The shape of Hindi inside a language model

A synthesis of conversation state-space mapping, multilingual interpretability and language-dependent values, with a plan for Hindi and other Indian languages

tp53 `inventcures.github.io`

17 November 2026 Draft 0.2, with pilot results

Abstract

A recent field note by operator0x projected the residual-stream states of Qwen3-1.7B, read after the final RMSNorm, for about a thousand templated conversations into one fixed PCA frame. English conversations collapsed onto a plane, Chinese sat offset, and Python sat near English. The Hacker News discussion raised sound objections: the last layer mostly encodes the next token, the frame may be shaped by token embeddings, end states hide trajectories, and length can dominate such maps. This note synthesises that experiment, its critique, roughly forty papers on multilingual interpretability published between 2020 and 2026, and Anthropic's 2025 and 2026 studies of the values Claude expresses across languages, which found Claude leans furthest toward warmth in Hindi. I argue that the final-layer picture is close to guaranteed by tied embeddings, that the literature predicts a shared middle-layer space with language-specific ends, and that Hindi is an unusually clean laboratory for separating meaning from surface form. Hindi and Urdu are one spoken language in two scripts, Hindi and Marathi are two languages in one script, romanized Hindi and Hinglish are what people actually type, and the pronouns *aap*, *tum* and *tu* encode social relationship in the grammar. I give a pre-registered design with nine parallel conditions, four PCA frames, full-dimensional metrics and causal checks, plus a list of further experiments for Indian and other non-English languages. A first pilot, 20 machine-translated items in 8 versions through Qwen3-1.7B, supports the token-cost prediction, shows Hindi to English retrieval peaking at 90% at layer 16 once each language's mean is removed, and finds Marathi closer to Hindi in the first four layers and Urdu closer from layer 20 on. Romanized Hindi was not closer to English than Devanagari Hindi, against my prediction. The pilot is too small to settle any representational claim. Exercises with answers are included.

Contents

1 Why this question **2**

2 The original experiment and its critics **3**
2.1 Setup . . . 3
2.2 The critique . . . 3
2.3 Reading 99.3% and 17.6% together . . . 3

3 Why the last layer shows the next token **3**
3.1 Tied embeddings . . . 3
3.2 Massive activations and RMSNorm . . . 4

4 What the literature predicts **4**
4.1 Three phases across depth . . . 4
4.2 Language as an offset . . . 5
4.3 Evidence specific to Hindi and Indian languages . . . 5

4.4 Tokenization . . . 5
4.5 Conversations and personas . . . 5

5 Values across languages **6**
5.1 What Anthropic measured . . . 6
5.2 What other studies add . . . 6
5.3 What an open model can add . . . 6

6 Why Hindi is a good laboratory **6**

7 Proposed experiment **7**
7.1 Conditions . . . 7
7.2 Models . . . 7
7.3 Read-outs . . . 7
7.4 Frames . . . 7
7.5 Metrics . . . 7
7.6 Causal checks . . . 8
7.7 Controls . . . 8
7.8 Pre-registered predictions . . . 9
7.9 Values and register . . . 9
7.10 Cost . . . 9

8 Pilot results **9**
8.1 Token cost . . . 9
8.2 Retrieval across depth . . . 10
8.3 Script or language . . . 11
8.4 Confounds . . . 11

9 Further experiments **12**

10 Limitations **13**

11 Exercises **14**

1 Why this question

Suppose four people ask a chatbot for feedback on a business plan. One writes in romanized Hindi, *mera business plan dekh ke batao kaisa hai.* One writes the same words in Urdu script, میرا بزنس پلان دیکھ کے بتاؤ کیسا ہے. One writes in Marathi, a different language in the same Devanagari script as Hindi. One writes in English. A Hindi speaker hears the first two as the same sentence. Does the model?

Two pieces of work make this question timely. The operator0x field note gave a picture of how one small open model arranges conversations in different languages [1]. Anthropic's July 2026 study measured how the values Claude expresses shift with the language of the conversation, and found Hindi at the warm extreme of the top 20 languages [4]. One is geometry and the other behaviour. I think they are views of the same underlying question: when a model reads a sentence, how much of what it does depends on meaning, and how much on script, token count and social register?

2 The original experiment and its critics

2.1 Setup

The page itself refused automated fetches from my research environment, so the details below come from the Hacker News thread [2] and a briefing that quotes the article [3].

Model	Qwen3-1.7B: 28 layers, hidden size 2048, tied input and output embeddings
Data	900 templated question and answer pairs over 100 topics in four languages, plus 90 Python examples
Read-out	residual stream after the final RMSNorm
Projection	one global PCA frame; every conversation projected into it
Result	inside the 3D view English collapses onto a plane holding about 99.3% of its variation; Chinese is offset; code is near English
Stated caveat	the 3D view keeps 17.6% of total variance; 90% needs 684 directions

2.2 The critique

The 29 comments contain most of the design requirements for a follow-up. One commenter argued that the last-layer result says little more than that Chinese tokens are followed by Chinese tokens, and asked for middle layers. Another said including raw token embeddings in the PCA compresses the conversations onto a plane. A third asked about trajectories rather than end points. Several suggested UMAP or t-SNE. A builder of a similar map of Steam reviews reported that text length dominated the layout until it was removed, and others noted that length bias has been known in search since BM25, and that nuisance components can be regressed out before plotting, as single-cell biologists do with read depth. Finally, one commenter wished for a third script family, and a reply proposed Hindi and Urdu.

2.3 Reading 99.3% and 17.6% together

These numbers look contradictory. My reading is that 17.6% is the share of pooled variance across all 2048 dimensions kept by the three plotted axes, and 99.3% describes the shape of the English cloud inside that view. A small example shows they can coexist. Let the pooled covariance have eigenvalues 10, 6 and 2 on the three frame axes, and 0.05 on each of the other 2045 axes. Total variance is $18 + 2045 \times 0.05 \approx 120$, so the frame keeps $18/120 = 15\%$. If English varies only along the first two frame axes, it is perfectly planar in the view. Exercise 1 asks you to build your own version.

3 Why the last layer shows the next token

3.1 Tied embeddings

Let $E \in \mathbb{R}^{V \times d}$ be the embedding matrix, with row e_t for token t. Qwen3-1.7B uses the same matrix to read out logits:

$$z_t = e_t^\top h_L, \qquad p(t \mid \text{context}) = \frac{\exp z_t}{\sum_s \exp z_s}, \tag{1}$$

where h_L is the state after the final RMSNorm. Training makes h_L point toward the embedding of the likely next token. So h_L sits near the region of embedding space that contains the output language's tokens.

A two-token toy. Let English "the" be $e_1 = (1, 0)$ and a Chinese particle be $e_2 = (0, 1)$. If the model assigns probability 0.7 to the Chinese token, then $z_2 - z_1 = \ln(0.7/0.3) \approx 0.85$, so $h_{L,2} - h_{L,1} \approx 0.85$. Every Chinese-continuing state lies on the Chinese side of the line $h_2 = h_1$. A

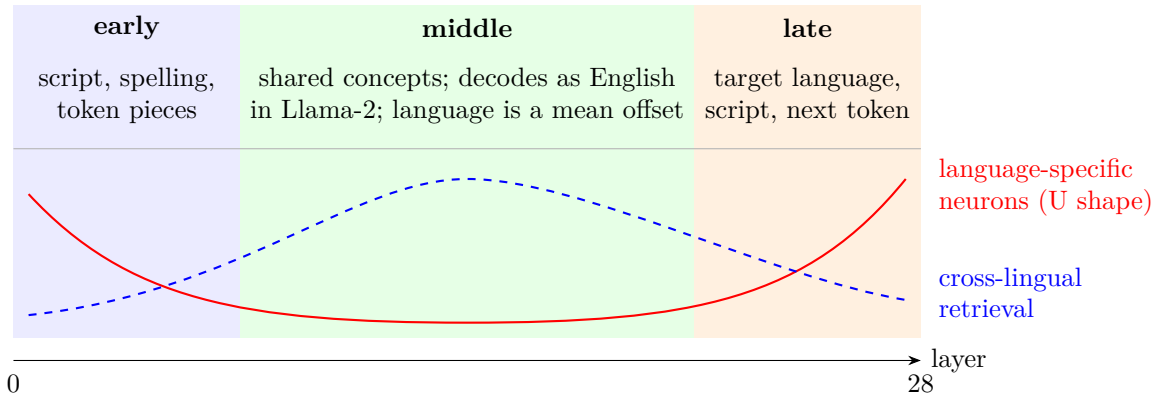


Figure 1: The picture the literature converges on, drawn schematically. Language-specific neurons pile up at both ends [10, 11, 12]; cross-lingual alignment peaks in the middle [13, 14, 15, 16]. The original field note read the network at the far right.

PCA of such states separates languages perfectly and tells us only that the model knows which language it is writing.

3.2 Massive activations and RMSNorm

Two further artifacts affect the final-norm reading. First, a few coordinates carry "massive activations" up to about 10^5 times typical size, mostly on the first token, acting as fixed biases [6]. In Qwen3-4B and 8B they appear abruptly at layer 7 [7]. Second, in Qwen3 the RMSNorm weights on those outlier dimensions are tiny, about 0.006 against 1, so the outliers inflate the root mean square and rescale every other coordinate [8]. Rogue dimensions of this kind can dominate cosine and Euclidean similarity without mattering for behaviour [9]. Any PCA must drop the first token and standardise dimensions first.

4 What the literature predicts

4.1 Three phases across depth

Evidence from logit lenses, neuron counts, sparse features and patching agrees on a picture with three parts (Figure 1).

Early layers are about the surface. Language-specific neurons form a U shape across depth; in Llama-2 70B about 7,000 sit at layer 2 against about 100 per middle layer [10]. Kojima et al. found the same with under 5% overlap between languages [11].

Middle layers are shared. Wendler et al. found Llama-2 decodes the right concept in English in its middle layers before switching to the target language near the end [13]. Dumas et al. patched middle-layer concepts across languages and found the output language is set around layers 12 to 16 of Llama-2 7B and the concept from about layer 16, with the result generalising to Qwen1.5 7B and others [14]. Wu et al. called this a semantic hub shaped by the dominant language: Llama-3 pivots on English, the Chinese-heavy Baichuan-2 on Chinese [15]. Anthropic found language-agnostic features in the middle of Claude 3.5 Haiku and language-specific ones at both ends, with the shared fraction growing with scale, most for pairs without a shared alphabet [16].

Late layers convert back into a language and script, which is where the original experiment looked.

4.2 Language as an offset

Chang, Tu and Bergen found that languages in XLM-R occupy similar affine subspaces once each is mean-centred, and that shifting by the difference of language means changes the output language [17]. Libovický et al. measured the consequence. In mBERT, centring each language raised parallel-sentence retrieval from 0.776 to 0.838 and lowered language identification from 0.960 to 0.853 [18]. Much of "being a language" is a constant vector.

The causal direction for language is the difference of means, not a principal component. A 2026 benchmark of steering in 32 languages including Hindi found difference-of-means directions succeed 84.5% of the time on average, against 15.1% for top PCA directions [19]. PCA finds variance, not use.

4.3 Evidence specific to Hindi and Indian languages

- Hindi has 781 language-specific neurons at the top-1% threshold in Llama-3.1-8B, second only to Tibetan among 21 languages [12].
- In Gemma-2 and Llama-2, the last six or seven layers pass through romanized subword pieces before Devanagari output, for Hindi and four other Indian languages but not Chinese. Patching gives near-identical concept curves for native and romanized input [20].
- Script is a late linear direction. One direction flips Hindi output between Devanagari and Latin while keeping meaning, and patching about five late heads switches script on every prompt [21].
- A competing result finds romanized and Devanagari Hindi activate almost disjoint units, Jaccard under 0.1 [22]. The two findings need reconciling, and a matched-content experiment can do it.
- Plain logit lens reads almost nothing for Indian languages in the middle layers of Sarvam-1, under 0.02 accuracy through layer 15, while a tuned lens is readable from layer 1 [23].
- In Serbian, the same language in Cyrillic and Latin script shares about 0.58 of active SAE features in Gemma-3 against 0.28 for random pairs, rising with scale [24]. I found no matched study of Hindi and Urdu.

4.4 Tokenization

The Qwen tokenizer charges Hindi heavily. On FLORES-200, Qwen3-8B, which shares the Qwen3 tokenizer, uses 4.75 tokens per Hindi word and 1.26 per English word; Gemma-3 uses about 1.39 and 1.24 [25]. The Qwen2.5 tokenizer, which Qwen3 inherits, has a Hindi to English parity of 4.43, rising to 7.02 for Telugu, and romanized Hindi costs 0.43 times native script [26]. A Hindi conversation is therefore about four times as long in tokens as its English twin, which is exactly the length confound the critics raised.

4.5 Conversations and personas

Anthropic's Assistant Axis work found the default assistant persona is a contrast direction in the middle layers of Qwen3-32B and other models, and that therapy-like and philosophical conversations drift away from it over 15 turns [27]. Persona vectors built from response-token averages predict and control traits [28]. Two 2026 papers give per-turn statistics that predict persistence and failure in Qwen3 models: subspace rotation angles [29] and trajectory curvature with covariance growth [30]. None of this has been tested in Hindi.

5 Values across languages

5.1 What Anthropic measured

"Values in the Wild" labelled the values Claude expressed in 308,210 subjective conversations from February 2025 and found 3,307 distinct values in five top-level groups [5]. Claude strongly supported the user's own values in 28.2% of conversations, reframed them in 6.6% and strongly resisted in 3.0%. The follow-up clustered these into 339 values, labelled 309,815 conversations from May 2026 across three models and 20 languages, dropped 18 values present in over 80% of conversations, and reduced the rest to four axes holding 15% of variance after controlling for task, topic and user values [4]:

deference vs caution warmth vs rigor depth vs brevity candor vs execution

Language moved Claude more than model choice did, mostly on warmth vs rigor and candor vs execution. Warmth was highest in Hindi and Arabic and rigor highest in English and Russian. Their illustration is two people asking for feedback on the same business plan in Hindi and in Russian, who may come away with different impressions of its quality.

5.2 What other studies add

Prompt language shifts value answers in open models about as much as explicit cultural framing [31]. Ethical reasoning in GPT-4 and Llama-2-70B was weakest in Hindi and Swahili and strongest in English and Russian [32]. Sycophancy rises in low-resource languages and tracks tokenizer fertility [33]. Most relevant here, a study of Hindi sycophancy found culturally adapted prompts explained about 14 points of the increase for Qwen2.5-7B and the language itself about 2 [34]. A warmer reply in Hindi may come from what people ask in Hindi rather than from Hindi.

5.3 What an open model can add

We cannot inspect Claude, but we can test whether an open model carries a similar lean in its states, and whether register cues like *aap* and *tu* share machinery with warmth. Section 7 gives the design.

6 Why Hindi is a good laboratory

Chinese and English differ in script, grammar, vocabulary, token cost and training share all at once. Hindi lets us vary one at a time.

Comparison	Held fixed	Varied
Hindi vs Urdu	spoken language, grammar, most words	script
Hindi vs Marathi	script	language
Devanagari vs romanized Hindi	language, words	script, token count
Hindi vs Hinglish	grammar	some content words
aap vs *tum* vs *tu*	everything but the pronoun	social relationship

The pronoun row is a social variable inside the grammar:

आप मेरी मदद करेंगे? तुम मेरी मदद करोगे? तू मेरी मदद करेगा?

All three ask "will you help me?". A Hindi speaker hears respect, familiarity and intimacy or rudeness.

7 Proposed experiment

The full specification is `docs/01-hindi-state-space-plan.md` in the repository. This section summarises it.

7.1 Conditions

Nine parallel arms, 100 topics by 6 templates, 600 items per arm, 5,400 conversations: `en`, `hi-deva`, `hi-latn`, `hien-latn`, `hien-deva`, `ur-arab`, `mr-deva`, `zh`, `code`. Arms 2 to 5 cross script with code-mixing; arms 2, 6 and 7 form a script against language triangle. Half the topics are universal and half Indian, so cultural content can be separated from language. One item:

English	How does a UPI payment actually move money from my account to a shopkeeper's?
Hindi	यूपीआई पेमेंट असल में मेरे खाते से दुकानदार के खाते तक पैसे कैसे पहुँचाता है?
Romanized	UPI payment asal mein mere khaate se dukaandaar ke khaate tak paise kaise pahunchaata hai?
Hinglish	UPI payment actually mere account se shopkeeper ke account tak paisa kaise transfer karta hai?
Urdu	یو پی آئی پیمنٹ اصل میں میرے کھاتے سے دکاندار کے کھاتے تک پیسے کیسے پہنچاتا ہے؟

Assistant answers are written once in English, translated, and teacher-forced, so content is identical across arms. Every result is repeated on natural parallel data from IN22-Conv and FLORES-200 to check that templating is not the cause.

7.2 Models

Qwen3-1.7B chat as primary; Qwen3-1.7B-Base, for which Qwen-Scope provides residual SAEs at all 28 layers [35]; Qwen3-0.6B, 4B and 8B for scale; Gemma-3-4B as a tokenizer control; Sarvam-1 as an Indian-trained control.

7.3 Read-outs

All 29 hidden states, at each turn's closing token, averaged over the assistant reply, and at the last user token. The first token and fixed template tokens are excluded.

7.4 Frames

F1, replication fitted on English, Chinese and code; Hindi projected in afterwards.

F2, balanced equal items per arm, standardised and mean-centred.

F3, content each arm centred on its own mean.

F4, difference PCA of paired differences, Hindi minus English.

7.5 Metrics

All in full dimension, with bootstrap intervals over topics.

Centred retrieval. For arms A and B with paired states a_i, b_i, $i = 1 \dots n$, let $\tilde{a}_i = a_i - \bar{a}$ and $\tilde{b}_i = b_i - \bar{b}$. Retrieval is

$$R(A, B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\Big[\, i = \arg\max_j \cos(\tilde{a}_i, \tilde{b}_j) \Big]. \tag{2}$$

With $n = 600$, chance is $1/600 \approx 0.17\%$. If raw retrieval is 40% and centred retrieval 85%, most of what kept the arms apart was one offset.

Linear CKA. For centred matrices $X, Y \in \mathbb{R}^{n \times d}$ of paired states,

$$\mathrm{CKA}(X, Y) = \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F}. \tag{3}$$

It is 1 when the two clouds have the same shape up to rotation and scale. It is sensitive to outliers [36], another reason to remove massive-activation dimensions first.

Participation ratio. With covariance eigenvalues λ_k, $\mathrm{PR} = (\sum_k \lambda_k)^2 / \sum_k \lambda_k^2$. Ten equal eigenvalues give 10; one large and nine tiny give close to 1.

Other metrics. Silhouette between arms, as used for refusal across languages [37]; planarity inside the 3D view; R^2 of log token count on the top 10 PCs; tuned-lens decoding per layer.

7.6 Causal checks

Add the Hindi minus English mean direction to English prompts at layers 10 to 20 and count Hindi replies. Patch middle-layer states between languages following Dumas et al. Rank Qwen-Scope features by a monolinguality score [38], ablate the top few, and check that only Hindi cross-entropy rises.

7.7 Controls

Every metric raw, with log length regressed out, and on a length-matched subset; with template tokens removed; with pairs shuffled for a chance floor; and on Gemma-3, where Hindi costs about as many tokens as English. UMAP only as an illustration with several seeds, because two-dimensional embeddings lose most nearest-neighbour structure [39].

7.8 Pre-registered predictions

	Prediction	Counts against
P1	At the final norm Hindi separates from English at least as clearly as Chinese	Hindi silhouette under half of Chinese
P2	Centred Hindi to English retrieval peaks between layers 10 and 20	peak at layer 24 or later
P3	Romanized Hindi and Hinglish sit closer to English late, and converge with Devanagari in the middle	romanized further from English at every layer
P4	Middle layers: Hindi closer to Urdu than to Marathi. Early layers: the reverse	Marathi closer at the middle-layer peak
P5	Hindi turns cost 3.5 to 5 times as many tokens as English	ratio under 2
P6	Without control, a top PC tracks log length with $R^2 > 0.5$	$R^2 < 0.2$
P7	A warmth direction built in English projects higher on Hindi replies, highest for *aap*	no difference, or English higher

7.9 Values and register

Build a warmth direction from 200 English feedback requests answered under warm and rigorous system prompts, as a difference of mean reply-token states at a middle layer. Check that it steers. Project unprompted replies in every arm onto it. Repeat with each Hindi request written with *aap*, *tum* and *tu*. Label replies with a judge model on a short rubric drawn from the Values in the Wild taxonomy and compare with the projections. Compare universal and Indian topics to separate language from content.

7.10 Cost

Qwen3-1.7B in bfloat16 needs about 3.4 GB; 5,400 forward passes of a few hundred tokens take under an hour on one consumer GPU. Turn-end states for all layers take about 1.3 GB in float16. The slow step is native-speaker review, about 20 hours per language for 600 items.

8 Pilot results

I ran a pilot on 17 November 2026 to test the pipeline end to end (Figure 2). It used 10 topics, five universal and five Indian, each with two templates, an explanation request and a request for feedback on a plan. That gives 20 items in each of 8 versions: English, Hindi in Devanagari, romanized Hindi, Hinglish in Roman and in Devanagari, Urdu, Marathi and Chinese, 160 conversations in all. Python was left out. Qwen3-1.7B ran in bfloat16 on one A10G GPU on Modal. Token counting took 8 seconds, extraction 55 seconds and analysis about 8 minutes. I read the residual stream at the end of each turn at all 28 layers, dropped the first token, and report the second turn. With 20 items, chance retrieval is $1/20 = 5\%$ and each result moves in steps of 5 points. Code, metrics and figures are in the repository under `results/` and `figures/pilot/`. All figures follow Saloni Dattani's guide to data visualisation.

8.1 Token cost

Figure 3 gives tokens per conversation relative to English. Devanagari Hindi cost 4.13 times as many tokens, Marathi 4.15, and Devanagari Hinglish 4.33. The same Hindi in Roman letters cost

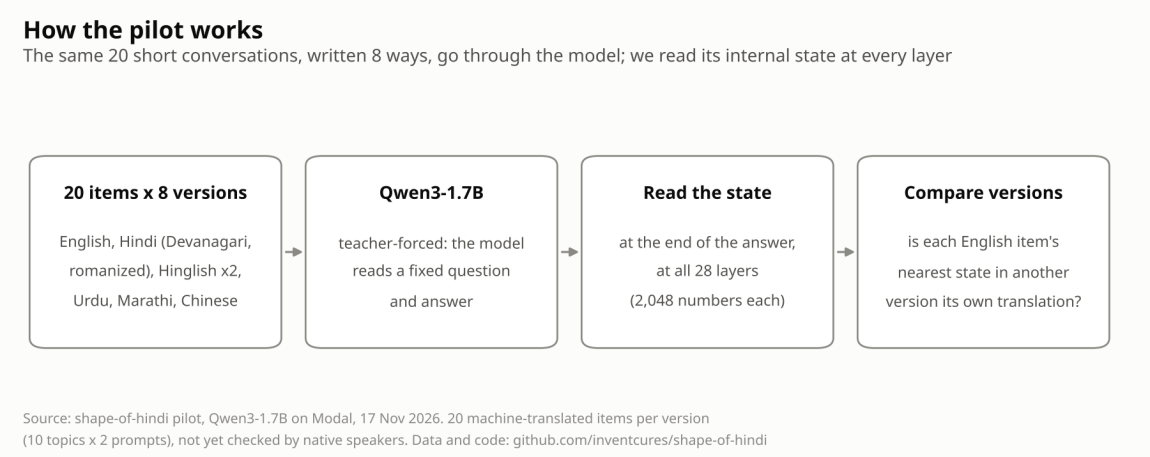


Figure 2: The pilot pipeline.

1.81 times and Roman Hinglish 1.38. Urdu cost 2.96 and Chinese 0.94. Per word, English took 1.15 tokens and Devanagari Hindi 4.63, close to the 4.75 reported for Qwen3-8B on FLORES [25]. P5 holds. The cost follows the script more than the language.

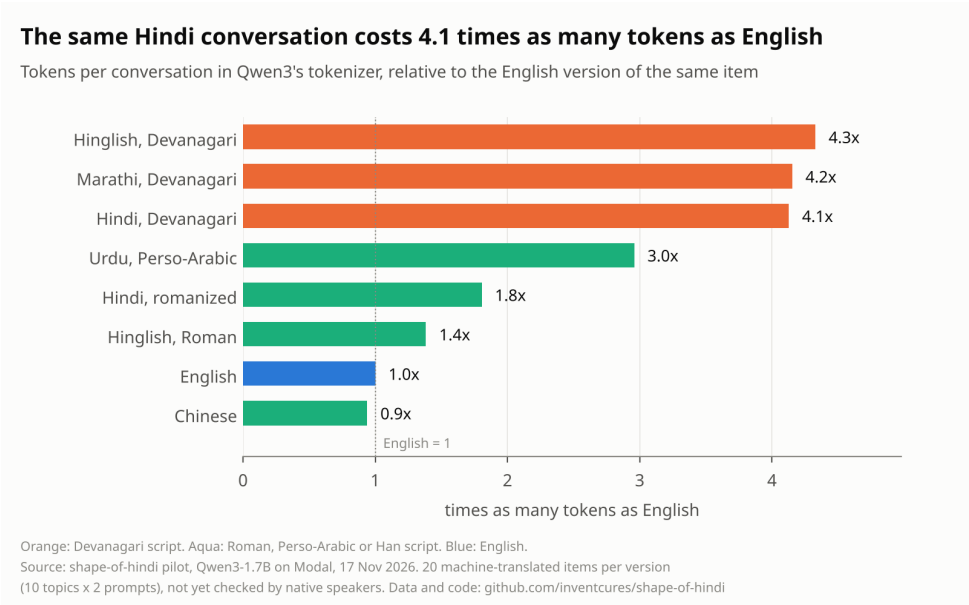


Figure 3: Tokens per conversation in Qwen3's tokenizer, relative to English.

8.2 Retrieval across depth

Figure 4 shows, for each English item, whether its nearest neighbour in another version is its own translation, by layer, before and after removing each version's mean. For Devanagari Hindi, raw retrieval reached 45% at layer 16 and fell to 10% at layer 28. Centred retrieval reached 90% at layer 16 and fell to 65% at layer 28. The peak falls in the predicted window, so P2 holds in this sample, and the gap between the two curves says most of the separation is a per-language offset, as the work reviewed in Section 4.2 predicts.

At layer 28, Hindi was further from English than Chinese was: raw retrieval 10% against 35%, centred 65% against 85%. This fits P1, though I measured it with retrieval, not the silhouette the table specifies. Chinese was at or near 100% centred retrieval through most of the middle,

which I suspect reflects close, near word-for-word translations. Roman Hinglish matched English at 80% or more at every layer, as expected from its shared English words.

P3 fails in its first half. Romanized Hindi had lower CKA with English than Devanagari Hindi at 23 of 28 layers, and at layer 28 matched English 50% of the time against 65% for Devanagari. Its second half holds: centred retrieval between Devanagari and romanized Hindi was 100% at layer 8, 95% at layer 16 and 75% at layer 28.

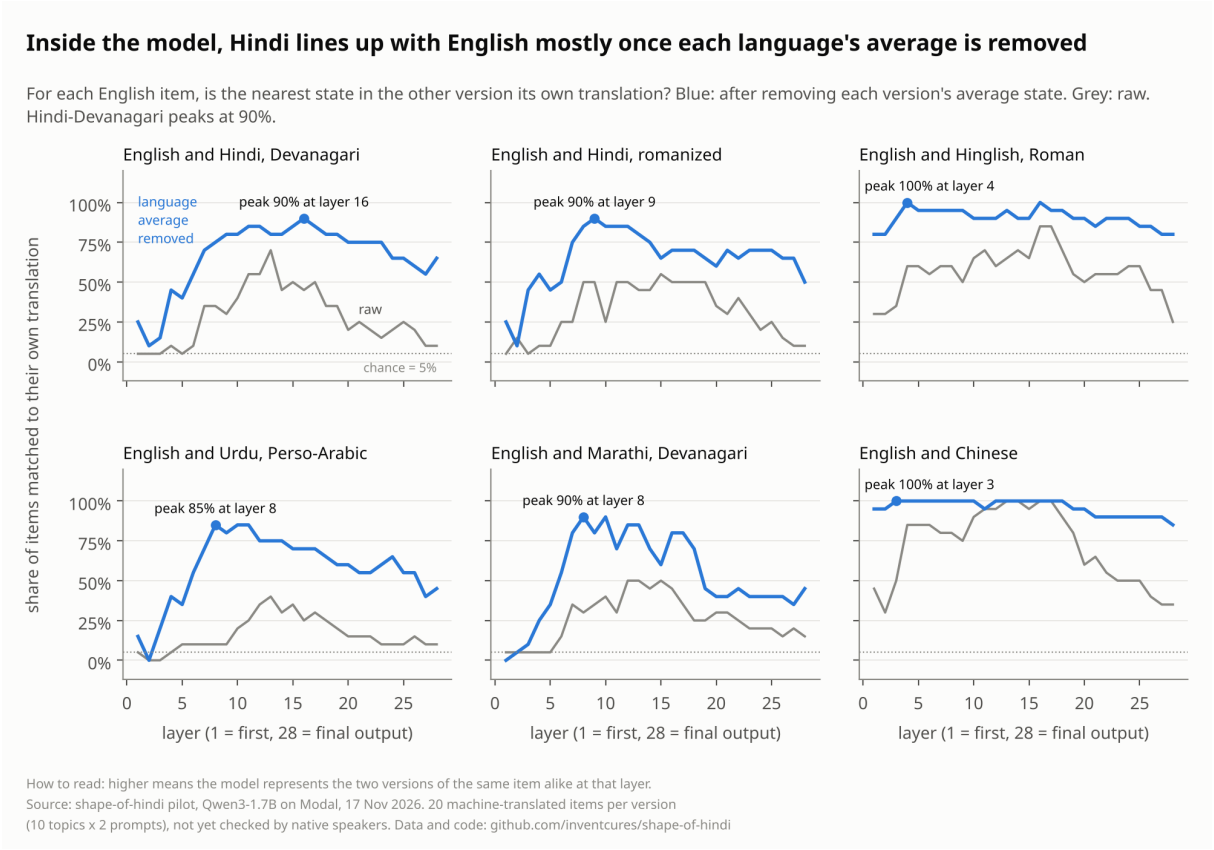


Figure 4: Retrieval of each English item's own translation, by layer, raw and after removing each version's mean.

8.3 Script or language

Figure 5 compares linear CKA between Hindi and Urdu, which share a language but not a script, with CKA between Hindi and Marathi, which share a script but not a language. Marathi was closer at layers 1 to 4, Urdu at layers 5 to 10, the two were within 0.015 of each other from layer 11 to 19, and Urdu was closer from layer 20 to 28, ending at 0.86 against 0.80. Urdu was closer at 18 of 28 layers. Centred retrieval at layer 28 was 95% for Urdu and 80% for Marathi, a difference of three items out of 20. The early half of P4 holds. The middle half is a tie, not a win, so P4 is unresolved, and the clearest separation is late, which the prediction did not anticipate.

8.4 Confounds

The top 10 principal components predicted log token count with R^2 between 0.94 and 0.98 at every layer (Figure 6, left), far above the 0.5 threshold in P6. But a logistic probe identified the version from the state with 5-fold accuracy between 0.93 and 0.98, and in this pilot the Devanagari versions are all about four times as long as the others, so length and language are nearly collinear. P6 cannot be tested here; the length-matched subset and Gemma-3 are needed.

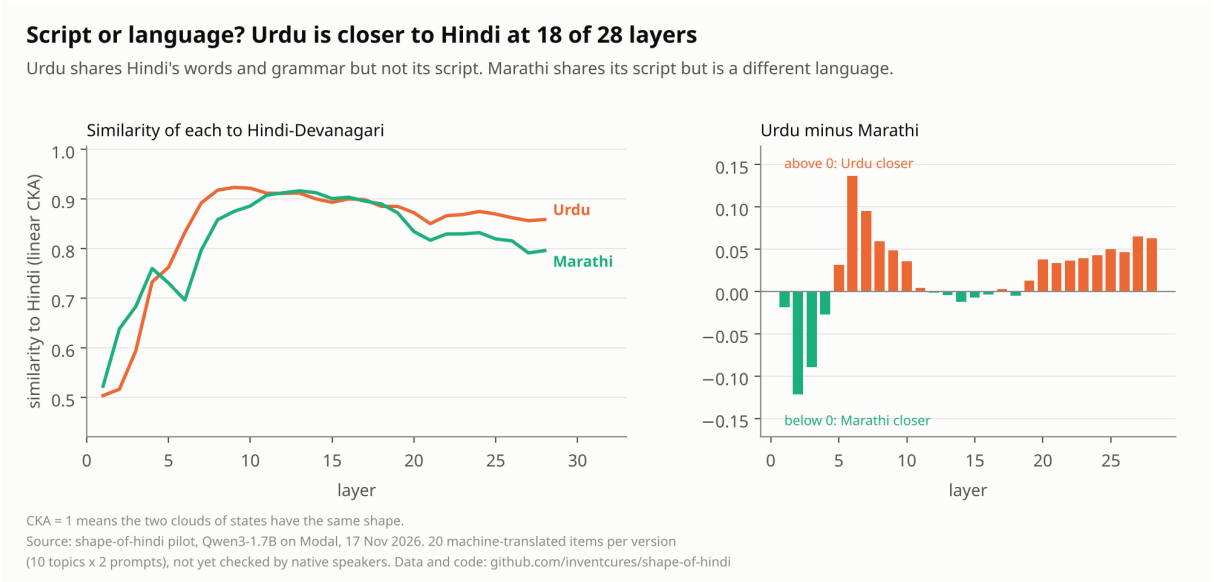


Figure 5: Similarity of Urdu and of Marathi to Hindi by layer, and their difference.

On the first token of the English conversations, the largest absolute value was 1,244 times that of a typical token at layer 3, falling to about 9 times by layer 27, and below 1 after the final norm (Figure 6, right). This matches the massive-activation literature [6, 7] and justifies dropping the token.

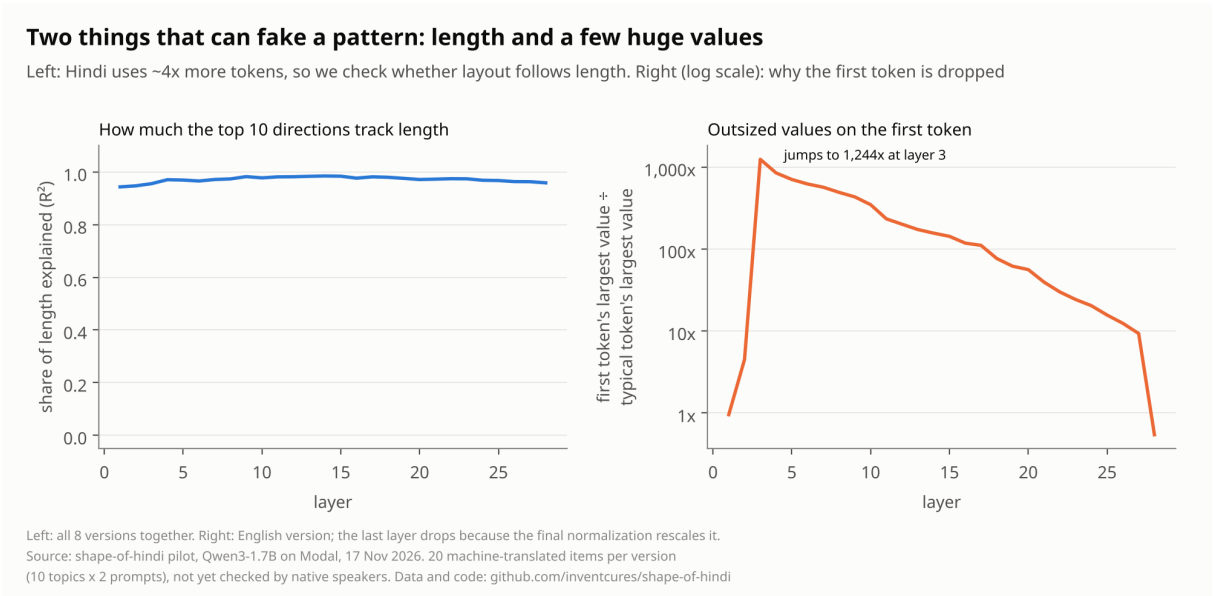


Figure 6: Length explained by the top 10 components, and the first token's outsized values.

Table 1 sets the pilot against the predictions.

9 Further experiments

The repository file `docs/02-further-experiments.md` gives designs, examples and costs. In brief:

1. **Hindi and Urdu** as a two by two of script and vocabulary, following the Serbian study

Table 1: Pre-registered predictions against the pilot.

	Pilot outcome
P1	Consistent: Hindi more separated from English than Chinese at layer 28
P2	Consistent: centred peak 90% at layer 16, 65% at layer 28
P3	Mixed: romanized Hindi not closer to English late; it does converge with Devanagari in the middle
P4	Mixed: Marathi closer early as predicted; a tie in the middle; Urdu closer late
P5	Holds: Devanagari Hindi 4.1 times English
P6	Not testable: length and version are collinear in the pilot
P7	Not run

[24].

2. **Same script, different language**: Hindi, Marathi, Nepali, Bhojpuri, Maithili and Sanskrit in Devanagari.
3. **The family tree**: build a tree from internal distances over 22 languages in IN22 and test whether it splits Indo-Aryan from Dravidian or clusters by token cost.
4. **Hinglish switch points**: track the state token by token through sentences like *kal mera exam hai aur maine abhi tak syllabus complete nahi kiya*.
5. **Honorifics**: *aap*, *tum* and *tu* as directions, compared with warmth and deference, and ported to Korean, Japanese and Tamil.
6. **User gender in verbs**: मैं गया and मैं गई both mean "I went". Does the model remember the speaker's gender and change its advice?
7. **Safety**: the English refusal direction transfers across 14 languages, none Indian [37]. GPT-4 attack success was 6.5% in Hindi against 0.96% in English [40], and Llama-3-8B rose from 0.06 in English to 0.63 in misspelled Hinglish [41].
8. **Culture versus language**: Diwali and Thanksgiving, each in English and Hindi.
9. **Token cost**: Devanagari against romanized Hindi within Qwen, and Qwen against Gemma-3 and Sarvam-1.
10. **Latent language**: which language the tuned lens decodes in Qwen's middle layers for Hindi input: English, Chinese, or romanized Hindi.
11. **Values profiles** for six Indian languages in an open model, following Anthropic's method.
12. **Indian number grouping**: lakh and crore against million.
13. **Other languages**: Serbian, Punjabi, Chinese simplified and traditional, and Malay as script pairs; Arabic and Tamil diglossia; Korean and Javanese speech levels; low-resource languages as a stress test; persona drift by language; the language of the reply under mixed-language prompts.

10 Limitations

I could not load the original field note, so its details rest on secondary sources. Many cited 2026 results are recent preprints. The pilot in Section 8 used 20 items per version that I wrote with a model's help and no native speaker has checked, read at the end of the second turn of one model, and its analysis does not standardise dimensions before PCA as the plan specifies. Teacher-forced answers measure how the model represents a fixed text, not what it would say. A

judge model labelling warmth may bring its own language biases, the same concern Anthropic raise about Claude labelling Claude [5].

11 Exercises

1. Build a toy covariance where a 3D frame keeps under 20% of pooled variance and one group of points is perfectly planar inside it.
2. In the two-token model of Section 3.1, the model is 90% sure the next token is Chinese. What is $h_{L,2} - h_{L,1}$? Why does this make the final norm a poor place to compare languages?
3. Raw Hindi to English retrieval is 12%; centred, it is 71%. What did centring remove?
4. A 50-word English reply has about how many Qwen3 tokens, and its Hindi translation about how many? Why does that matter if you average states over the reply?
5. Which comparison isolates script: Hindi vs Marathi, Hindi vs Urdu, or Hindi vs Hinglish? Which isolates language?
6. Compute the participation ratio for eigenvalues $(4, 4, 1, 1)$.
7. The warmth projection is higher for Hindi replies than English ones, but only on Indian-topic items. What does that suggest about where the lean comes from, and which experiment would test it?
8. In the pilot, centred retrieval at layer 28 was 95% for Hindi and Urdu and 80% for Hindi and Marathi, out of 20 items each. How many items is the difference? Is it evidence that the model groups Hindi with Urdu?

Answers

1. Eigenvalues 10, 6 and 2 on the frame axes and 0.05 on each of 2045 others give total about 120 and a frame share of 15%. A group varying only along frame axes 1 and 2 is planar in the view.
2. $\ln(0.9/0.1) = \ln 9 \approx 2.20$. The final state is determined largely by the next-token distribution, so it separates languages by output vocabulary whatever the model's internal representation of meaning.
3. A constant per-arm offset, the language mean, shared by every item.
4. About $50 \times 1.26 \approx 63$ English tokens. A Hindi translation of similar word count is about $50 \times 4.75 \approx 240$ tokens, nearly four times as many. A mean over many more, smaller pieces mixes in more sub-word and positional structure, and position has its own axes in these models [17], so length differences can masquerade as language differences.
5. Hindi vs Urdu isolates script, largely. Hindi vs Marathi isolates language with script fixed. Hindi vs Hinglish varies vocabulary within one grammar.
6. $(4 + 4 + 1 + 1)^2/(16 + 16 + 1 + 1) = 100/34 \approx 2.94$.
7. The lean comes from content and context rather than from the language itself, in line with the Hindi sycophancy result [34]. Test it by patching only the Hindi language direction into a rigorous English reply on a universal topic and checking whether warmth rises.
8. 19 against 16 items, a difference of three. The two scores are on the same 20 English items, so the right test is paired, such as McNemar's test on the items where the two disagree. With at most a few discordant items it cannot reach significance: if Urdu wins on all three of three discordant items, the exact two-sided p is $2 \times 0.5^3 = 0.25$. It is a direction worth checking at 600 items, not evidence.

References

[1] operator0x. What thousands of conversations look like inside a language model. Field note, November 2026. https://operator0x.com/field-notes/qwen-conversation-state-space/

[2] Hacker News discussion, item 50481321, November 2026. https://news.ycombinator.com/item?id=50481321

[3] The Context, Hacker News front-page briefing, 15 November 2026. https://thecontext.dev/en/briefing/

[4] M. Kearney, M. Zhang, et al. Claude's values across models and languages. Anthropic, 13 July 2026. https://anthropic.com/research/claude-values-models-languages

[5] S. Huang et al. Values in the wild: discovering and analyzing values in real-world language model interactions. Anthropic, 2025. arXiv:2504.15236.

[6] M. Sun, X. Chen, J. Z. Kolter, Z. Liu. Massive activations in large language models. arXiv:2402.17762, 2024.

[7] Shi et al. A single layer to explain them all. ICML 2026. arXiv:2605.08504.

[8] Qiu et al. A unified view of attention and residual sinks. arXiv:2601.22966, 2026.

[9] W. Timkey, M. van Schijndel. All bark and no bite: rogue dimensions in transformer language models. EMNLP 2021. arXiv:2109.04404.

[10] T. Tang et al. Language-specific neurons: the key to multilingual capabilities in large language models. ACL 2024. arXiv:2402.16438.

[11] T. Kojima et al. On the multilingual ability of decoder-based pre-trained language models: finding and controlling language-specific neurons. NAACL 2024. arXiv:2404.02431.

[12] D. Gurgurov et al. Language arithmetics: towards systematic language neuron identification and manipulation. IJCNLP-AACL 2025. arXiv:2507.22608.

[13] 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.

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

[15] Z. Wu et al. The semantic hub hypothesis. ICLR 2025. arXiv:2411.04986.

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

[17] T. Chang, Z. Tu, B. Bergen. The geometry of multilingual language model representations. EMNLP 2022. arXiv:2205.10964.

[18] J. Libovický, R. Rosa, A. Fraser. On the language neutrality of pre-trained multilingual representations. Findings of EMNLP 2020. arXiv:2004.05160.

[19] D. Gurgurov et al. CLaS-Bench: a cross-lingual alignment and steering benchmark. arXiv:2601.08331, 2026.

[20] A. Saji et al. RomanLens: the role of latent romanization in multilinguality in LLMs. arXiv:2502.07424, 2025.

[21] D. Gurgurov, A. Saji, et al. The Latin substrate: how language models represent and mediate script choice. arXiv:2605.31363, 2026.

[22] Verma et al. Multilingual language models encode script over linguistic structure. ACL 2026.

[23] Panchal et al. Indic-TunedLens: interpreting multilingual models in Indian languages. VarDial 2026. arXiv:2602.15038.

[24] Karne. One language, two scripts. arXiv:2603.08869, 2026.

[25] Shravan. BrahmicTokenizer-131K. arXiv:2605.29379, 2026.

[26] Johny, Ghosh, Sam. The Indic token tax in 2026. arXiv:2611.13058.

[27] C. Lu, J. Gallagher, J. Michala, K. Fish, J. Lindsey. The assistant axis: situating and stabilizing the default persona of language models. arXiv:2601.10387, 2026.

[28] R. Chen, A. Arditi, H. Sleight, O. Evans, J. Lindsey. Persona vectors: monitoring and controlling character traits in language models. arXiv:2507.21509, 2025.

[29] Simhi et al. Old habits die hard: how conversational history geometrically traps LLMs. ICML 2026. arXiv:2603.03308.

[30] Liang et al. Geometry of divergence: tracking hidden-state trajectories for adaptive multi-turn reasoning. arXiv:2608.30650, 2026.

[31] Bulté, Rigouts Terryn. LLMs and cultural values: the impact of prompt language and explicit cultural framing. arXiv:2511.03980, 2025.

[32] U. Agarwal et al. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. LREC-COLING 2024. arXiv:2404.18460.

[33] Shah et al. Multilingual sycophancy. arXiv:2606.08451, 2026.

[34] Sattigeri. Extending Beacon to Hindi. arXiv:2602.00046, 2026.

[35] Qwen Team. Qwen-Scope: turning sparse features into development tools for LLMs. arXiv:2605.11887, 2026.

[36] M. Davari et al. Reliability of CKA as a similarity measure in deep learning. ICLR 2023. arXiv:2210.16156.

[37] X. Wang et al. Refusal direction is universal across safety-aligned languages. NeurIPS 2025. arXiv:2505.17306.

[38] B. Deng et al. Unveiling language-specific features in large language models via sparse autoencoders. ACL 2025. arXiv:2505.05111.

[39] T. Chari, L. Pachter. The specious art of single-cell genomics. PLoS Computational Biology, 2023.

[40] Z.-X. Yong, C. Menghini, S. Bach. Low-resource languages jailbreak GPT-4. arXiv:2310.02446, 2023.

[41] Aswal, Jaiswal. Haet Bhasha aur Diskrimineshun: phonetic perturbations in code-mixed Hinglish. arXiv:2505.14226, 2025.