Understood in Hindi, answered worse

Middle-layer alignment and answer quality for health questions in English, Hindi, romanized Hindi and Hinglish

tp53 `inventcures.github.io`

18 November 2026 Draft 0.3, with two controls run and awaiting grading. Claude drafted the text from the study's outputs; four models complete, Sarvam-105B on MedQA only.

Abstract

Open LLMs line up Hindi text with its English translation in their middle layers. I ask whether that alignment predicts how well a model answers a Hindi question. I posed 100 HealthBench consensus questions, 45 HealthBench cancer questions and 284 MedQA (USMLE) questions to four open models (Qwen3-1.7B, Qwen3.8-27B, Gemma 4 31B, gpt-oss-120b) in four forms: English, Devanagari Hindi, romanized Hindi and Hinglish. Gemini 3.8 Flash graded the HealthBench answers against the physician rubrics, and I scored MedQA by exact letter. Before each model wrote its answer, I recorded its residual stream at every layer. I find three things. (1) Every model answers worse in Hindi. The mean HealthBench drop from English is 0.03 for Gemma 4 and 0.10–0.20 for Qwen3.8-27B and gpt-oss-120b, and Qwen3-1.7B's Hindi mean falls below zero. Romanized Hindi is the hardest form for every model. (2) Every model identifies the question. At its peak layer, 93–100% of Hindi questions retrieve their own English twin among 100 (chance 1%). (3) Per-question alignment with the English twin barely predicts the per-question score drop. Spearman ρ at the peak layer lies between 0 and 0.25, with bootstrap intervals that mostly include zero. So the loss happens after the model has matched the question to its English twin. On MedQA, accuracy in romanized Hindi falls by 13–20% of English accuracy for the three larger models, about half the relative HealthBench loss for gpt-oss and Qwen3.8, and alignment again barely predicts correctness (ρ 0.03–0.10). Two limits matter most. Qwen3-32B wrote every non-English question and no native speaker has checked them, and Gemini 3.8 Flash alone graded every answer. To separate model, translation and grader effects I ran two controls, an answer-in-English condition (800 answers, 100% compliant) and an English translation of 900 Hindi answers for regrading. Both are run and await grading.

1 Introduction

Work on Llama, Claude and other models describes three phases. Early layers track surface form, middle layers share a space across languages, late layers return to one language's tokens [2, 3, 4]. My companion study [1] found the same rise and fall for Hindi in five open models. An English conversation's nearest Hindi state is its own translation 55% (Qwen3-1.7B) to 99% (Gemma 4) of the time at mid-depth.

A shared middle space suggests that a model which represents a Hindi question like its English twin should answer it like the English twin. If so, mid-layer alignment would be a cheap, label-free predictor of answer quality in a low-resource form, and a target for intervention. I test this on health questions, where a worse answer can hurt someone, against four hypotheses for each model and question. **H1** represented well, answered well; **H2** represented well, answered poorly; **H3** represented poorly, answered poorly; **H4** answer quality proportional to alignment.

Contributions.

1. A four-form parallel set of 429 health questions (HealthBench general and oncology with physician rubrics; MedQA with oncology and specialty-stratified general items), with code to regenerate it.
2. Rubric-graded answers and pre-generation states at every layer for four open models.
3. Evidence that cross-lingual retrieval of the question is near-saturated while answer quality drops by up to 0.4 rubric points, and that per-question alignment explains at most about 6% of the variation in the drop.

2 Data

HealthBench general. From HealthBench [5] I kept English, single-turn questions of 60–900 characters with a physician rubric, dropped the `health_data_tasks` theme (note-writing rather than patient questions), took consensus-subset questions first, and balanced the six remaining themes of communication, complex responses, context seeking, emergency referrals, global health and hedging. This gives 100 questions with 1,238 rubric items.

HealthBench oncology. The same filter restricted to questions matching a list of cancer terms: 45 questions, 39 from the consensus subset.

MedQA. From the 1,273-question four-option US test split [6]: every question matching the cancer term list (188; 186 kept after translation checks), plus 98 general questions drawn round-robin across 21 specialties labelled by Gemini. Gemini labels 30 of the keyword-matched questions as mainly oncology ("strict oncology"); the rest touch cancer in passing.

Forms. Qwen3-32B wrote everyday Devanagari Hindi and Hinglish from English, then romanized Hindi from the Devanagari, with instructions to keep drug names, doses, units and lab values verbatim. A version was kept only if its script share and length ratio passed checks; a question was kept only if all four forms passed. Option letters A–D stay in Latin script. For example, the instruction appended to every MedQA prompt is "Answer with the letter of the correct option only" in English and "केवल सही विकल्प का अक्षर लिखकर उत्तर दें।" in Devanagari.

3 Method

Models. Qwen3-1.7B, Qwen3.8-27B, Gemma 4 31B (QAT w4a16) and gpt-oss-120b, through Hugging Face `transformers` with the chat template, thinking disabled (low reasoning effort for gpt-oss, final channel only), greedy decoding, at most 1,536 new tokens (512 for MedQA). I planned Sarvam-105B too; it answered only MedQA (Section 7).

States. For each question I render the chat with the generation prompt and record the residual stream at every layer (embeddings, each block, final norm), averaged over the question's own tokens. Let $h^f_{q,\ell}$ be this state for question q in form f at layer ℓ. I centre within form, $\tilde{h}^f_{q,\ell} = h^f_{q,\ell} - \frac{1}{n}\sum_{q'} h^f_{q',\ell}$, and normalise.

Retrieval and alignment. For a Hindi form f, *alignment* is $a_{q,\ell} = \cos(\tilde{h}^f_{q,\ell}, \tilde{h}^{\mathrm{en}}_{q,\ell})$. *Retrieval* at layer ℓ is the fraction of questions whose nearest English state is their own twin; chance is $1/n$. The *peak layer* $\ell^\star$ maximises retrieval.

Grading. Gemini 3.8 Flash graded each HealthBench answer against all its rubric items in one call; the per-item prompt follows HealthBench [5], with a note to judge the answer in whatever language it is written. The score is the points of met items over the total positive points, so negative items can push it below zero; means are clipped to $[0, 1]$ as in the original. In 1,200 grouped calls for Qwen3-1.7B, Qwen3.8-27B and gpt-oss-120b, 25 rubric items failed to parse, under 1% of items; Gemma 4's 400 calls had none. A per-item check on a 300-answer sample is reported in the next draft. The grader is not GPT-4.1, so scores are comparable within this study only. MedQA answers are scored 1 if the first option letter found matches the key.

Analysis. The per-question *gap* is $g_q = s_q^f - s_q^{\text{en}}$. I report Spearman $\rho(a_{\cdot,\ell^\star}, g)$ with a 2,000-sample bootstrap interval over questions. The next draft will add the per-layer profile.

4 Results

4.1 Every model answers worse in Hindi

	English	Hindi	Roman Hindi	Hinglish
Qwen3-1.7B	0.25	0.00 (−0.37)	0.00 (−0.40)	0.00 (−0.31)
Qwen3.8-27B	**0.55**	0.46 (−0.10)	0.36 (−0.20)	0.45 (−0.11)
gpt-oss-120b	0.49	0.32 (−0.17)	0.30 (−0.20)	0.36 (−0.13)
Gemma 4 31B	0.49	**0.46** (−0.03)	**0.41** (−0.08)	**0.46** (−0.03)

Table 1: Gemma 4 loses least in Hindi; the best English model is not the best Hindi one. HealthBench score on 100 general questions (clipped mean), with the mean per-question gap from English in brackets (unclipped). Qwen3-1.7B's unclipped Hindi means are −0.06 to −0.15.

Table 1 shows a tenfold spread in the Hindi penalty. Romanized Hindi is the lowest-scoring form for all four models; Hinglish is within 0.01 of Devanagari or above it. Qwen3-1.7B's Hindi answers meet fewer positive than negative rubric points on average.

4.2 Every model identifies the question

At each model's peak layer, retrieval of the English twin is 93–100% in every Hindi form (chance 1%). Matching the question to its English twin is saturated even for Qwen3-1.7B, whose Hindi answers score zero. This shows the models recognise which question was asked. It does not show clinical comprehension.

4.3 Alignment barely predicts the gap

At the peak layer, Spearman ρ between alignment and gap is between 0 and 0.25 for most model–form pairs, and most bootstrap intervals include zero. The two clearly positive cases are Gemma 4 on romanized Hindi ($\rho = 0.25$, 95% CI 0.04–0.43) and Qwen3-1.7B on Hinglish ($\rho = 0.24$, 0.05–0.43). A ρ of 0.25 corresponds to about 6% of rank variance.

H1 describes Gemma 4 (aligned, small gap); H2 describes Qwen3-1.7B (aligned, large gap) and, to a lesser degree, gpt-oss and Qwen3.8 on romanized Hindi; H3 hardly occurs because almost no question is poorly aligned; H4 receives weak support at best.

4.4 Multiple choice shrinks the gap

Relative to English, romanized Hindi costs gpt-oss 20% of its MedQA accuracy against 41% of its HealthBench score, and Qwen3.8 19% against 36%. Gemma 4 loses about the same on both,

Set	Model	English	Hindi	Roman Hindi	Hinglish
All (284)	Qwen3-1.7B	48.2	14.8	33.1	45.8
	Qwen3.8-27B	84.2	69.0	68.3	79.6
	gpt-oss-120b	90.1	78.9	71.8	85.6
	Gemma 4 31B	83.8	79.6	72.9	81.7
Oncology (26)	Qwen3-1.7B	65.4	23.1	38.5	50.0
	Qwen3.8-27B	92.3	73.1	76.9	84.6
	gpt-oss-120b	96.2	88.5	80.8	84.6
	Gemma 4 31B	88.5	80.8	80.8	80.8
General (93)	Qwen3-1.7B	45.2	16.1	24.7	45.2
	Qwen3.8-27B	75.3	62.4	64.5	75.3
	gpt-oss-120b	83.9	71.0	66.7	79.6
	Gemma 4 31B	79.6	78.5	67.7	80.6

Table 2: Romanized Hindi is the hardest MedQA form for the three larger models. Accuracy (%) on four-option questions; chance is 25%. "Oncology" is the 26 questions Gemini labels as mainly oncology; "General" is the stratified general sample minus oncology-labelled questions.

13% against 16%. Choosing an option needs only recognition; a HealthBench answer must also give the advice in Hindi, including when to see a doctor. A multiple-choice benchmark alone would therefore understate the Hindi penalty for gpt-oss and Qwen3.8.

Qwen3-1.7B's Devanagari score reflects instruction failure: in 129 of 284 questions it writes a Hindi explanation without choosing a letter until it hits the 512-token limit. I score these as wrong.

Sarvam-105B. Sarvam answered 258 MedQA questions in all four forms before running out of memory on the longest prompts. On those 258 its accuracy is 58.1% in English, 50.8% in Devanagari, 48.1% in romanized Hindi and 60.9% in Hinglish, against 83.7, 80.6, 74.4 and 82.2% for Gemma 4 and 90.3, 79.8, 73.3 and 86.4% for gpt-oss on the same questions. Its ordering of forms matches the other models. In 126 answers it reasons step by step and reaches the 512-token limit without naming a letter, which I score as wrong, so its accuracy is a lower bound. A letter parser that allows a newline after the leading letter recovers 13–27 of its answers per form and changes no other model by more than 0.8 points; the numbers above use it.

Retrieval at the peak layer is 97–100% for every model and form, and Spearman ρ between per-question alignment and correctness lies between 0.03 and 0.10, with every interval including zero.

On the 45 HealthBench oncology questions, Gemma 4 scores 0.54 in English, 0.47 in Devanagari and 0.41 in both romanized forms. The other models' oncology grades are pending.

5 Controls for translation and grading

A lower Hindi score can come from three sources: a translation that changes the question, a grader that marks Hindi text harder, or a model that writes worse advice in Hindi. Only the third is the effect of interest. Two controls separate them. Both are run; grading waits on the grader's API quota, and the scores will appear in the next draft.

Control A: Hindi question, English answer. Each of the 100 HealthBench general questions, in all four forms, receives one line asking for the answer in English, written in the question's own form ("कृपया उत्तर अंग्रेज़ी में दें।" in Devanagari, "Kripya jawab English mein dein." in romanized

Hindi, "Please answer in English." in Hinglish and English). The English condition measures the instruction's own effect. Qwen3-1.7B and Qwen3.8-27B answered with the settings of Section 3. An answer is compliant when more than 90% of its letters are Latin script; all 800 answers are compliant. One Qwen3-1.7B and 20 Qwen3.8-27B answers reach the 1,536-token limit. If Hindi-question scores recover to the English level, the model read the Hindi question and the loss lies in Hindi generation or in the grader; if they stay at the Hindi level, the loss arises before generation.

Control B: translate and regrade. Qwen3-32B, the model that wrote the question translations and not the grader, translated the 900 Hindi-form answers of Qwen3-1.7B, Qwen3.8-27B and gpt-oss-120b into English at temperature 0, with instructions to keep every fact, dose and warning. The translations are graded with the same grader and rubrics. Scores near the original Hindi scores indicate that the grader treats Hindi fairly; scores near English indicate a grader penalty on Hindi text. Two caveats: the translator turns Qwen3-1.7B's often ungrammatical Hindi into fluent English, which biases Control B upward for that model, and 92 of its 300 translations reach the output limit because the source answers repeat (against 0 for Qwen3.8 and 2 for gpt-oss). Spot checks of the translations find facts, lists and tables preserved. gpt-oss sometimes answers romanized-Hindi questions in Devanagari.

Not yet run. A bilingual clinician check of translation fidelity (30 questions $\times$ 3 forms) and of grader agreement (40 answers, Cohen's κ against Gemini) needs no compute and is planned.

6 Discussion

The models match each question to its English twin. The loss arises between a shared mid-layer representation and a good Hindi answer. Two explanations fit. Either the model cannot write good Hindi advice, or the meaning degrades in the late layers between the shared representation and the output. Control A (Section 5) separates them by asking in Hindi and requiring an English answer; patching the Hindi question's mid-layer state into the English forward pass [4] would localise the loss further. For deployment, the results argue for testing in romanized Hindi, the form most users type, and against choosing a model for Hindi on its English score.

7 Limitations

Translation. Qwen3-32B wrote every non-English question, and no native speaker has read them. Near-saturated retrieval suggests meaning survived, but register and medical terms may not.

Grader. One model grades all four forms. Any language bias in Gemini 3.8 Flash enters the gap directly. Control B (Section 5) tests this directly; a doctor's spot check of a sample is still needed.

Sarvam-105B. The Indian open model ran about seven times slower per token than gpt-oss in my setup (bf16 weights, a per-expert Python loop with a CPU sync per layer, four GPUs used in sequence) and was preempted twice. Both HealthBench runs failed before writing an answer and were dropped; its MedQA run answered 258 questions in all four forms.

Scale. With 100 questions the intervals on ρ are wide; detecting $\rho = 0.25$ reliably needs a few hundred.

Pending. HealthBench oncology grades for the two Qwen models and gpt-oss, whose answers are complete, and the scores for both controls. All wait on the grader's API quota.

Reproducibility

Code is in `https://github.com/inventcures/shape-of-hindi` under `src/health/` (`prepare.py`, `medqa.py`, `answer.py`, `grade.py`, `analyze.py`) with Modal entry points in `src/modal_app.py`. Control code is `src/health/controls.py` (described in `docs/health-controls.md`). The study cost about US$63 of GPU time and grading, of which about $7 was the two controls.

References

[1] tp53. Where Hindi meets English inside five open LLMs. inventcures.github.io, November 2026.

[2] C. Wendler, V. Veselovsky, G. Monea, R. West. Do Llamas work in English? On the latent language of multilingual transformers. ACL 2024. arXiv:2402.10588.

[3] J. Lindsey, E. Ameisen, et al. On the biology of a large language model. Transformer Circuits, 2025.

[4] C. Dumas et al. Separating tongue from thought: activation patching reveals language-agnostic concept representations in transformers. ACL 2025. arXiv:2411.08745.

[5] R. K. Arora et al. HealthBench: evaluating large language models towards improved human health. arXiv:2505.08775, 2025.

[6] D. Jin et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences 11(14), 2021. arXiv:2009.13081.