic-language hallucination patterns can be distinguished empirically and have meaningful implications for both research and deployment. Third, the work to test these hypotheses is achievable for a research team based in India within a year, using existing models, modest compute, and the language access advantages of being in the relevant linguistic context.

Section 2 reviews what is known about multilingual and Indic-specific hallucination. Section 3 argues for the typology-aware framing. Section 4 develops four testable hypotheses. Section 5 proposes a measurement framework. Section 6 sketches a concrete experimental protocol. Section 7 makes the case for India as the right context for this work. The contribution is methodological and conceptual, but unlike most position papers in the area, the experimental programme it proposes is designed to be actually executable.

2. Background

2.1 Hallucination in LLMs: General

The standard taxonomy distinguishes intrinsic hallucinations (output contradicts the input) from extrinsic hallucinations (output adds claims unsupported by the input or by world knowledge), and faithfulness from factuality. Survey work (Ji et al., 2023) catalogues many specific failure modes; benchmarks like TruthfulQA, HaluEval, and FActScore operationalise different aspects. The mechanisms are partially understood: spurious associations from training data, over-confident next-token sampling on rare claims, and the model's lack of explicit knowledge about its own uncertainty all contribute. The relevant point for this paper is that hallucination is not a single phenomenon — it is a family of failure modes whose relative prevalence depends on input, model, and task.

2.2 Multilingual Hallucination Work

Mehri et al. (2024) extended FACTSCORE to 19 languages by translating biographies and comparing against multilingual Wikipedia sources, revealing significant gaps especially in low-resource languages. Vamvas et al. (2025) trained a multilingual hallucination detector and used it to estimate rates across 30 languages and six LLM families on open-

domain question answering — the MFAVA benchmark. Poly-FAVER (Lou et al., 2025) extended FAVER, Climate-FAVER, and SciFact to 11 languages with 77,973 claims, evaluating intrinsic hallucination detection. HalluVerseM3 (recent work, 2026) introduced fine-grained hallucination annotations across English, Arabic, Hindi, and Turkish, with the pattern that detection accuracy degrades sharply in lower-resource languages and Hindi exhibits the lowest detection accuracy of the four.

These works share a common structure: take a benchmark or methodology developed for English, extend it to N languages, report the gap. They are valuable. They have one shared limitation for our purposes: they treat hallucination as a single measured rate, even when separately measuring intrinsic and extrinsic categories, and do not connect rate variation to language typology beyond the broad high-resource-versus-low-resource axis.

2.3 What Has Been Done Specifically for Indic

Indic-specific work has expanded substantially. AI4Bharat at IIT Madras has produced IndicTrans, IndicNLP, and most recently IndicIFEval, a verifiable instruction-following benchmark across 14 Indic languages (Jayakumar et al., 2025). Google Research India released IndicGenBench (Singh et al., 2024), the largest generation benchmark covering 29 Indic languages across four language families. The "Better to Ask in English?" study (Singh et al., 2025) compared factual accuracy across 19 Indic languages including Tamil, Telugu, Kannada, and Malayalam, finding systematic accuracy gaps between English and Indic responses across multiple model families.

On hallucination specifically, BHRAM-IL (Terdalkar et al., 2026) introduced a benchmark covering Hindi, Gujarati, Marathi, Odia, and English with 36,047 questions across nine categories. This is the most directly relevant existing work and is well-constructed, but it covers Indo-Aryan languages and English; it does not include any Dravidian language. The combination of "Better to Ask in English?" (Dravidian languages included, but only aggregate factual accuracy) and BHRAM-IL (fine-grained hallucination categories, but no Dravidian languages) leaves the typology-aware comparison this paper proposes effectively unaddressed.

3. Why Categories, Not Rates

Three observations motivate the shift from rate-based to category-based questions.

3.1 The Indic Landscape Is Linguistically Rich

Indo-Aryan and Dravidian languages share the Indian subcontinent and have influenced each other for millennia,

but they are typologically distinct. Dravidian languages are agglutinative — words are built from morphemes glued together, often producing long single-word forms that English would express as phrases. Indo-Aryan languages are fusional, with shorter words but more complex inflection. Dravidian languages lack grammatical gender for inanimate nouns; Hindi and Marathi mark gender extensively. Tamil and Malayalam use distinct scripts (Tamil and Malayalam script), while Telugu and Kannada share a closely-related script family. These differences plausibly affect what the model can express well, what kinds of mistakes it makes, and where its confidence is calibrated.

3.2 Different Languages May Hallucinate Differently

Several mechanisms suggest that hallucination categories may not be uniform across languages even when controlling for rate. A model that has seen relatively little Telugu training data may produce many extrinsic hallucinations (added unsupported claims) because it has limited grounding; the same model in Hindi, with more training data, may instead produce primarily intrinsic hallucinations (logical contradictions) because the surface fluency masks deeper reasoning errors. A Dravidian language without grammatical gender may not produce the gender-confusion hallucinations that show up in Hindi or Marathi. A model that pivots through English internally — generating in English and then translating — will produce a third pattern: outputs that are factually correct relative to English knowledge but that introduce translation artefacts at the linguistic surface. None of this is currently established empirically; it is hypothesis space.

3.3 Code-Mixing Is the Real Use Case

Real Indic-language interaction with LLMs is heavily code-mixed. A user in Bengaluru asking a question in Kannada will routinely use English technical terms in the same sentence; the response, to be useful, must handle this. Existing benchmarks largely use monolingual prompts, which is fine for measurement comparability but does not capture the deployment conditions where most user-facing failures actually occur. Code-mixed inputs plausibly create distinctive failure modes: the model may correctly handle the English content while hallucinating around the Kannada content, or vice versa, in ways that pure monolingual evaluation will miss.

4. Four Testable Hypotheses

Four hypotheses are worth distinguishing. Each implies a different research and deployment response.

4.1 H1: Rate Scales Inversely with Training Data

The simplest hypothesis: hallucination rates are determined primarily by the quantity of training data in each language, with all other factors secondary. Under H1, Tamil and Telugu — both with substantial native-speaker populations and reasonable web presence — would hallucinate at similar rates, while Konkani or Sindhi would be substantially worse, with the gap to English a function of relative corpus size. This is the implicit assumption behind much current work and it is partially supported by existing aggregate measurements.

4.2 H2: Categories Differ Even at Similar Rates

A more interesting hypothesis: even when overall hallucination rates are matched between two languages, the distribution across categories differs systematically. Tamil and Hindi may both hallucinate at, say, 20% on a particular benchmark, but Tamil's failures are concentrated in entity attribution while Hindi's are concentrated in numerical claims. Under H2, mitigations would need to be language-specific and aggregate metrics would obscure deployment-relevant variation.

4.3 H3: Translation Pivoting Introduces a Distinctive Pattern

Many multilingual LLMs are believed to perform internal reasoning in English (or in a language-agnostic representation that is closer to English than to other languages), and to produce non-English outputs partly through implicit translation. If H3 holds, hallucinations in Indic languages should fall into two distinct types: those inherited from the underlying English reasoning (factually wrong in any language) and those introduced by the translation step (linguistically incorrect, culturally inappropriate, or factually distorted by translation choice). Distinguishing these would require evaluation pipelines that can attribute errors to their source.

4.4 H4: Code-Mixing Creates Distinctive Failure Modes

Code-mixed inputs may produce hallucinations that monolingual evaluation cannot detect: confused entity resolution when the same concept is mentioned in two languages, asymmetric handling of mixed-language constraints, or failure to maintain consistency between English technical content and Indic-language explanation. This hypothesis has the strongest deployment relevance because code-mixed interaction is the dominant real-world pattern in Indian LLM use.

5. A Measurement Framework

5.1 Hold the Underlying Question Constant

The first methodological commitment: every test instance should exist in parallel translations across all evaluated languages, asking the same underlying question. This separates language-of-prompt effects from underlying-question difficulty. IndicGenBench established the parallel-multi-way structure for several Indic languages; the same approach can be extended to a hallucination-focused benchmark.

5.2 Stratify by Hallucination Category

A defensible category decomposition for Indic hallucination evaluation: entity hallucination (wrong person, place, organization), numerical hallucination (wrong dates, quantities, amounts), reasoning hallucination (logically inconsistent claims), cultural hallucination (incorrect or fabricated culturally-specific content — important in the Indic context), and linguistic hallucination (grammatically wrong or unidiomatic, distinct from factual errors). Each category should report its own rate per language.

5.3 Include Code-Mixed Conditions

Each test instance should be evaluated in at least three conditions: pure monolingual (Indic only), pure English, and naturally code-mixed (Indic with English technical terms inserted at attested rates). This is more work but it captures deployment-relevant failure modes that monolingual evaluation misses.

5.4 Native-Speaker Evaluation Is Required

Automated hallucination metrics in low-resource languages are themselves unreliable; the meta-detector problem compounds the underlying detection problem. For a study to produce trustworthy results, native-speaker evaluation of at least a stratified sample is essential. This is operationally easier in India than almost anywhere else, since the relevant native-speaker populations are large, well-educated, and accessible to research institutions.

6. An Experimental Protocol

6.1 Language Selection

A defensible eight-language selection: four Dravidian (Tamil, Telugu, Kannada, Malayalam), three Indo-Aryan (Hindi, Marathi, Bengali), and English as the high-resource baseline. This provides the typological contrast (Dravidian vs Indo-Aryan), within-family variation (the four Dravidian languages differ in script and corpus size), and a strong English anchor.

6.2 Question Categories and Scale

Five hundred test items per language category (entity, numerical, reasoning, cultural, linguistic), parallel-translated across all eight languages, totalling roughly 4,000 items per language and 32,000 across the full benchmark. Half the items should also have code-mixed variants. This is large enough for statistical claims and small enough to be feasible in a year with appropriate native-speaker collaboration.

6.3 Models to Evaluate

A defensible set: GPT-4-class proprietary models, Gemini (which has had specific Indic-language attention from Google Research India), open-weight Llama-3 family, Gemma-2 family, and at least one Indic-focused open model such as Krutrim or Sarvam-class. Comparing across this set separates model-specific findings from general patterns.

6.4 Predictions Paired with Falsification

- Prediction A. Aggregate rates will roughly track training-data quantity (H1 partially supported), but category distributions will differ significantly across languages even at matched rates (H2 also supported). Falsified if categories track rates within statistical noise.
- Prediction B. Cultural hallucination will show the largest cross-language variation, with Indic languages systematically more prone to fabricating culturally-specific content. Falsified if cultural-category errors are uniform across languages.
- Prediction C. Code-mixed conditions will produce distinctive error patterns not predicted by monolingual evaluation, with at least 20% of code-mixed errors falling into categories that pure monolingual evaluation under-counts. Falsified if code-mixed and monolingual error distributions match.
- Prediction D. Models that show evidence of English-internal reasoning (via probe analyses) will show H3-style distinctive translation-error patterns; models trained more natively on multilingual data (Aya, Sarvam-class) will not. Falsified if all models show similar error distributions.

7. Why India Is the Right Place to Do This Work

The case is structural, not parochial. India has the languages — multiple Dravidian and Indo-Aryan languages with sufficient digital infrastructure to be evaluable. India has the institutions: AI4Bharat at IIT Madras, Google Research India, IIIT Hyderabad's NLP groups, and a maturing ecosystem of Indic-focused industry teams (Krutrim, Sarvam, Karya). India

has the native-speaker access necessary for trustworthy evaluation, with educated bilingual or trilingual annotators reachable through existing research and crowdsourcing pipelines. And India has the deployment pressure: hundreds of millions of users interacting with LLMs in Indic languages, with code-mixing as the default.

The specific work proposed in this paper — eight languages, 32,000 items, native-speaker evaluation, multi-model comparison — is achievable within the budget and timeline of a single well-organised research group. The infrastructure, the linguistic access, and the policy stakes are all favourably aligned. We argue this is one of the more tractable high-impact research programmes available to Indian researchers in AI evaluation today.

8. Conclusion

Multilingual hallucination evaluation has reached the stage where rate measurements are well-established and the marginal value of additional rate studies is small. The next informative move is to ask which kinds of hallucinations dominate in which languages and why. The Indic context is unusually well-suited to address this — typological diversity, varying corpus sizes, pervasive code-mixing, strong institutional infrastructure, and direct deployment relevance all converge.

This paper has proposed a typology-aware research agenda, four falsifiable hypotheses, a measurement framework that takes code-mixing and category stratification seriously, and a concrete experimental protocol calibrated to be executable within a year. Unlike most position papers in the multilingual-AI space, the work proposed here is small enough that a single team can do it. The empirical study that would test these hypotheses is, in our view, the natural next step — and it is the kind of study that is best done by researchers with direct linguistic access to the languages in question, which is to say, in India.

References

- Jayakumar, T., Khan, M. S. U. R., Dabre, R., Puduppully, R., & Kunchukuttan, A. (2025). IndicIFEval: A benchmark for verifiable instruction-following evaluation in 14 Indic languages. AI4Bharat / IIT Madras Technical Report.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).
- Lou, H., Liu, S., Zhang, R., Choe, K., Bang, J., & Sridhar, A. (2025). Poly-FEVER: A multilingual fact verification benchmark for hallucination detection in large language models. arXiv:2503.16541.
- Mehri, M., Mathur, A., Vassallo, V., & Suhr, A. (2024). Multilingual hallucination gaps in large language models. arXiv:2410.18270.
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., et al. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. *EMNLP 2023*.
- Singh, H., Gupta, N., Bharadwaj, S., Tewari, D., & Talukdar, P. (2024). IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. *ACL 2024*, 11047–11073.
- Singh, S., Naik, A., Garimella, A., & Iyer, R. (2025). Better to ask in English? Evaluating factual accuracy of multilingual LLMs in English and low-resource Indic languages. arXiv:2504.20022.
- Terdalkar, H., Bhojani, K., Dongare, A., & Behera, O. A. (2026). BHRAM-IL: A benchmark for hallucination recognition and assessment in multiple Indian languages. arXiv:2512.01852.
- Vamvas, J., Aepli, N., & Sennrich, R. (2025). How much do LLMs hallucinate across languages? On realistic multilingual estimation of LLM hallucination. arXiv:2502.12769.
- HalluVerseM3 authors. (2026). HalluVerseM3: A multitask multilingual benchmark for hallucination in LLMs. arXiv:2602.06920.
- Mishra, A., Asai, A., Balachandran, V., Wang, Y., Neubig, G., Tsvetkov, Y., & Hajishirzi, H. (2024). FAVA: Fine-grained hallucination detection and editing for language models. *NAACL 2024*.
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *ACL 2022*.
- Kunchukuttan, A., Kakwani, D., Golla, S., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). AI4Bharat-IndicNLP corpus: Monolingual corpora and word embeddings for Indic languages. arXiv:2005.00085.