

Why Calques Look Rare in Hindi Model Output: A Human-Anchored Measurement Study

Abdur Rahman, Vasu Sharma, Mohammad Saifullah

Pocket FM

Correspondence: abdur75648@pocketfm.com

Abstract

Large language models write Hindi that carries the shape of English, and measuring how often is the open problem, because the phrase-level instruments available are lexical and brittle. We build a 26-pair Hindi contrast set over eight idiom families, verified by three native Hindi speakers. Prompt framing moves supplied-candidate scores toward the calque: an English-literal cue shifts them in 8 of 8 families, and a separate translate-from-English instruction shifts them again, with the direction holding on a family split frozen before the outcome was seen. Free generation appears to show far less calquing, so we audit the lexical indicator behind that appearance against native-speaker judgement. Its precision on its own positives is 1.00, while 24 of 60 sampled negatives are calques by majority verdict, and the misses concentrate in the families where its exact-match trigger was never observed to fire. The apparent scarcity is substantially a property of the instrument. We recommend that detectors of this form report coverage alongside hits.

1 Introduction

A model asked for Hindi can return sentences that are fluent and grammatical and still built on English scaffolding. Where a Hindi writer says रोंगटे खड़े हो गए, the model may assemble a literal rendering of *a shiver ran down his spine* out of Hindi words (Figure 1). Borrowing of this kind is a calque, catalogued as a contact phenomenon by Weinreich (1953) and traced through Hindi by Kachru (1989). What has changed is the scale and the source, and how much of it a model produces is now an evaluation problem (Li et al., 2025).

Related work. Detecting translationese is an old task, and its history says where the difficulty sits. Document-level classifiers do well: Ilisei et al. (2010) reach high accuracy on translated Spanish



Figure 1: What a calque is, and what the lexical calque indicator misses. The indicator requires complete coverage of the tokens that distinguish the two references, which is also what makes its precision near-perfect. Here it found four of five and scored the output negative; all three native Hindi speakers judged it a calque. Twenty-four of sixty sampled negatives fell short of their trigger set and were missed (Section 4).

from simplification features, though Rabinovich and Wintner (2015) show such detectors are sensitive to the domain they were fitted on. Both read aggregate statistics of a whole document. Our unit is one phrase inside two sentences of narrative, where those statistics do not exist, and matching against known realisations is what is left. Riley et al. (2020) treat natural and translated text as distinct modes at corpus level, and Liu et al. (2025) give the graded, likelihood-ratio alternative a phrase-level study might reach for instead of an exact-match rule.

Whether models lean toward translationese phrasing is a question prior work already owns. Kunz (2026) probes Swedish models with idiomatic-versus-translationese minimal pairs and finds a preference for idiomatic phrasing that is acquired slowly, holds only on length-controlled and unambiguous pairs, and is undone quickly by

instruction tuning on machine-translated data; on the unfiltered set the models score below chance and favour the literal form. The companion dataset of [Kunz et al. \(2026\)](#) reports the same direction for English-to-Swedish. [Zhou and Ackerman \(2026\)](#) describe the general shape, where a measurable preference fails to become a behaviour. Our framing results are a Hindi instance of this established phenomenon and we present them as such. The form a task takes shapes multilingual behaviour beyond lexical choice: [Ranaldi et al. \(2025\)](#) find that natural-language chain-of-thought demonstrations transfer poorly to lower-resource languages where program-style demonstrations hold up, and [Ranaldi and Pucci \(2025\)](#) improve multilingual step-wise reasoning by having a model restructure its reasoning and then self-refine. Our framing manipulations are the same kind of sensitivity, measured on a lexical outcome instead of a reasoning one.

We could not find anyone asking whether the production side of that gap is partly a property of the detector. A lexical calque indicator can be perfectly precise and still recover a small fraction of what a native speaker would call a calque, and if it does, a low measured rate of calquing in free generation is evidence about the instrument as much as about the model. We audit one such indicator for precision and recall against native judgement, on a Hindi contrast set three native speakers verified. The audit is the contribution; the contrast set is a secondary one, and it matters because Indic evaluation infrastructure is built for other tasks ([Kakwani et al., 2020](#); [Doddapaneni et al., 2023](#)) and offers nothing at this granularity. The closest Hindi neighbour, [Choudhary et al. \(2026\)](#), compares model perplexity against native judgement on Hindi-English code-mixed minimal pairs and finds the two come apart, which is part of why we anchor on human verdicts throughout.

2 A verified Hindi contrast set

The contrast set holds 26 Hindi item pairs across eight idiom families. Each pair puts a calque completion and a natural completion in the same short narrative context. The two members differ in the idiom and agree on what they say. Items were generated from a human-curated Hindi seed lexicon by an automated pipeline, then verified by three native Hindi speakers, who judged calque status, naturalness, and propositional equivalence for every item.

Set	<i>n</i>	Measure	Value
Contrast pairs	26	blind label recovery, majority	26/26
		propositional equivalence agreed	26/26
Detector positives	69	judged calque, unanimous precision	69/69 1.00
Detector negatives	60	judged calque, majority	24/60
Retest control	12	judged calque, majority three-way agreement	12/12 65/72

Table 1: Verification and detection summary. Contrast-pair rows check the items (Section 2); detector rows are the audit (Section 4). Recovery is by majority of three annotators; the retest control is 12 known positives re-shown unlabelled.

The source units are attested and the context prefixes shown to the model are adaptations of them, which is how the item metadata labels them. They are anchored to revision-pinned Premchand passages from Wikisource, public domain with a CC BY-SA transcription, and we record a revision ID and scan page per item.

Three native Hindi speakers annotated all items independently, with labels hidden, A/B order randomised, and no sight of one another’s answers. Individual blind recovery of the intended label was 25, 24, and 25 of 26; by majority it is 26 of 26, and all three agree exactly on 22 of 26 (Table 1). Asked whether the two members of a pair say the same thing, every annotator answered yes for all 26. Pairwise agreement between annotators runs 23 to 24 of 26.

Three items drew a flag that both members sounded off. All three are *adj* आवाज़ में voice-adjective constructions, so the weakness is localised to one construction. Appendix H prints them. Appendix H also reports a surface search of attested Premchand prose for the seed phrases.

3 Framing shifts, and the apparent absence of production

We score both members of each pair under Gemma 4 12B Instruct ([Gemma Team, 2026](#)) and Qwen3 14B ([Yang et al., 2025](#)), reporting the mean per-token log-probability of the calque completion minus that of the natural completion, in nats per token. Positive means the model scores the calque higher. Intervals are family-level bootstrap percentile intervals, $B = 10,000$.

3.1 Two framings move the score toward the calque

Prompt framing shifts the score. For Gemma 4 on the 15 development items, replacing a neutral English cue with an English-literal one moves the calque-minus-native score by +0.578 nats/token under a direct framing and +0.769 under an author framing, positive in 8 of 8 families in both cases (Table 2, rows 1 and 2). Qwen3 moves the same way and less far (+0.254, +0.290), with its direct-framing interval crossing zero. Both are differences of differences: they say how far the framing moves the model and nothing about where it starts. Under the neutral English cue the absolute calque-minus-native score is +0.161 and +0.162, positive in only 5 of 8 families, so the level is modest where the shift is uniform.

A second manipulation points the same way. Holding the cue fixed and changing the instruction from write-idiomatic-Hindi to translate-from-English moves the score by +0.276 on the families we developed on and +0.413 on a family split we had committed to before any of it was scored (Table 2, rows 3 and 4). The two splits share no families and both intervals exclude zero. The sign also survives dropping any single family. This is the only positively replicating result in the programme. The absolute levels again qualify it: on the held-out families the model scores the natural form higher under both instructions (−1.554 idiomatic, −1.141 translate), so what carries across the split is the direction of the shift. Cue literalness and instruction type are two separate manipulations of the prompt, and neither is a replication of the other.

3.2 The cue-language contrast keeps its sign after a rewrite

Annotators judged several of our original Hindi control cues grammatical but not idiomatic, which would have confounded the language of a cue with its naturalness. They rewrote thirteen and we re-ran with nothing else changed. The neutral-English against neutral-Hindi contrast keeps its sign in both framings (−0.1374 to −0.1260 direct, −0.2238 to −0.3106 author). Conditions whose cue we did not edit differ from the baseline by exactly 0.0, so nothing but the intended edit changed between runs. The English cues were never touched, which leaves the estimands above unperturbed by the rewrite; we make no claim that it stress-tested them. Appendix D has the proto-

Shift in calque – native score	Mean (nats/tok)	95% CI	Fam. > 0
literal-En – neutral-En cue, direct	+0.578	[+0.380, +0.785]	8/8
literal-En – neutral-En cue, author	+0.769	[+0.355, +1.214]	8/8
translate – idiomatic instr., development	+0.276	[+0.041, +0.534]	6/8
translate – idiomatic instr., held-out	+0.413	[+0.166, +0.682]	5/6

Table 2: Framing shifts in the calque-minus-native score, in nats per token, under Gemma 4. Each row is a difference between two prompt conditions, so a positive value means the first moves the model toward the calque. Rows 1 and 2 vary cue literalness; rows 3 and 4 vary instruction type on disjoint family splits. Intervals are family-level bootstrap percentile intervals, $B = 10,000$, seed 20260727. Per-family values: Appendix C.

col, the thirteen pairs, and the leave-one-family-out check.

3.3 Free generation appears not to follow

Under free generation the lexical calque indicator of Section 4 fires rarely. Across the decoder matrix, greedy decoding yields hits in 3 of 8 and 4 of 8 families for Gemma 4, and Qwen3 sampled reaches 2 of 8 and 1 of 8; all eight predeclared family coverage criteria fail. Truncation and stemming do not explain it: a stemmer ablation changes 0 of 600 labels, and raising Qwen3’s cap from 96 to 256 gives 83 of 83 valid sampled terminations and 23 of 23 greedy with the criteria still failing. These counts are indicator firings, and whether they are calque rates is the question of the next section.

4 What the indicator measures

The lexical calque indicator works per item. For a given pair it takes the stemmed tokens present in the calque reference and absent from the natural one, typically one to three words, and flags a generated output only if the output reproduces that entire set. We put its verdicts in front of the same three native Hindi speakers, blind.

4.1 Precision

All three annotators called every one of the 69 flagged outputs English-shaped. Precision on the indicator’s own positives is therefore 1.00, a figure close to tautological given how narrowly it selects. It earns its place here for one reason: whatever the indicator finds, it finds correctly, so any shortfall

against human judgement is a shortfall in coverage. The same narrowness that makes this precision close to tautological is what bounds the recall reported below.

4.2 Recall

We drew 60 indicator negatives, stratified across four model-by-decoder cells and all eight families, and mixed in 12 known positives re-shown unlabelled as a control against annotators defaulting to no. Every annotator marked all 12, so they were discriminating. Of the 60 negatives, **24 are calques by majority verdict and 19 are unanimous**. Three-way agreement across the whole 72-card task is 65 of 72. Projecting that rate onto the 531 negatives in the sampled frame gives an estimated recall of 0.233, 95% bootstrap interval [0.184, 0.317]. We call that a cell-weighted projection and treat the observed 24 of 60 as the measurement, because the selection procedure, its seed, and the inclusion probabilities were not recorded, so the design cannot be reconstructed. The weighting choice barely moves it, and thirty separately audited Hindi-control rows sit outside the frame and are excluded. Six of the 24 come from a family whose trigger is the single stem नीच, narrow enough that the natural reference form scores negative too; discounting all six leaves 18 of 60 and a projection near 0.28.

4.3 The misses are concentrated

The shortfall is not spread evenly. In the three families where no exact-match hit was observed anywhere in the 600 decoder-matrix outputs, native speakers judged 16 of 23 sampled negatives to be calques, a miss rate of 70%; where the trigger did fire at least once it is 8 of 37, or 22% (Figure 2). Because each item’s trigger is exact-match to the words of one reference realisation, an output rendering the same English idiom through different Hindi words scores negative, and in those three families no output reproduced the required set. We report zero observed hits and make no claim that a hit was impossible.

Appendix B prints all 24 misses with their required, found, and missing tokens, plus glossed exemplars. Figure 1 is one of them, from fear shiver, a family where the indicator does fire elsewhere, so that miss is not a matter of the family going uncovered.

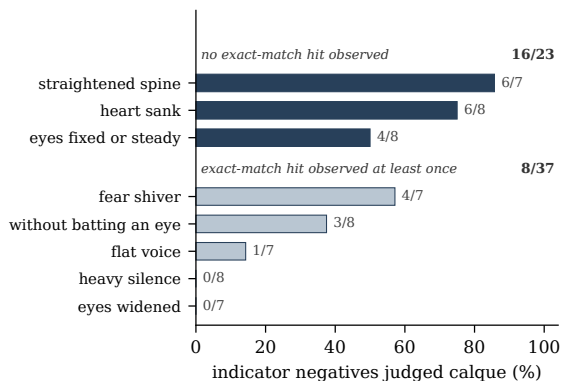


Figure 2: Sampled indicator negatives that native speakers judged calques, by idiom family, grouped by whether the indicator’s trigger was ever observed to fire in that family across the 600 decoder-matrix outputs. Bar labels give the raw fraction, bold figures the pooled fraction per group.

4.4 Consequence

Part of the gap in Section 3 belongs to the instrument. How the remainder divides between instrument and model behaviour is not something this study estimates, and we claim nothing about it. An indicator of this form should carry its coverage wherever it reports its hits, and until its recall has been measured against human judgement its counts will not support a claim about how often a model writes calques. A graded measurement answers a question this instrument cannot (Liu et al., 2025).

5 Conclusion

Native speakers found calques in 24 of the 60 outputs our indicator passed over, clustered in the families whose trigger tokens no generated output reproduced. Prompt framing reliably moves the score toward the calque, and what reaches free generation is read through an instrument whose recall nobody had measured. Measure it, and report coverage with every count.

6 Limitations

The contrast set holds 26 pairs in eight idiom families, scored on two models in one narrative domain and annotated by three native speakers. The audited indicator is our own, and our claims cover it and others built on the same per-item exact-match rule; extending them to lexical calque detection in general is an argument we have not made. Recall rests on a 60-item sample whose selection procedure was not recorded, which is why we report the

observed fraction first and label the projection.

We report no successful mitigation and no causal control. Six inference-time branches produced no usable reduction without damage to the output, and the attention head that best localises calque preference moves the score less than a matched orthogonal control does. Appendix F records both, with the screening procedure and the control arms. Twenty runs failed to execute for operational reasons and none is reported as a null result.

Ethical considerations

The narrative contexts derive from public-domain literary prose, and no item contains personal or sensitive information. Annotation was carried out by native speakers of Hindi participating as colleagues in the course of their work, on material that presents no foreseeable risk to them. The indicator we audit is a measurement instrument, used here to count and never to filter or moderate output, and our finding is a caution against reading its counts as a rate at which models produce calques.

References

- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- Mukund Choudhary, Madhur Jindal, Gaurja Aeron, and Monojit Choudhury. 2026. [Do LLMs model human linguistic variation? a case study in Hindi-English verb code-mixing.](#) In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5491–5509, Rabat, Morocco. Association for Computational Linguistics.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. [Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.
- Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O’Brien. 2025. [Improving LLM abilities in idiomatic translation.](#) In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Bryan Eikema and Wilker Aziz. 2022. [Sampling-based approximations to minimum Bayes risk decoding for neural machine translation.](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. [Quality-aware decoding for neural machine translation.](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.
- Gabriel Franco, Lucas M. Tassis, Azalea Rohr, and Mark Crovella. 2026. [Finding interpretable prompt-specific circuits in language models.](#) *Preprint*, arXiv:2602.13483.
- Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M., Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. [IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages.](#) *Transactions on Machine Learning Research*.
- Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. [Evaluating models’ local decision boundaries via contrast sets.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Gemma Team. 2026. [Gemma 4 technical report.](#) *Preprint*, arXiv:2607.02770.
- Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. [Statistical power and translationese in machine translation evaluation.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.
- Sarah Griebel and Ted Underwood. 2026. [Fluency and faithfulness in human and machine literary translation.](#) In *Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities*, pages 178–189, San Diego, USA. Association for Computational Linguistics.
- Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. [Do large language models have an English “Accent”? evaluating and improving the naturalness of multilingual LLMs.](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838, Vienna, Austria. Association for Computational Linguistics.
- Jennifer Hu and Roger Levy. 2023. [Prompting is not a substitute for probability measurements in large language models.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.
- Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. [Identification of translationese: A machine learning approach.](#) In *Computational Linguistics and Intelligent Text Processing (CICLing 2010)*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511, Berlin, Heidelberg. Springer.
- Abigail Z. Jacobs and Hanna Wallach. 2021. [Measurement and fairness.](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385. Association for Computing Machinery.
- Yamuna Kachru. 1989. [Corpus planning for modernization: Sanskritization and Englishization of Hindi.](#) *Studies in the Linguistic Sciences*, 19(1):153–164.

- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. [Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Raghvendra Kumar, Devankar Raj, and Sriparna Saha. 2026. [BhashaSutra: A task-centric unified survey of Indian NLP datasets, corpora, and resources](#). *Preprint*, arXiv:2604.18423. Accepted to ACL 2026 (Main Conference).
- Jenny Kunz. 2026. [Preferences for idiomatic language are acquired slowly – and forgotten quickly: A case study on Swedish](#). *Preprint*, arXiv:2602.03484. Accepted to TACL.
- Jenny Kunz, Anja Jarochenko, and Marcel Bollmann. 2026. [A dataset for probing translationese preferences in English-to-Swedish translation](#). *Preprint*, arXiv:2603.08450.
- Théo Lasnier, Armel Zebaze, Djamé Seddah, Rachel Bawden, and Benoît Sagot. 2026. [Translation heads: Disentangling meaning from language in LLM-based machine translation](#). *Preprint*, arXiv:2602.04613.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. [Contrastive decoding: Open-ended text generation as optimization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.
- Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. [Lost in literalism: How supervised training shapes translationese in LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria. Association for Computational Linguistics.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. [DExperts: Decoding-time controlled text generation with experts and anti-experts](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.
- Yikang Liu, Wanyang Zhang, Yiming Wang, Jialong Tang, Pei Zhang, Baosong Yang, Fei Huang, Rui Wang, and Hai Hu. 2025. [Translationese-index: Using likelihood ratios for graded and generalizable measurement of translationese](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12521–12538, Suzhou, China. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of what art? a call for multi-prompt LLM evaluation](#). *Transactions of the Association for Computational Linguistics*, 12:933–949.
- Ella Rabinovich and Shuly Wintner. 2015. [Unsuper-vised identification of translationese](#). *Transactions of the Association for Computational Linguistics*, 3:419–432.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. [When natural language is not enough: The limits of in-context learning demonstrations in multilingual reasoning](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7384–7411, Albuquerque, New Mexico. Association for Computational Linguistics.
- Leonardo Ranaldi and Giulia Pucci. 2025. [Multilingual reasoning via self-training](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11566–11582, Albuquerque, New Mexico. Association for Computational Linguistics.
- Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. [Explainability and interpretability of multilingual large language models: A survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20454–20486,

- Suzhou, China. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. [Translationese as a language in “Multilingual” NMT](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.
- Aaquib Syed, Can Rager, and Arthur Conmy. 2024. [Attribution patching outperforms automated circuit discovery](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. [Investigating gender bias in language models using causal mediation analysis](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Uriel Weinreich. 1953. *Languages in Contact: Findings and Problems*. Number 1 in Publications of the Linguistic Circle of New York. Linguistic Circle of New York, New York.
- Frank Zhengqing Wu, Francesco Tonin, and Volkan Cevher. 2026. [Demystifying variance in circuit discovery of LLMs](#). *Preprint*, arXiv:2606.16920.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Hongbin Zhang, Kehai Chen, Xuefeng Bai, Youcheng Pan, Yang Xiang, Jinpeng Wang, and Min Zhang. 2026. [Mitigating translationese bias in multilingual LLM-as-a-judge via disentangled information bottleneck](#). *Preprint*, arXiv:2603.10351. Accepted to ICML 2026.
- Shukang Zhang and Chaoyong Zhao. 2026. [Machine translationese of large language models: Dependency triplets, text classification, and SHAP analysis](#). *PLOS One*, 21(1):e0339769.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Yujun Zhou and Christopher M. Ackerman. 2026. [When preferences fail to become incentives: A utility–behavior gap in large language models](#). *Preprint*, arXiv:2606.22974.

A Annotation protocol

Three native Hindi speakers annotated all items independently. None saw another annotator’s answers, and none was shown the intended label. Three instruments were used.

Blind pair judgement. For each of the 26 contrast pairs the annotator saw the shared context and the two completions in randomised A/B order, and answered two questions: which of the two reads as English-shaped Hindi, and whether the two say the same thing. The first carried a third option for an annotator who judged that both members sounded wrong, which is how the naturalness flag in Section 2 was recorded. The propositional-equivalence question is what licenses treating the pair as a minimal contrast at all.

Detector-positive review. All 69 outputs the indicator flagged were shown without the indicator’s verdict, and the annotator judged whether the main descriptive idea used English-shaped phrasing.

Recall sample with retest controls. Sixty indicator negatives, stratified across four model-by-decoder cells and all eight families, were mixed with 12 known indicator positives re-shown unlabelled. Annotators were asked to judge one phrase and not the whole sentence. The retest items are the control against an annotator answering no by default; every annotator marked all 12.

Rater identifiers are not linked across the two annotation rounds, so we report per-round agreement and never a per-annotator trajectory.

The design follows the practice of stating what a benchmark operationalises before reporting a number on it (Jacobs and Wallach, 2021; Bowman and Dahl, 2021), and the contrast-pair format follows the minimal-pair tradition in linguistic evaluation (Warstadt et al., 2020; Ribeiro et al., 2020). On the risk of substituting a model score for a human verdict see Gardner et al. (2020) and, for judge models

specifically, Zheng et al. (2023) and Zhang et al. (2026).

B What the indicator missed

Table 3 lists all 24 indicator negatives that annotators judged calques, with the tokens each item’s trigger required and the tokens the output failed to supply. Nineteen are unanimous. Six families are represented and two are absent, which matches Figure 2.

Table 4 glosses five of them. Two are worth reading together. Examples 3 and 4 share the English cue *a cold shiver ran down his spine at the sight* and differ only in model. Qwen3 wrote ठंडा झटका (*a cold jolt*) reaching छाती (*the chest*), which supplies 1 of the 5 required tokens. Gemma 4 wrote एक ठंडी सिहरन running through शरीर (*the body*) where the reference has रीढ़ की हड्डी (*spine-bone*), which supplies 3 of 5 and leaves रीढ़ and हड्डी missing. Both are English-shaped by annotator verdict and both score negative, by different routes. A third, example 5, is a miss of a different kind: the output token is two characters long, and tokens shorter than three characters are dropped before stemming, while the reference’s four-character form stems down and stays in the required set. Two of the 24 misses arise this way. We name it because a reviewer recomputing our numbers would find it, and because it is a property of the instrument of exactly the kind this paper is about.

The dominant failure mode is plainer. Eighteen of the 24 are ordinary paraphrase, where the model renders the English idiom through Hindi words the reference pair never used; most of the remainder are near misses of the kind in Figure 1. A graded measure of translationese (Liu et al., 2025) would score these on a continuum, and corpus-level naturalness measures (Guo et al., 2025) would register them in aggregate, where an exact-match rule returns a hard zero. Zhang and Zhao (2026) and Griebel and Underwood (2026) characterise translationese in model output more broadly, and Donthi et al. (2025) treat the Hindi and Urdu idiom case. On reporting a measure whose error is asymmetric, Graham et al. (2020) and Kaster et al. (2021) are the relevant guidance; Resck et al. (2025) argue for saying what an evaluation measure does and does not see.

C Per-family values

Table 5 gives the cue shift and the absolute level behind every aggregate in Table 2. The macro means over the eight families reproduce the body figures exactly: +0.578 and +0.769 for the shift, +0.161 and +0.162 for the level. The shift is positive in all eight families under both framings. The level is positive in five of eight under each, and the three families where it is negative are the same three under both.

Two disclosures belong here. The shift is quoted against the neutral-English cue; against the alternative neutral-Hindi cue the same contrast gives +0.440 (6 of 8 families) and +0.545 (7 of 8), so the neutral-English control is the more favourable of the two and we name it wherever the number appears. Rows 1 and 2 of Table 2 are computed on the 15 development items under Gemma 4; Qwen3 gives +0.254 direct and +0.290 author, with the direct interval crossing zero.

D Cue rewrite record

Annotators judged several of the original Hindi control cues grammatical but not idiomatic. Thirteen were rewritten by the same three speakers and are given in Table 6. A rewrite was accepted only if it preserved the propositional content of the original and introduced none of the words the indicator matches on, so that the edit could not create or destroy a detection by itself.

The re-run changed the Hindi cue conditions and nothing else. Conditions built on the unedited English cues returned values identical to the baseline to every recorded digit, which is the check that no unintended difference entered with the edit. Prompt wording moves measured behaviour on its own (Hu and Levy, 2023; Mizrahi et al., 2024), which is the reason for running the control at all.

Figure 3 gives the contrast family by family. Two things in it qualify the aggregate quoted in Section 3. The per-family values are mixed in sign, and the reported figure is a macro mean over families: under the direct framing 2 of 8 families are positive before the rewrite and 3 of 8 after. The author-frame aggregate also grows by about 39%, from −0.2238 to −0.3106, while holding its sign. Section 3 claims the sign survives, which is a weaker claim than the rewrite having changed nothing.

Family	Model	Dec.	Cov.	Required tokens	Missing
eyes fixed or steady	gemma	sam	1/3	आँख थीं स्थिर	थीं स्थिर
eyes fixed or steady	gemma	gre	2/3	आँख थीं स्थिर	स्थिर
eyes fixed or steady	gemma	sam	1/3	आँख थीं स्थिर	थीं स्थिर
eyes fixed or steady	qwen	sam	1/3	आँख थीं स्थिर	थीं स्थिर
fear shiver	gemma	sam	3/5	ठंड दौड़ रीढ़ सिहरन हड्डु	रीढ़ हड्डु
fear shiver	qwen	gre	1/5	ठंड दौड़ रीढ़ सिहरन हड्डु	दौड़ रीढ़ सिहरन हड्डु
fear shiver	qwen	sam	1/5	ठंड दौड़ रीढ़ सिहरन हड्डु	दौड़ रीढ़ सिहरन हड्डु
fear shiver	gemma	sam	4/5	ठंड दौड़ रीढ़ सिहरन हड्डु	हड्डु
flat voice	qwen	gre	0/1	सपाट	सपाट
heart sank	gemma	gre	0/1	नीच	नीच
heart sank	qwen	sam	0/1	नीच	नीच
heart sank	gemma	sam	0/1	नीच	नीच
heart sank	qwen	sam	0/1	नीच	नीच
heart sank	qwen	gre	0/1	नीच	नीच
heart sank	qwen	gre	0/1	नीच	नीच
straightened spine	qwen	sam	1/3	चल रीढ़ सीध	चल रीढ़
straightened spine	qwen	gre	1/3	चल रीढ़ सीध	रीढ़ सीध
straightened spine	gemma	gre	2/3	चल रीढ़ सीध	चल
straightened spine	gemma	gre	0/3	चल रीढ़ सीध	चल रीढ़ सीध
straightened spine	gemma	sam	2/3	चल रीढ़ सीध	चल
straightened spine	gemma	sam	0/3	चल रीढ़ सीध	चल रीढ़ सीध
without batting an eye	qwen	sam	1/3	झपक पलक बि	झपक पलक
without batting an eye	qwen	gre	1/3	झपक पलक बि	झपक पलक
without batting an eye	qwen	gre	2/3	झपक पलक बि	पलक

Table 3: All 24 indicator negatives that annotators judged calques by majority. *Cov.* is the fraction of the item’s required token set the output supplied; the indicator fires only at complete coverage.

No.	Model output, indicator verdict, and annotator verdict
1	उसकी आँखें उस दिशा में ही जमी रहीं। <i>his eyes stayed set in that very direction</i> cue: <i>His eyes remained fixed in that direction.</i>; required आँख थीं स्थिर, supplied 1/3, missing थीं स्थिर; indicator negative, annotators 3-0 calque
2	उसकी आँखें किनारे पर टिकी रहीं। <i>his eyes remained resting on the shore</i> cue: <i>His eyes stayed fixed on the shore.</i>; required आँख थीं स्थिर, supplied 1/3, missing थीं स्थिर; indicator negative, annotators 3-0 calque
3	उसे फ़तहचंद के सामने डंडा सँभाले खड़ा देखते ही ठंडा झटका उसकी छाती तक पहुँच गया। <i>seeing him standing before Fatehchand holding a stick, a cold jolt reached as far as his chest</i> cue: <i>A cold shiver ran down his spine at the sight.</i>; required ठंड दौड़ रीढ़ सिहरन हड्डु, supplied 1/5, missing दौड़ रीढ़ सिहरन हड्डु; indicator negative, annotators 3-0 calque
4	साहब ने फ़तहचंद को सामने डंडा सँभाले और पक्का इरादा किए खड़ा देखा, जिसे देखते ही उनके शरीर में एक ठंडी सिहरन दौड़ गई। <i>seeing which, a cold shiver ran through his body</i> cue: <i>A cold shiver ran down his spine at the sight.</i>; required ठंड दौड़ रीढ़ सिहरन हड्डु, supplied 3/5, missing रीढ़ हड्डु; indicator negative, annotators 3-0 calque
5	वह अपनी रीढ़ सीधी करके चल रहा था। <i>he was walking having made his spine straight</i> cue: <i>He walked with his spine straight.</i>; required चल रीढ़ सीध, supplied 2/3, missing चल; indicator negative, annotators 3-0 calque

Table 4: Glossed exemplars, numbered as the text refers to them. Examples 3 and 4 share an English cue and are missed by different routes; example 5 is one of the two token-length cases described in the text.

Family	cue shift		level, neutral-En	
	direct	author	direct	author
eyes fixed or steady	+1.078	+1.075	+0.808	+0.132
eyes widened	+0.180	+0.303	+1.623	+1.118
fear shiver	+0.732	+1.410	+0.471	+1.225
flat voice	+0.745	+0.337	+0.132	+0.281
heart sank	+0.414	+0.390	-0.127	-0.411
heavy silence	+0.394	+0.325	-1.904	-1.617
straightened spine	+0.260	+1.975	-0.169	-0.051
without batting an eye	+0.819	+0.337	+0.450	+0.616

Table 5: Per-family calque-minus-native values in nats per token, Gemma 4. *Cue shift* is the literal-English cue minus the neutral-English cue. *Level* is the absolute score under the neutral-English cue.

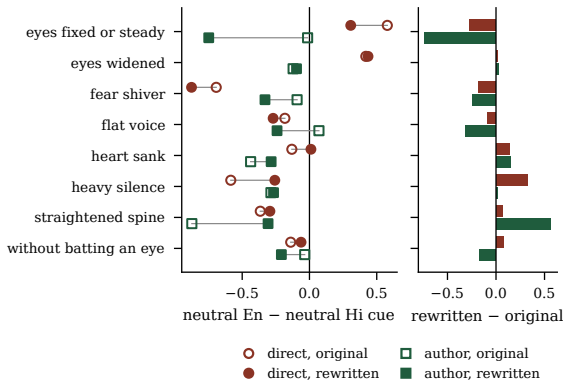


Figure 3: Cue-language contrast by family, before and after thirteen non-idiomatic Hindi control cues were rewritten by the annotators. Left panel: the neutral-English minus neutral-Hindi score in nats per token, hollow markers before the rewrite and filled after. Right panel: the change itself. Conditions built on the unedited English cues are identical between the two runs, and the original baseline was reproduced byte-for-byte before either run was compared.

E Reproduction record

We re-ran the original experiment before the rewritten cues went in. Its output reproduced the stored baseline byte-for-byte under SHA-256. Scoring ran on Gemma 4 12B Instruct and Qwen3 14B with greedy and sampled decoding on a single A100-class device, with the model weight hash, package versions, and configuration hash recorded per run. Every number in this paper regenerates from the released artifacts with a single script, together with nine independent verification scripts that recompute the frozen study’s figures from raw run data.

F Mechanism and mitigation negatives

These results do not support the paper’s claim, and we record them so the released artifacts are not read as hiding them.

Localisation without control. Following the activation-patching and head-attribution literature (Vig et al., 2020; Meng et al., 2022; Syed et al., 2024), we screened all 768 attention heads of Gemma 4 with an analytic max-statistic test on cached activations. Thirteen passed a localisation gate. The strongest, L16H13 at $p = 0.00293$, was carried to a hooked intervention against 19 controlled arms covering four dose levels together with shuffled, orthogonal, wrong-layer, wrong-position, and dense-baseline controls and split-half checks. Its true +1 shift is +0.0049 nats/token while a matched orthogonal control reaches +0.0110, and the three orthogonal controls average +0.0069. Decodability without control is the outcome, which is the pattern reported for prompt-specific and template-specific circuits generally (Wu et al., 2026; Franco et al., 2026). A positive causal result for a different factor is given by Lasnier et al. (2026). Scope is one model and one head carried to full intervention.

Suppression without repair. Counterfactual decoding in the style of contrastive and expert-ensemble methods (Liu et al., 2021; Li et al., 2023) drove indicator positives from 12/30 to 5/30 to 0/30 across two doses, while observed invariant failures reached 9/30 and 17/30. Perturbation strengths across arms are unmatched and baseline invariants were not assessed, so we report a trade-off and not a causal harm estimate. Quality-aware and utility-based decoding (Fernandes et al., 2022; Eikema and Aziz, 2022) were not tried.

Why every mitigation failed. A headroom analysis explains the pattern. 92 of 96 candidates are indicator-positive and 11 of 12 cells contain no alternative to move toward, so six inference-time branches frequently had nothing to move to.

Twenty runs in the wider programme failed to execute for operational reasons. None of them measured an absence of effect and none is reported here.

G A superseded result

An earlier version of this work reported a reversal on a changed-span decomposition of the held-out set. It does not survive. The effect is carried almost entirely by cold voice, a single-item fam-

Family	Original Hindi cue	Rewritten by annotators
heavy silence	सवाल के बाद किसी ने कोई आवाज़ नहीं की।	प्रश्न के बाद वहाँ से कोई आवाज़ नहीं आई।
heavy silence	बँगले के आसपास कोई आवाज़ सुनाई नहीं दी।	बँगले के आसपास कुछ भी सुनाई नहीं दे रहा था।
eyes fixed or steady	वह लगातार उसी दिशा की ओर देख रहा था।	वह उसी तरफ़ लगातार देखता रहा।
eyes fixed or steady	वह लगातार दूर जाती जमीन की सीमा की ओर देख रहा था।	वह लगातार किनारे की ओर देखता रहा।
without batting an eye	उसकी प्रतिक्रिया तुरंत हुई और उसमें कोई हिचक नहीं थी।	जवाब देने में उसे एक पल की भी देर नहीं लगी।
eyes widened	उसके चेहरे पर आश्चर्य साफ़ दिखाई दिया।	उसके चेहरे पर हैरानी साफ़ झलक रही थी।
eyes widened	खबर सुनकर उसका आश्चर्य साफ़ दिखाई दिया।	खबर सुनते ही उसकी हैरानी चेहरे पर झलक उठी।
straightened spine	उसकी चाल में डर की जगह दृढ़ निश्चय दिख रहा था।	उसकी चाल अब पहले जैसी सहमी हुई नहीं, बल्कि भरोसे से भरी थी।
heart sank	नाम सुनते ही वह अचानक निराश हो गया।	नाम सुनते ही उसकी उम्मीद टूट गई।
heart sank	क्रोध भरा सवाल सुनते ही वह अचानक निराश हो गया।	गुस्से भरा सवाल सुनते ही उसकी उम्मीद टूट गई।
flat voice	उसने बिना कोई भाव प्रकट किए तथ्य बताया।	उसने बिना किसी हाव-भाव के यह बात कह दी।
flat voice	उसने बुलाए जाने का कारण जानना चाहा और कोई भाव प्रकट नहीं किया।	उसने बिना किसी हाव-भाव के पूछा कि उसे क्यों बुलाया गया है।
fear shiver	डर से उसके पूरे शरीर में अचानक कंपकंपी हुई।	वह डर से अचानक थर-थर काँपने लगा।

Table 6: All thirteen Hindi control cues that annotators rewrote, before and after. The experiment was re-run with nothing else changed.

ily, and leave-one-family-out drops it to -0.106 and flips its sign, with all three bootstrap intervals crossing zero. The development set shows the mirror sign pattern. We dropped the result and record it here because it is recoverable from the released artifacts.

H Item-level data

Table 7 gives the three items annotators flagged as having both members sound off; the 26 contrast pairs are released in full with the annotation record. All three are voice-adjective constructions of the form *adj* आवाज़ में, where the calque and natural members differ only in the adjective and the natural adjective is itself contestable. We print them so the flag can be checked instead of taken on trust.

We also searched 152,673 characters of attested Premchand narrative for the 129 seed phrases of the frozen idiom map, and none appear. Those 129 cover the natural alternatives as well as the calques, and neither side matched, so the search bounds what this small historical corpus attests and says nothing that separates a calque from its natural counterpart. The run’s own sufficiency threshold, at least five families with a natural-form match, was not met, and we report it on that footing. In-

dic evaluation resources at other granularities are surveyed by [Kakwani et al. \(2020\)](#), [Ramesh et al. \(2022\)](#), [Gala et al. \(2023\)](#), and [Kumar et al. \(2026\)](#).

Calque completion	Natural completion
उसने सपाट आवाज़ में कहा, “तू मेरा खोया हुआ भाई है।”	उसने सूखी आवाज़ में कहा, “तू मेरा खोया हुआ भाई है।”
उसने सपाट आवाज़ में पूछा, “मुझे क्यों बुलाया है?”	उसने बेरंग आवाज़ में पूछा, “मुझे क्यों बुलाया है?”
उसने ठंडी आवाज़ में कहा, “पानी अब कल मिलेगा।”	उसने रूखी आवाज़ में कहा, “पानी अब कल मिलेगा।”

Table 7: The three items annotators flagged. Each pair differs only in the adjective modifying आवाज़ (*voice*).