IndEuph-170: Benchmarking Cultural Pragmatics through Euphemism Detection in Indian English

Debamita Samajdar
Jawaharlal Nehru University
New Delhi, India
debamita.samajdar@gmail.com

Abstract

Large Language Models (LLMs) have shown remarkable proficiency in standard English benchmarks, yet their ability to navigate the sociopragmatic cues of non-Western English varieties remains underexplored. This paper introduces IndEuph-170, a novel benchmark dataset focused on Indian English (IndE) euphemisms — expressions whose roots lie in local social hierarchies, politeness norms, and cultural taboos (e.g., "setting," "loose character," "suitable boy"). IndEuph-170 comprises 170 curated IndE sentences, against which the performance of two distinct architectures was evaluated: a fine-tuned BART model and GPT-4. The findings reveal a significant "cultural gap". While GPT-4 achieves 82.5% accuracy, it struggles with authoritative and punitive nuances. BART achieves 55.3% accuracy but exhibits a high rate of false positives by over-classifying general Indianisms as euphemisms. The paper argues that current multilingual benchmarks such as MME (Fu et al., 2025) and GLUE (Wang et al., 2018) fail to capture these dialectal pragmatics, and that a culturally-aware evaluation framework for Global Englishes is necessary.

Keywords: Indian English, Euphemism Detection, LLM Evaluation, Cultural Pragmatics, NLP Benchmarking

1. INTRODUCTION

Euphemisms act as linguistic shields — a buffer that allows speakers to navigate sensitive topics like death, sex, and social status without being overtly direct. In the Indian sociolinguistic context, these expressions diverge from Western standards by incorporating local social hierarchies and politeness norms (Sailaja, 2012).

Indian English is recognised as a distinct, institutionalised variety of English (Sailaja, 2012) with its own unique sociopragmatic norms; however, NLP models often exhibit a dialectal bias, wherein they fail to distinguish between standard regional variations (e.g., 'passed out' for graduation) and intentional meaning-masking euphemisms (e.g., 'loose character'). Far from being mere slang, these phrases flout Gricean maxims of Manner or Quantity to negotiate or preserve social harmony. For example, "loose character" encodes societal morality, while "convent-educated" signals high social standing and English fluency. The role of pragmatics — the meaning between the lines — is central to such expressions, making them a difficult test for AI, which may interpret 'adjusting' literally rather than as a socially-enforced compromise.

Large Language Models (LLMs) are predominantly trained and tested on Western datasets such as MME (Fu et al., 2025) or GLUE (Wang et al., 2018). Consequently, models like GPT-4, while excellent at standard English, often fail to internalise regional pragmatics. "Indianisms" are frequently labelled as errors by AI when they are, in fact, purposeful euphemisms.

Current research by Hu et al. (2024) introduced the JointEDI framework, which improved bilingual euphemism identification but remains restricted to Standard American English and Mandarin, thereby overlooking the 1.4 billion speakers of Global Englishes. This study aims to bridge this gap by evaluating how models like BART and GPT-4 handle the specific ways in which Indian English encodes and softens meaning.

This paper introduces a benchmark of 170 Indian English sentences categorised by euphemistic type (e.g., Understatement, Humor, Indirectness). Comparing the detection accuracy of a fine-tuned BART model with GPT-4's zero-shot performance reveals that, while LLMs are improving, they still struggle with the authoritative and punitive tones unique to Indian social contexts. This benchmark specifically targets the pragmatic competence of models rather than mere semantic similarity.

2. METHODOLOGY

2.1 Dataset Collection

This study presents IndEuph-170, a curated benchmark dataset of 170 Indian English (IndE) sentences. Designed through scraping Reddit,

Twitter, Quora, and Indian-English media articles, IndEuph-170 targets pragmatic expressions common in the Indian subcontinent that are absent from existing benchmarks such as MME (Fu et al., 2025) and GLUE (Wang et al., 2018). The dataset comprises 101 euphemistic instances and 69 literal "Indianisms" (e.g., "passed out," "missed call"). To supplement

natural samples and ensure diversity, 40 sentences were synthetically generated and subsequently validated by native speaker annotators against the same annotation framework applied to the full dataset. Loanword examples were included where widely accepted (e.g., jugaad, masala). The balance of cultural slang and true euphemisms tests whether models can distinguish between the two categories.

Category	Sentences	Pragmatic Focus
Euphemisms	101	Meaning Masking / Taboo
Literal / Indianisms	69	Dialectal Variation
Total	170	

Table 1: Dataset Composition

2.2 Categorisation Framework

The euphemisms were classified into five distinct types, following the taxonomy of Allan and Burridge (1991):

(1) Indirectness: Expressions avoiding direct reference (e.g., "loose character").

(2) Understatement: Expressions minimising the intensity of a situation (e.g., "small scene").

(3) Humor / Slang: Witty or informal local terms to soften a taboo (e.g., "doing timepass," "chutney").

(4) Politeness: Terms maintaining social harmony or respect (e.g., "suitable boy," "good name").

(5) Social Status: Expressions masking class or educational hierarchies (e.g., "convent educated").

2.3 Experimental Design: Pilot and Scaling Phase

For the pilot phase, a double-blind annotation was conducted on a subset of 40 sentences to ensure the reliability of euphemism labels. A native speaker of Indian English performed the primary annotation; a second native speaker validated the classification. Both annotators possess extensive experience in educational linguistics and student language assessment.

The evaluation was conducted in two distinct phases:

Pilot Phase (GPT-4 Evaluation): A subset of 40 sentences was tested using GPT-4 via manual prompt-based interaction. This phase evaluated two tasks: Detection (binary classification) and Paraphrasing (rewriting euphemisms in literal language).

Scaling Phase (BART Evaluation): The full 170-sentence dataset was processed using the facebook/bart-large-mnli model via Hugging Face in a Google Colab environment. This phase focused exclusively on binary detection to measure model robustness at scale.

It is acknowledged that this two-phase design does not permit a direct, head-to-head comparison between GPT-4 and BART under a unified protocol. The asymmetry was a deliberate methodological choice. It was driven by two constraints: (1) the cost and rate-limiting of GPT 4's API made full 170-sentence evaluation impractical within this study's scope, and (2) the two phases serve distinct research purposes. The pilot phase assesses qualitative pragmatic competence (detection and paraphrasing). And the scaling phase stress-tests binary detection robustness at volume. Table 2 metrics are indicative of each model's characteristic failure modes, and not a direct performance race. Future work will evaluate both architectures under a unified experimental protocol on an expanded dataset in order to enable fair comparison.

2.4 Annotation and Validation

To ensure label reliability, a double-blind annotation was conducted. A primary native speaker of IndE and a secondary native speaker, both with expertise in educational linguistics, annotated the pilot subset. A Cohen's kappa of 0.79 was achieved, indicating "substantial agreement."

Annotators identified: (a) whether a euphemism was present, (b) the target taboo or sensitive word, and (c) a literal paraphrase of the sentence. Each entry was thus labelled with a binary label, the target word, and a standard English translation of the intended meaning. Discrepancies were resolved through consensus based discussion to ensure that the final "Gold Standard" labels reflect a shared cultural understanding of IndE.

3. RESULTS AND DISCUSSION

3.1 Pilot Phase Results (GPT-4)

In the pilot phase, GPT-4 demonstrated high proficiency, achieving 82.5% accuracy (7 mismatches) in the detection task. In the paraphrasing task, it achieved 90% accuracy, with

only 4 semantic mismatches where the model failed to capture the specific Indian social gravity of an expression.

Model	Acc.	Prec.	Rec.	F1
BART Large	55.3%	58.6%	84.1%	69.0%
GPT-4	82.5%	80.5%	84.6%	82.5%

Table 2: Performance Metrics

3.2 Scaling Phase Results (BART) The scaling phase revealed a significant performance drop. As detailed in Table 2, BART Large yielded an accuracy of only 55.3%. While the model showed high Recall (84.1%), its Precision was low (58.6%) due to excessive false positives.

3.3 Error Analysis: The "Indianism" Bias

The most striking result from the BART evaluation is the high rate of False Positives (FPs). Out of 69 literal control sentences, BART incorrectly flagged 60 as euphemisms, revealing a "dialectal bias": the model treats any non-standard English construction as a euphemism, despite such constructions being standard usage in the Indian subcontinent.

Phrases such as "Give me a missed call," "I passed out of college," and "What is your good name?" were classified as meaning-masking expressions. BART's training on Standard American/British English evidently causes it to treat regional pragmatic variation as inherently "suspicious" or indirect.

Three compounding factors appear to drive this bias. First, BART-large-mnli was pre-trained predominantly on Western English corpora (BookCorpus, CC-News, OpenWebText), meaning its entailment priors have no representation of IndE as a grammatically coherent variety. Second, class imbalance in the IndEuph-170 dataset — with euphemistic instances (101) outnumbering literal controls (69) — may have reinforced a positive-classification tendency. Third, BART's zero-shot NLI framing (hypothesis: "This sentence is a euphemism") is an imprecise instrument for the pragmatic detection task. It requires sensitivity to context and speaker intent rather than surface lexical cues. Future work offers a promising avenue to address this bias through targeted fine-tuning on Indic English corpora, or by substituting Indic specific architectures such as IndicBERT (Kakwani et al., 2020) or MuRIL (Khanuja et al., 2021).

3.4 Qualitative Gaps in GPT-4: Pragmatic Blind-Spots

Despite its higher accuracy, GPT-4 failed on culturally specific expressions involving Indian authority and social dynamics. It missed the punitive tone in "The police launched a campaign to crack down…" and the corruption-related nuance in "He has a setting with the manager."

These results highlight a Western-centric bias and an absence of deep sociopragmatic mapping of the Indian context, even in the most advanced LLMs.

A closer breakdown of GPT-4's seven detection errors reveals a consistent pattern: the model struggles specifically with Social Status and authority-coded expressions. Three of the seven errors involved Social Status euphemisms (e.g., "convent-educated," "decent family"), where GPT 4 identified the literal meaning correctly but failed to recognise the hierarchical subtext. Two errors involved Indirectness expressions with punitive or institutional authority registers (e.g., "the officer asked him to cooperate"). GPT-4 parsed these as literal requests rather than coercive softening. One error involved a Humour/Slang category item ("jugaad fix") where the loanword was treated as untranslatable rather than euphemistic. The final error involved a code-mixed token where the English frame around a Hindi noun misled the classifier. These patterns suggest that GPT-4's failures are not random but cluster around power asymmetric social registers that are culturally specific to the Indian context. Few-shot prompting with culturally annotated exemplars may improve performance on these categories in future evaluations. Techniques such as explicitly encoding the euphemism taxonomy as part of the system prompt could also prove helpful.

3.5 Type-Specific Performance

A category-wise breakdown of the 101 euphemisms reveals that Humour and Slang (e.g., "chutney," "timepass") achieved the highest detection rates across both models. Understatement (e.g., "small scene," "adjusting") proved most difficult to paraphrase correctly, frequently resulting in literal translations that strip the expression of its social gravity.

4. CONCLUSION

This paper demonstrates that Indian English euphemisms pose a significant challenge for current NLP models. Through the curation of IndEuph-170, a 170-sentence benchmark, it has been shown that state-of-the-art models consistently either over-classify dialectal variations as euphemisms (BART) or miss

culturally specific taboos related to authority and social negotiation (GPT-4).

The interpretative failure is twofold: (1) smaller models over-classify dialectal variations as euphemisms, and (2) larger models miss the social power dynamics embedded in Indian expressions. It is not sufficient for an AI to be "multilingual"; it must also be trained to be "multicultural."

Future work will expand IndEuph-170 to over 1,000 samples, incorporating Bengali-English and Tamil-English code-mixed euphemisms to determine whether the bias persists in mixed language settings. To ensure annotation reliability and cultural authenticity at scale, expansion will employ a structured protocol: each regional variety will be annotated by a minimum of two native speakers with demonstrated competency in the target variety, with Cohen's kappa ≥ 0.75 required for inclusion. Taxonomic consistency will be maintained by anchoring annotations to the five-category Allan and Burridge (1991) framework used in the current dataset, with an additional code-mixing category to capture intra sentential switching. Additionally, training Indic specific models such as IndicBERT (Kakwani et al., 2020) or MuRIL (Khanuja et al., 2021) on the expanded dataset, and evaluating both models under a unified protocol, represents the primary next step toward fair cross-architecture comparison.

Once this gap is bridged, it becomes possible to move toward NLP systems that better address the complex indirectness that defines human communication in the Global South.

ACKNOWLEDGEMENTS

The authors thank the native speaker annotators whose cultural expertise was indispensable to the dataset validation process. We also thank the reviewers for their constructive feedback, which has substantially improved this work. The author specially thanks Dr. Ashwini Vaidya (IIT Delhi) for the intellectual grounding and guidance provided through her course on 'Computational Models of Meaning', which inspired the framing of this work.

BIBLIOGRAPHICAL REFERENCES

Allan, K. and Burridge, K. (1991). Euphemism and Dysphemism: Language Used as Shield and Weapon. Oxford University Press.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., et al. (2025). MME: A comprehensive evaluation benchmark for multimodal large language models. In Proceedings of the Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.

Grice, H. P. (1975). Logic and conversation. In Syntax and Semantics, Vol. 3: Speech Acts (pp. 41–58). Academic Press.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J. (2021). Measuring massive multitask language understanding. In Proceedings of the International Conference on Learning Representations (ICLR).

Hu, Y., Li, J., Wu, M., Huang, Z., Chen, G. and Sha, Y. (2024a). A unified generative framework for bilingual euphemism detection and identification. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 403–417). Association for Computational Linguistics.

Hu, Y., Li, J., Wu, M., Huang, Z., Chen, G. and Sha, Y. (2024b). Uncovering and mitigating the hidden chasm: A study on the text-text domain gap in euphemism identification. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 18270–18278. https://doi.org/10.1609/aaai.v38i16.29786

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Iyer, A., Khapra, M. M. and Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 4948–4961).

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., Gupta, S., Gali, S. C., Subramanian, V. and Talukdar, P. (2021). MuRIL: Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730.

Sailaja, P. (2012). Indian English. Edinburgh University Press.

Srivastava, V. and Singh, M. (2021). Challenges and considerations with code-mixed NLP for multilingual societies. arXiv preprint arXiv:2106.07823.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S. R. (2018). GLUE: A multi task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP (pp. 353–355). Association for Computational Linguistics.

Zeng, L. (2024). Leveraging large language models for code-mixed data augmentation in sentiment analysis. In Proceedings of the 2024 Symposium on Social Influence and Conversational AI (pp. 1–17).

Zhang, R. and Eickhoff, C. (2024). CroCoSum: A benchmark dataset for cross-lingual code switched summarization. In Proceedings of LREC-COLING 2024 (pp. 367–382). European Language Resources Association.