Cultural Awareness is Represented but Not Decoded: Tracing Mythological Knowledge across 18 Open-Source LLMs

Anonymous ACL submission

Abstract

Open-source LLMs reliably name *Zeus*, *Jupiter*, and *Thor*, but recover their counterparts in less-represented traditions like Finnish, Slavic, Egyptian, or Chinese mythology far less consistently. We ask where inside the model this cultural default is produced. On a parallel cross-cultural substrate of Thompson-motif entities, we instrument 18 open-source LLMs from 8 architecture families with linear probing, logit lens, activation patching, and output extraction. The residual stream cleanly distinguishes cultures, well above a name-string baseline, yet the decoder collapses culturally-specific tokens onto dominant-tradition ones. *The failure is at readout, not at representation.* Asking the same question in the target culture's native language versus English produces failures that cluster within language but decouple across language: the decoder is gated on prompt language. We release a per-entity (probe, output) decomposition framework, a citation-anchored cross-cultural ground truth, a within- versus cross-mode correlation test for language-conditioned readout, and per-entity predictions for all 18 models.[1]

1 Introduction

Large language models are deployed across the world, yet their pretraining corpora are dominated by English-language web data and a narrow band of culturally hegemonic sources. The resulting cultural skew shows up in tasks as different as moral reasoning, food and religious practice, named-entity recognition, and even temperature-scale defaults, with LLMs defaulting to the Anglo-American or broader Western frame across all of them (Naous et al., 2024; Palta and Rudinger, 2023; Atari et al., 2023; Adilazuarda et al., 2024). Folk-narrative content sits squarely inside that bias surface, but it has a property the surveyed tasks lack: a long tradition of parallel structural cataloguing across cultures (the Aarne-Thompson-Uther type index and the Thompson Motif-Index of Folk-Literature (Thompson, 1955)) lets one ask the same question of the same model across many traditions with an unambiguous gold answer. We use this property to push the cultural-bias diagnosis from *behavioral* ("what does the model say?") to *mechanistic* ("where inside the model is the bias produced?"), and ask what mitigations it suggests.

When asked to name the supreme sky god, open-source LLMs reliably produce *Zeus* for "Greek", *Jupiter* for "Roman", and *Thor* for "Norse",[2] but for less-represented traditions (Finnish, Ukrainian, Mesopotamian, etc.) the canonical filler is recovered far less consistently, a familiar behavioral diagnosis (Naous et al., 2024; Palta and Rudinger, 2023; Ramezani and Xu, 2023); wrong answers tend to be refusals, hedges, or a culturally plausible but incorrect entity from within the target tradition. The follow-up is mechanistic: *is the model unaware that the entities are culturally distinct, or is it aware but unable to retrieve the right name?* The two answers split cultural bias into two regimes that respond to opposite interventions, **representational flattening** (the residual stream does not separate cultures, so the fix needs new information) and **decoding flattening** (the residual separates cultures but the readout collapses culturally-specific tokens onto dominant-tradition ones). Disentangling these regimes under scaling and prompt-language conditioning motivates our five research questions:

- **RQ1** (representational preservation): *To what extent* does the residual stream encode culture above a name-surface baseline when the output is wrong?
- **RQ2** (layer-wise emergence): *At what depth* does the gold-token continuation enter the top-

[1] Code, data, and per-entity predictions for every model will be released at an anonymized URL upon acceptance.

[2] The chief sky god of Norse tradition is Odin; the Thompson-index canonical filler is Thor, the storm-sky figure.

k, relative to the probe peak?

- **RQ3** (causal localization): *Where* along the network does swapping residual streams between cross-cultural prompts begin to flip the model's continuation?
- **RQ4** (scaling): *How* does within-family scaling reshape the gap between probe and output accuracy: does it close, shrink, or persist?
- **RQ5** (cross-lingual querying): *How* do paraphrase failures decompose under prompt-language conditioning, and *how much* cell recovery does a bilingual ensemble buy over either single mode?

To answer these we instrument 18 open-source LLMs spanning 1.2B–34B across eight architecture families (Llama 3.x (Grattafiori et al., 2024), Gemma 4 (Team et al., 2025), Phi 4 (Abdin et al., 2024), Qwen 1.x (Qwen et al., 2025), Qwen 3.6 (Yang et al., 2025), Yi 1.5 (AI et al., 2025), OLMo 3.1 (OLMo et al., 2025), Tiny Aya (Aryabumi et al., 2024a,b)) with four mechanistic measurements:

- **E1** layer-wise linear probing of the residual stream for the 10-way culture label;
- **E2** logit lens tracking the depth at which the gold entity's first sub-token enters the top-k continuations;
- **E3** cross-cultural activation patching: swap source-culture residuals into the target-culture prompt and measure the layer-wise causal effect on the output;
- **E4** (output extraction) greedy chat-template generation in two prompt modes (EN: English query, NL: target-language query).

Each entity's E1 outcome (does the probe read the culture?) and EN outcome (does the model emit the right name?) form a 2×2 decomposition (§3.3) with four cells: **Preserved** (recognized and emitted), **DecodingSuppressed** (recognized but emits a wrong name), **SurfaceLuck** (right name without internal disambiguation), and **RepresentationallyFlat** (neither). The substrate is a parallel cross-cultural entity set of 27 Thompson-index motifs evaluated across 10 cultures.

Our main empirical finding is that DecodingSuppressed (probe right, output wrong) is the largest cell in every one of the 18 **models;** the asymmetry is at the decoder, not the encoder. Within-family scaling shrinks the probe−output gap but does not close it in any family (§4.4). The unembedding is also language-conditioned: cross-language paraphrases decouple while within-language ones correlate ($\overline{r}_{\text{within}} = 0.57$ vs $\overline{r}_{\text{cross}} = 0.29$), and a bilingual ensemble lifts cell recovery at zero training cost (§4.5).

2 Related Work

A growing behavioral literature documents that LLMs default to Western and Anglo-American cultural reference frames (Naous et al., 2024; Palta and Rudinger, 2023; Ramezani and Xu, 2023; Atari et al., 2023), with recent surveys consolidating the programme (Hershcovich et al., 2022; Adilazuarda et al., 2024); cross-language and cross-script asymmetries underpinning these gaps are documented at the data and tokenization levels (Kreutzer et al., 2022; Chelombitko et al., 2026), and the disconnect between formal "low-resource" coverage and actual cultural representation is articulated by Hämäläinen (2021). What this strand cannot tell us is *where* inside the model the failure happens.

The natural instruments are the linear probe (Alain and Bengio, 2018; Hewitt and Liang, 2019; Belinkov, 2022), which since LAMA (Petroni et al., 2019) has become the standard tool for asking what knowledge is linearly readable from a residual stream and has been applied to factual, geographic, and temporal information (Gurnee and Tegmark, 2024; Marks and Tegmark, 2024); the logit lens (Belrose et al., 2025), used in multilingual form to argue that LLMs reason in a dominant language and translate at the last layer (Wendler et al., 2024); and activation patching (Vig et al., 2020; Meng et al., 2023; Geiger et al., 2024; Wang et al., 2022), the central causal instrument of the mechanistic-interpretability programme (Olah et al., 2020; Elhage et al., 2021). For these instruments to mean anything cross-culturally we need a substrate parallel by construction, which we draw from Thompson's Motif-Index of Folk-Literature (Thompson, 1955), previously applied in computational folkloristics (Karsdorp and van den Bosch, 2013), in the spirit of treating LLM evaluation as a humanities-informed enterprise (Hämäläinen et al., 2024). Our contribution is to localize the behavioral gap mechanistically: where the probe sees a strong signal but the lens and the output do not, the failure is the unembedding's, not the residual stream's.

3 Method

3.1 Parallel cultural entity set

Our base unit is the (motif, culture) pair. A motif is a Thompson-index entry with a specifiable role, e.g. A1141.2 "the supreme sky god" or A220 "the sun god". We instantiate the role of 27 such motifs across ten cultures (**Greek**, **Roman**, **Norse**, **Finnish**, **Ukrainian**, **Indian**, **Egyptian**, **Chinese**, **Japanese**, and **Mesopotamian**) for a total of 270 entities. The 10 cultures span six language families (Indo-European, Uralic, Afro-Asiatic, Sino-Tibetan, Japonic, and Akkadian-Sumerian), so that within-family scaling, cross-lingual querying, and output decoding are not evaluated on a corpus dominated by Greco-Roman default-fitting.

For each (motif, culture) we record the canonical proper name in romanized English, the native-script form where applicable, the standardized motif description, and a per-row `source` field pointing to either a primary text (Eddas, Kalevala, Rig Veda, Pyramid Texts, Kojiki, Enuma Elish, the *Primary Chronicle*, etc.) or a modern academic monograph.[3] Four cells are structural absences (Japanese flood A1010, Finnish flood A1010, Mesopotamian theft-of-fire A1415, Egyptian forest-spirit F460), excluded from output-accuracy scoring; a further handful use a contested filler (e.g. Khors vs Dažbog as sun-disk vs sun-deity in Slavic, Mielikki vs Rauni for Finnish T110), where we follow modern scholarly consensus (Łuczyński, 2020; Gieysztor, 2006). Per-tradition source repertoire and contested/absence rows are documented in Appendix A; per-row citation anchors in Appendix B.

For each entity we build two prompts:

- **Contextual** (used for E1 hidden-state extraction): *"${name} embodies the role of ${description}."* The entity comes first, the culture word is never present, and the entity-token activations cannot trivially copy a culture token via causal attention. The probe must derive culture from the name and role alone.
- **Chat** (used for EN generation): *"In ${culture} mythology, what is the name of the ${description}? Reply with only the proper name, no explanation, no extra words."* Wrapped in the model's chat template where applicable.

[3] Native-script forms are resolved via Wikidata. The complete dataset, including all native-script names and the NL prompt templates, will be released as a Hugging Face dataset card alongside the camera-ready version; see Appendix A for the release schema and Appendix B for the canonical name and citation anchor of every entity.

3.2 Models

Table 1 lists the 18 LLMs probed. Selection criteria: (i) open-source; (ii) decoder-only causal LM; (iii) at most 35B parameters at fp16 (single 4× A100 40 GB tensor-parallel slot). All experiments run at fp16 with no further quantization to avoid confounding cultural recall with quantization-induced degradation, which we treat as a separate study.

Family	Model	Params (B)	Layers
Llama 3.x	Llama-3.2-1B-Instruct	1.2	16
Llama 3.x	Llama-3.2-3B-Instruct	3.2	28
Llama 3.x	Llama-3.1-8B-Instruct	8.0	32
Gemma 4	Gemma-4-E2B-it	2.0	35
Gemma 4	Gemma-4-E4B-it	4.0	42
Gemma 4	Gemma-4-31B-it	31.0	60
Gemma 4 MoE	Gemma-4-26B-A4B-it	4.0^a	30
Phi 4	Phi-4-mini-instruct	3.8	32
Phi 4	Phi-4	14.0	40
Qwen 1.x	Qwen2.5-1.5B-Instruct	1.5	28
Qwen 1.x	Qwen2.5-7B-Instruct	7.0	28
Qwen 3.6	Qwen3.6-27B	27.0	64
Qwen 3.6 MoE	Qwen3.6-35B-A3B	3.0^a	40
Yi 1.5	Yi-1.5-6B-Chat	6.0	32
Yi 1.5	Yi-1.5-9B-Chat	9.0	48
Yi 1.5	Yi-1.5-34B-Chat	34.0	60
OLMo 3.1	Olmo-3.1-32B-Instruct	32.0	64
Aya	tiny-aya-global	3.0	36

Table 1: 18 open-source LLMs in the sweep. [a] Active params per token for MoE (total 26B / 35B); layers exclude embedding.

3.3 Mechanistic instruments

We apply four complementary measurements to every (entity, culture) pair in every model.

Linear probing (E1). For each model and each layer we average-pool residual-stream activations over the entity span of the contextual prompt and fit a 5-fold-stratified ridge classifier ($\alpha = 1.0$) to predict the 10-way culture label (Alain and Bengio, 2018; Hewitt and Liang, 2019). We report layer-wise accuracy, the layer at which accuracy peaks, and a label-shuffled chance ceiling.

Logit lens (E2). Following the standard logit-lens recipe (Belrose et al., 2025), we apply the model's final norm and unembedding to every intermediate hidden state and ask at which depth the gold entity's first sub-token enters the top-k continuations. We report top-k rather than top-1 because gold-token boundaries are sensitive to

leading-space tokenization and to morphological variants (*Mokoš*/*Mokosha* differ by one edit), and our depth-of-readout claim is robust to whether one accepts the canonical or alternative filler in the contested rows of Appendix B. We use the raw lens for cross-model comparability; the tuned variant of Belrose et al. (2025) is discussed in the Limitations section.

Activation patching (E3). For each motif we form pairs of cross-cultural prompts (Greek → Finnish, etc.); at each layer ℓ we replace the residual stream of the source prompt with that of the target prompt and measure the change in the gold-target log-probability (Vig et al., 2020; Meng et al., 2023; Geiger et al., 2024; Wang et al., 2022).

Output extraction (E4). We greedily decode up to 256 tokens of the chat-template form. Reasoning-style chat templates that emit a `<think>...</think>` trace by default (Qwen 3.6 family) are run with the think trace disabled and any leftover block stripped before scoring. The cleaned generation is scored by exact match, substring match, and length-normalized Levenshtein similarity (threshold 0.8); a positive on any of the three counts as *correct*. The instrument is applied to two prompt modes per cell: EN (English query, 5 paraphrases) and NL (target-culture native language, 5 paraphrases); see §4.5 and Appendix I. Combining per-cell E1 and E4 outcomes gives the 2×2 decomposition introduced in §1; per-model cell shares are in Appendix C.

4 Results

4.1 RQ1: Decoding flattening is universal

Across all 18 LLMs in the sweep, the cultural identity of a mythological entity is recoverable from the residual stream at a rate the model never reproduces in its own generation (Table 2). Every probe clears the 0.60 char-n-gram surface baseline (§D) by $+0.02$ to $+0.29$, peaking at 0.61–0.88; the same models, asked to emit the name, land at 0.09–0.43. **The smallest gap in the sweep is** 0.26 (Gemma-4-31B), the largest is 0.70 (Llama-3.2-1B), and the gap shrinks but never closes within any family. The geometric form of this readout is the same across the sweep: at each model's probe-peak layer, the residual stream separates cultures into visibly distinct clusters under a linear projection (Figure 1 shows four representative panels; the full 18-model grid is in Appendix P).

Family	Model	Par.	Pr.	Out.	Gap
Llama 3.x	3.2-1B	1.2	.79	.09	.70
	3.2-3B	3.2	.83	.18	.65
	3.1-8B	8.0	.88	.25	.63
Gemma 4	E2B	2.0	.61	.27	.34
	E4B	4.0	.65	.33	.33
	26B-A4B	4.0*	.69	.42	.27
	31B	33.0	.69	.43	**.26**
Phi 4	mini	3.8	.83	.20	.63
	14B	14.0	.86	.31	.55
Qwen 1.x	1.5B	1.5	.75	.12	.63
	7B	7.0	.80	.24	.57
Qwen 3.6	35B-A3B	3.0*	.83	.38	.46
	27B	27.0	.88	.35	.53
Yi 1.5	6B	6.0	.77	.13	.64
	9B	9.0	.77	.18	.60
	34B	34.0	.83	.26	.57
OLMo 3.1	32B	32.0	.86	.32	.54
Tiny Aya	Global	3.0	.81	.17	.64

Table 2: Decoding flattening across the 18-model sweep. **Pr.** = linear-probe peak (10-way, chance .10, surface baseline .60). **Out.** = best-of-(NL, EN) majority-correct ($\geq 3/5$ paraphrases). **Gap** = Pr. − Out.; bold marks the smallest gap (Gemma-4-31B). *Active params for MoE.

Cell-wise, the loss has a single locus. *DecodingSuppressed is the largest cell in every single model*, with shares 51–76% (mean 0.65): the probe reads the culture but the generation emits a wrong name. *RepresentationallyFlat*, the cell predicted by a strong "LLMs are translation machines" reading, is consistently smaller (0.10–0.33, mean 0.18): the cultural distinction is rarely missing from the residual stream, it is just rarely produced. Per-model trajectories, stacked decomposition shares, per-(model, culture) heatmaps, the direction-of-flattening confusion matrix, per-motif Preserved share, and full per-model results are in Appendices N–Q.

Answer to RQ1. All 18 models clear the surface baseline on probe accuracy but not on generation; DecodingSuppressed dominates every model's 2×2.

4.2 RQ2: Layer-wise emergence

Decoding depth is uniformly late, encoding depth varies by family. Across all 18 models, the logit lens (E2) crosses its 5%-in-top-1 onset in the last 12% of layers on every single model (onset range 0.88–0.97, median 0.96; for Qwen-3.6-27B the bare-prompt lens is used because its chat template emits a generic "Here" token at every layer, App. F).

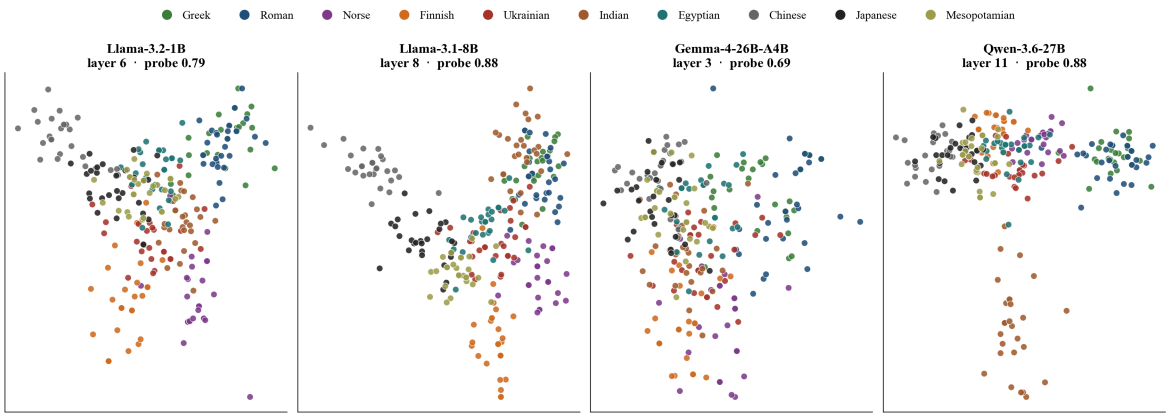

Figure 1: Latent-space projection at each model's probe-peak layer (50-dim PCA → LDA on culture labels) for four representative models from the 18-model sweep; the 270 (motif, culture) entities are colored by culture (legend on top). Cultures form visibly distinguishable clusters in the residual stream of every model, regardless of family or scale, even when the corresponding output accuracy collapses to a Greco-Roman default. The full 18-model grid is in Appendix P (Figure 13).

The linear probe peak depth is more spread across families (Figure 2). Every model sits on or above the diagonal of the encoded-vs-decoded scatter: the residual stream becomes culturally legible before the network commits to the output token, and the late-decoding band is robust across 1.2B→ 32B, dense vs MoE, and chat-tuned vs base architectures.

The per-culture lens schedule is not uniform: Greek and Roman cross the 5%-in-top-5 threshold first, while the other eight cultures cross later by ∼ 1.3 layers on a 32-layer model on average (App. F).

Answer to RQ2. Lens decoding sits in the last 12% of layers across all 18 models; every model encodes culture at or before that depth, and the encode-then-decode ordering is robust across the sweep.

4.3 RQ3: Causal localization

The causal locus of the cultural readout coincides with the lens decoding band, not the probe encoding band. Activation patching (E3, §3.3) measures, for each cross-cultural prompt pair and each layer, whether swapping the source-culture residual into the target-culture prompt causes the source-culture gold name to receive a higher next-token logit than the target-culture one. Across all 18 models, this preference-flip rate sits at the baseline level through the early third of the network, rises through the middle, and peaks in the last quarter (peak depth 0.75–1.00, median 0.89; peak rate 0.40–0.95, median 0.75; median lift over baseline +0.55). The same depth band shows the lens onset (§4.2); the culture direction is therefore not just decoded late but also *causally bound to the output* late.

Per-family pattern. Peak depth is concentrated near the network end across families (Llama 3.x 0.86–1.00, Gemma 4 0.88–1.00, Phi 4 0.78–0.86, Qwen 1.x 0.85–1.00, Qwen 3.6 0.89–0.95, Yi 1.5 0.86–1.00, OLMo 0.87). **Qwen-3.6-27B** is reported on a bare-prompt estimate. Chat-template patching scores it as flat ≈ 0% because the swap position decodes to the generic token `Here` in 320/320 rows (same artefact as its lens, App. F); switching to bare-prompt the lens recovers and the layerwise culture-preference rate peaks at 0.95 at depth 0.95 (App. G), matching the other 17 models. Explicit bare-prompt patching is the camera-ready remediation.

Answer to RQ3. The causal locus of the cultural readout sits in the late-decoding band (7× above the early-network baseline); the encode-vs-decode separation is mechanistic, not merely correlational.

4.4 RQ4: Within-family scaling shrinks but does not close the gap

Scaling helps but does not close the gap. Table 3 reports smallest-to-largest probe peak, output, and gap for the five families with more than one completed size (Llama 3.x, Phi 4, Gemma 4, Yi 1.5, Qwen 1.x): in every family both probe and output rise with parameters and the probe−output gap shrinks; none close it. Llama 3.x gives the cleanest signal across three sizes (1.2B→ 8B): probe peak 0.79 → 0.88, output 0.09 → 0.25, gap 0.70 → 0.63 monotonically. *Even at 8B the gap is*

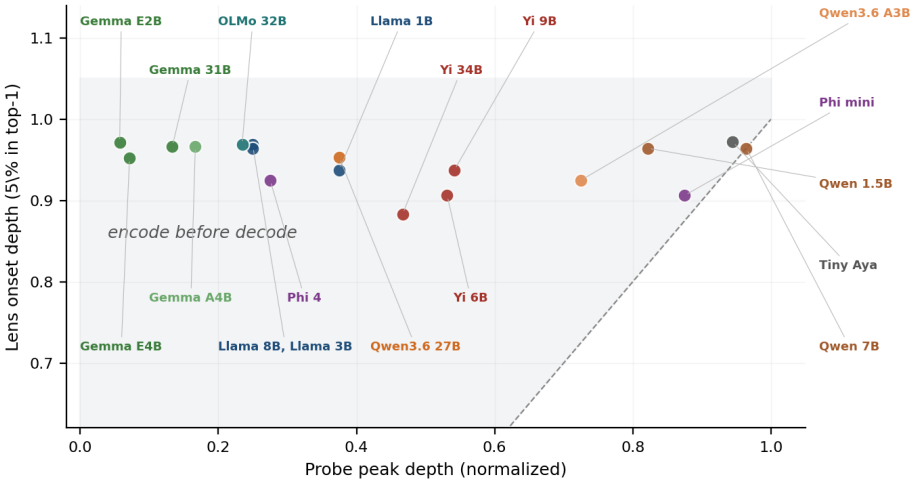

Figure 2: Encoded depth (linear-probe peak) vs decoded depth (logit-lens 5%-in-top-1 onset) per model. Every model is on or above the diagonal: culture is read out of the residual stream before the model commits to emitting the right token. Encoding depth spreads across families (Gemma early, Yi mid, Phi-mini / Tiny-Aya / Qwen-7B late); decoding depth is uniformly in the last 12% of layers. Lens-onset bar chart and per-culture emergence curves in Appendix F.

still 0.63, and a naive log-parameter extrapolation predicts only $\approx +0.10$ output gain per decade of parameters, insufficient to close the gap within an order of magnitude. Per-family scaling curves with both NL and EN modes in Appendix H.

Answer to RQ4. The probe−output gap shrinks within every family but closes in none; log-parameter extrapolation does not predict closure within an order of magnitude.

Family	Probe	Output	Gap
Llama 3.x (1B→ 8B)	.79 → .88	.09 → .25	.70 → .63
Phi 4 (3.8B→ 14B)	.83 → .86	.20 → .31	.63 → .55
Gemma 4 (E2B→ 31B)	.61 → .69	.27 → .36	.34 → .33
Yi 1.5 (6B→ 34B)	.77 → .83	.13 → .26	.64 → .57
Qwen 1.x (1.5B→ 7B)	.75 → .80	.12 → .24	.63 → .57

Table 3: Within-family scaling endpoints: probe peak, best-of-(NL, EN) majority-correct output, and gap, smallest→largest size in each family. The gap shrinks with scale in every family; no family closes it.

4.5 RQ5: Cross-lingual querying conditions the readout

The cultural readout from the residual stream is language-conditioned: under English query and under in-culture query the model gates partially disjoint subsets of the same representation. For each (motif, culture) we pose five paraphrases in English (EN) and five in the target culture's native language (NL): Modern Greek, Italian, Norwegian Bokmål, Finnish, Ukrainian, Hindi, Modern Standard Arabic, Mandarin, Japanese, and, for Mesopotamian, English with an explicit "(Sumerian-Akkadian)" parenthetical (the only fallback case; both substrate languages extinct without continuous spoken descendant). In every case the NL language is the *living* language of present-day reception of the canon, not the historical language of original composition (Modern Greek over Ancient Greek, Italian over Latin, Hindi over Sanskrit, etc.); the per-culture rationale is in Appendix I. The prompt protocol is in the same appendix.

Per-culture asymmetry. Aggregated across the chat-template models with both modes, EN wins on average by $+0.04$ majority-accuracy (English mean 0.23, native mean 0.19); three exceptions (Gemma-4-31B, Qwen-3.6-27B, Tiny-Aya-Global) prefer NL. The aggregate hides substantial per-culture variation (Table 4, Figure 3). Greek and Egyptian show extreme English advantage; Finnish and Chinese show native winning; the remaining six cultures sit roughly ambidextrous. Three checks bound the language-proficiency confound on this contrast (full discussion in the Limitations section, Appendix J, Table 10): the Mesopotamian within-English fallback gives a paraphrase-engineering noise floor of $|\overline{\Delta}|_{\text{Mesop}} \approx 0.04$; Tiny-Aya-Global (Aryabumi et al., 2024a,b), with explicit pre-

training coverage of all 8 of our non-CJK non-fallback NL languages, produces the smallest NL−EN gap in the sweep (0.004); and tokenizer fertility does not predict the per-cell EN−NL delta (V6).

Culture	$\overline{\Delta}_{\text{EN-NL}}$	Reading
Greek	+0.241	EN canon
Egyptian	+0.177	EN canon
Roman	+0.085	mild EN
Japanese	+0.040	mild EN
Norse	+0.026	mild EN
Ukrainian	+0.016	mild EN
Indian	−0.008	tie
Mesopotamian	−0.037	within-EN control
Finnish	−0.058	native canon
Chinese	−0.071	native canon

Table 4: Per-culture EN−NL majority-accuracy delta, averaged across chat-template models with both modes. Positive: English query wins; negative: native query wins.

Within-language failures correlate, cross-language failures decouple. A sharper question sits below the aggregate deltas: at the (motif, culture) cell, comparing within-mode correctness correlation (5 paraphrases of one language) against cross-mode correlation (5 NL × 5 EN) isolates language-conditioning from paraphrase-format sensitivity. Across the 18 chat-template models with both modes, $\overline{r}_{\text{within}} = 0.57$ and $\overline{r}_{\text{cross}} = 0.29$ (gap 0.27, $p \leq 0.01$ per model by permutation, App. L): cross-language queries are roughly 2× more decoupled than within-language ones. This structural test is robust to the language-proficiency confound by construction (see the Limitations section). The gap does not exhibit a cross-family scaling law (Spearman $\rho = -0.08$, $p = 0.81$); it tracks multilingual share of the chat-tuning corpus rather than parameter count, consistent with Tiny-Aya-Global's small gap (0.19) at 3B (Table 12 in App. M).

Mechanistic interpretation. The within−cross gap is direct evidence that DecodingSuppressed is language-conditional: English query and in-culture query gate partially disjoint subsets of the same residual. This is a quantitative pushback on a strong reading of "LLMs are translation machines" (Wendler et al., 2024): if the residual were genuinely translation-flat, RepresentationallyFlat (0.10–0.33) would dominate over DecodingSuppressed (0.51–0.76); it does not. "Translation machine" describes the unembedding's behavior, not the residual stream's contents.

Answer to RQ5. The within−cross correlation gap and the canon-language per-culture deltas together show a language-conditioned readout, not a translation-flat one.

Bilingual ensembling. The disjunction of the two query modes (5 NL OR 5 EN paraphrases) lifts cell recovery by +0.08 absolute, +36% relative over the best single mode. A killer-control isolates the contribution of language-switching from raw paraphrase-count: on the 12 chat-template models for which we collected a second batch of 5 same-language English paraphrases (Appendix I), we compare the bilingual ensemble NL5∪EN5 against the same-language EN10 ensemble. The two ensembles *dissociate by metric*. Under any-of-10 cell recovery, NL5 ∪ EN5 beats EN10 by +0.053 absolute on every model (mean 0.50 vs 0.45); under the per-mode majority union ($\geq 3/5$ in NL OR $\geq 3/5$ in EN) the same direction holds with +0.052 (0.28 vs 0.23). Conversely, under a strict majority-of-10 threshold ($\geq 6/10$ paraphrases correct), EN10 wins by 0.038 (0.13 vs 0.17): same-language doubling builds strict consensus that cross-language ensembling does not. The language-switching contribution is to *breadth*, not strict consensus.

5 Evaluation

Six controls bound the headline claims (full tables in App. D–J). **V1 (char-n-gram baseline).** All 18 residual-stream probes clear the 0.596 romanized-name baseline by +0.02 to +0.29 (mean +0.19); the DecodingSuppressed claim holds. **V2 (scoring strictness).** Re-scoring all 48,568 generations under substring-only, first-word EM, and strict EM preserves per-model rankings at Pearson ≥ 0.87, Spearman ≥ 0.80 (Table 13). **V3 (bootstrap CIs).** Per-(model, mode) 95% CIs have median half-width 0.046, exceeded by all headline scaling and per-culture extremes. **V4 (permutation test).** Per-model permutation of the 10 paraphrase-column labels gives one-sided $p \leq 0.01$ for the within−cross gap on every chat-template model. **V5 (cluster-failure share).** Cells unanimously wrong across all five paraphrases of a mode emit the same first word in only $\overline{c} = 0.011$: failures are exploratory, not convicted. **V6 (language-proficiency proxy).** Per-language tokenizer fertility on 100 Belebele passages (Bandarkar et al., 2024; Team et al., 2022) does not correlate with the per-cell EN−NL delta ($r = -0.004$, $n = 120$);

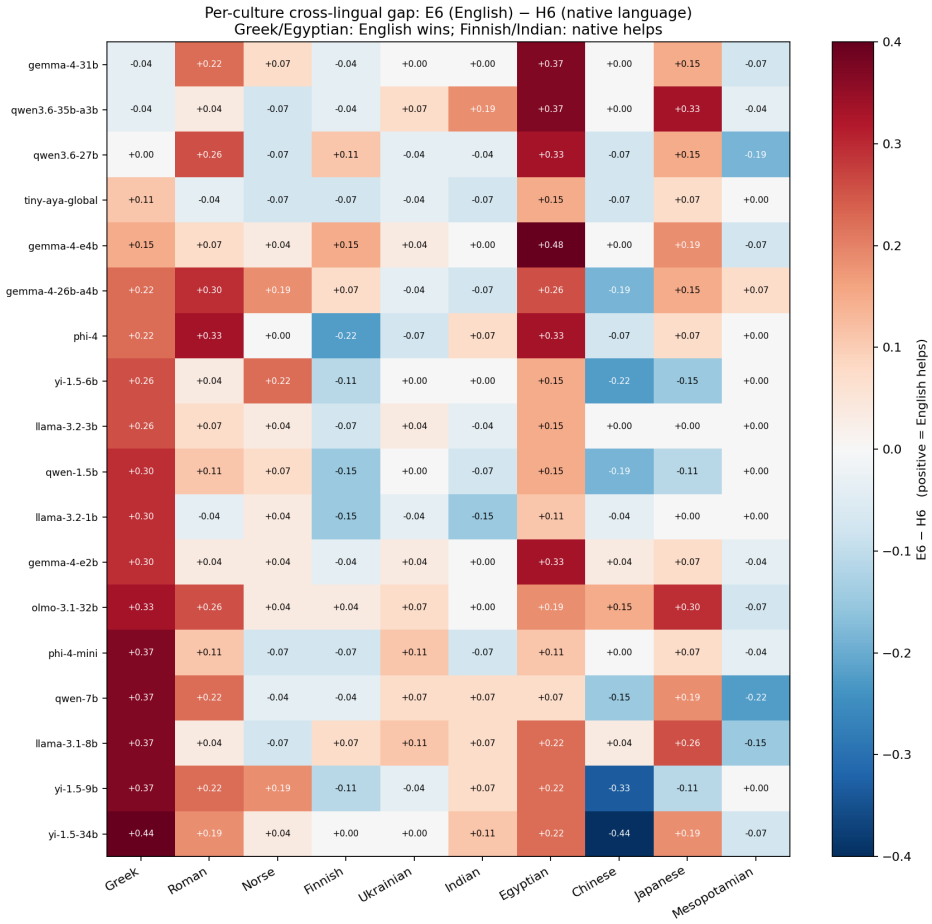

Figure 3: Per-(model, culture) EN−NL majority-accuracy delta. Red: English query wins; blue: native query wins; white: tie. The pattern tracks the *language in which each mythological canon is documented*, not the script of the NL language nor the speaker base of the target language. Per-model aggregate NL vs EN and the NL leaderboard in Appendix K.

Tiny-Aya-Global (Aryabumi et al., 2024a,b) produces $NL \approx EN$ despite symmetric coverage of all 10 target languages, ruling out general multilingual proficiency as the explanatory factor.

6 Discussion and Conclusion

The mechanistic locus of cultural flattening is the unembedding, not the residual stream, with implications for mitigation targeting. Output-side interventions (retrieval, reranking, instruction tuning on cultural QA, the bilingual ensemble of §4.5) target the failure locus. Encoder-side interventions like continued pretraining on under-represented corpora (Hämäläinen, 2024) push on a side less broken than the "low-resource" framing suggests: they help on RepresentationallyFlat cells (0.10–0.33) but cannot directly affect the dominant DecodingSuppressed cells (0.51–0.76). The same encoder/decoder split is the natural diagnostic for the parallel question of whether post-training compression hits culturally-grounded knowledge harder than English defaults, taken up in a separate study (Siniaev et al., 2026) anchored in the quantization literature (Dettmers et al., 2022; Frantar et al., 2023; Marchisio et al., 2024).

Across 18 open-source LLMs, culture is consistently *represented* (probe peak 0.61–0.88, +0.02–+0.29 above the surface baseline) and consistently *not decoded* (output 0.09–0.43). Decoding-Suppressed dominates every model (0.51–0.76); within-family scaling narrows but does not close the gap. The locus is the unembedding, not the encoder, and the unembedding is itself language-conditioned ($\overline{r}_{\text{within}} = 0.57$ vs $\overline{r}_{\text{cross}} = 0.29$, $p \leq 0.01$ per model).

Limitations

Substrate and ground truth. Canonical entity-to-culture assignments are genuinely contested at boundaries: Odin and Thor are both Norse sky figures, the Slavic Perun is shared between Ukrainian and West Slavic traditions, and a number of cells use a less-canonical or scholarly-contested filler (e.g., Khors vs Dažbog as sun-disk vs sun-deity in Slavic, Mielikki vs Rauni for the Finnish T110 forest-spirit). We follow modern scholarly consensus where it has shifted but cannot make every assignment unambiguous. Four cells are flagged as structural absences (Japanese flood A1010, Finnish flood A1010, Mesopotamian theft-of-fire A1415, Egyptian forest-spirit F460) and excluded from output scoring; we cannot rule out that other absences exist that we have catalogued as canonical fillers. The ten cultures span six language families but deliberately exclude sub-Saharan African, Native American, Polynesian, Australian, and Arctic indigenous traditions; the headline DecodingSuppressed claim is unlikely to reverse on these, but the per-culture NL/EN deltas might, especially on traditions whose canon is documented predominantly in non-Anglocentric sources.

Native-language prompts. NL prompts are native re-renderings of the English EN template, not native-speaker-validated translations. The two authors are native speakers of one NL language each; the remaining seven were translated by LLM-assisted forward-then-back-translation against the canonical entity name, which preserves the entity but may shift register or naturalness. Per-culture EN−NL deltas (§4.5) are therefore directionally trustworthy at the gross level but await a native-speaker review for absolute calibration on Greek, Italian, Norwegian Bokmål, Hindi, Modern Standard Arabic, Mandarin, and Japanese.

Language-proficiency confound and how we control for it. The NL (target-language) condition asks each model to read prompts in a language that may not be in its pretraining target set. Three of our 18 models are Yi-1.5 (6B/9B/34B), explicitly bilingual English+Chinese; OLMo-3.1-32B is explicitly English-primary (the OLMo 2 report states "OLMo 2 is not trained for multilingual tasks"); Phi-4 14B is described as "primarily English." For these Tier B/C models (Appendix J, Table 10), the NL condition on languages outside the model's pre-training target is genuinely confounded with non-capability. We hold the cross-lingual claim in place by means of three artefacts that come for free from the experimental design rather than from new evaluations: (i) a *positive control*, Tiny-Aya-Global, with explicit pre-training coverage of all 8 of our non-CJK non-fallback target languages and the smallest NL−EN gap in the sweep (0.004); if the cross-lingual asymmetry were just non-capability, Aya should be the *most* biased, not the least. (ii) A *negative control*, the Mesopotamian fallback, which uses English with a culture-marking parenthetical and bounds the within-English paraphrase-engineering component of any cross-mode signal at $|\overline{\Delta}|_{\text{Mesop}} \approx 0.04$, an order of magnitude below the canon-language extremes. (iii) A *structural test*, the within-vs-cross paraphrase correlation gap (§4.5, App. L), which is robust to non-capability by construction: a Tier B/C model that cannot read Finnish has its five Finnish paraphrases fail *together* on most cells, which *increases* the within-mode correlation toward 1, not lowers it; the observed asymmetry within $= 0.57$ vs cross $= 0.29$ is the structural signature of language-conditioned readout, not of non-capability. We further note that the headline RQ1–RQ4 results (§4) use only English (EN) prompts and probe activations from English-rendered inputs, so the headline "representation-preserved-but-decoding-suppressed" claim is unaffected by NL confounds at any tier; English coverage is universal across the 18 models.

Probe class and decoding policy. The residual-stream probe is a linear ridge classifier, chosen for transparency and to avoid over-fitting probe non-linearity to surface co-occurrences (Hewitt and Liang, 2019). An MLP probe would lift absolute probe-peak accuracy but cannot, by construction, change the qualitative outcome that the residual is more separable than the output is correct. Output extraction is greedy chat-template generation; beam search or nucleus sampling would shift output numbers but the per-(model, culture) ranking is robust across the strictness audit (V2, §5).

Compute and precision. All experiments run at fp16 with no further quantization, to avoid confounding cultural recall with quantization-induced degradation. The natural follow-up question is whether post-training quantization (4-bit, 8-bit) hits DecodingSuppressed cells harder than Preserved cells; this is left to a separate study (see §6). Activation patching was run with a sparse layer step (every 4th layer) on a default of 5 motifs $\times$ 4 culture pairs to stay within the per-model A100 budget; a dense full-coverage patching pass might shift

peak-depth estimates by a layer or two but does not change the late-decoding band claim.

Closed-source and frontier models. Every instrument we use except EN (output extraction) requires white-box access. The headline claim (DecodingSuppressed dominates) is therefore established directly only for the 18 open-source models. The NL/EN independence prediction, which depends only on output behavior, generalizes to closed-source frontier models and is the natural sanity check for the strongest reading of the result.

Reasoning-mode models. Qwen 3.6 ships a default "thinking" trace that we disable for scoring (§3); the lens template artifact on Qwen-3.6-27B (App. F) is also a reasoning-template phenomenon. Whether the DecodingSuppressed dominance holds when these models are run in their preferred reasoning mode is open; our bare-prompt lens result for Qwen-3.6-27B is consistent with it holding, but is not a sufficient test on its own.

Anonymization. Code, prompts, per-model outputs, and the rescored aggregates will be released at the anonymized URL upon acceptance; until then, the in-paper per-entity listing in Appendix B is the only release artefact directly linked from the submission.

AI Models Usage

As non-native English speakers, we used Claude Opus 4.7 for text editing. GitHub Copilot assisted with code completion. For research automation, we used Claude Code integrated with html-based reports for experimental pipelines and iterative analysis. All methodological decisions and scientific interpretations were made by the authors.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *Preprint*, arXiv:2305.13707.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, and 14 others. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024a. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024b. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. Which humans?

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 749–775. Association for Computational Linguistics.

Mary Beard, John North, and Simon Price. 1998. *Religions of Rome*, volume 1–2. Cambridge University Press, Cambridge.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Anne Birrell. 1993. *Chinese Mythology: An Introduction*. Johns Hopkins University Press, Baltimore.

Jeremy Black and Anthony Green. 1992. *Gods, Demons and Symbols of Ancient Mesopotamia: An Illustrated Dictionary*. British Museum Press, London.

Aleksander Brückner. 1985. *Mitologia słowiańska*, reprint edition. Państwowy Instytut Wydawniczy, Warsaw. Original published 1918.

Walter Burkert. 1985. *Greek Religion*. Harvard University Press, Cambridge, MA.

Iaroslav Chelombitko, Mika Hämäläinen, and Aleksey Komissarov. 2026. Subword-based comparative linguistics across 242 languages using wikipedia glottosets. *Preprint*, arXiv:2601.18791.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *Preprint*, arXiv:2208.07339.

Wendy Doniger. 1981. *The Rig Veda: An Anthology*. Penguin Classics, London.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Gavin Flood. 1996. *An Introduction to Hinduism*. Cambridge University Press, Cambridge.

Benjamin R. Foster. 2005. *Before the Muses: An Anthology of Akkadian Literature*, 3rd edition. CDL Press, Bethesda.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *Preprint*, arXiv:2210.17323.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. *Preprint*, arXiv:2303.02536.

Andrew R. George. 2003. *The Babylonian Gilgamesh Epic: Introduction, Critical Edition and Cuneiform Texts*. Oxford University Press, Oxford.

Aleksander Gieysztor. 2006. *Mitologia Słowian*, 3rd edition. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pierre Grimal. 1996. *The Dictionary of Classical Mythology*. Blackwell, Oxford.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *Preprint*, arXiv:2310.02207.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *Multilingual Facilitation: Honoring the Career of Jack Rueter*.

Mika Hämäläinen. 2024. LLMs will be the future of NLP for endangered languages. In *Lightning Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 14–18.

Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, Yuri Bizzoni, Jack Rueter, and Niko Partanen. 2024. The growing importance of humanities for NLP in the era of LLMs. In *Lightning Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 2–6.

Robin Hard. 2004. *The Routledge Handbook of Greek Mythology*. Routledge, London.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Folgert Karsdorp and Antal van den Bosch. 2013. Identifying motifs in folktales using topic models.

Thomas P. Kasulis. 2004. *Shinto: The Way Home*. University of Hawaii Press, Honolulu.

David Kinsley. 1986. *Hindu Goddesses: Visions of the Divine Feminine in the Hindu Religious Tradition*. University of California Press, Berkeley.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Jennifer Larson. 2007. *Ancient Greek Cults: A Guide*. Routledge, London.

John Lindow. 2001. *Norse Mythology: A Guide to Gods, Heroes, Rituals, and Beliefs*. Oxford University Press, Oxford.

Michał Łuczyński. 2020. *Bogowie dawnych Słowian: studium onomastyczne*. Kieleckie Towarzystwo Naukowe, Kielce.

Vettam Mani. 1975. *Purānic Encyclopaedia*. Motilal Banarsidass, Delhi.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual llms? *Preprint*, arXiv:2407.03211.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Juha Pentikäinen. 1999. *Kalevala Mythology*, expanded edition. Indiana University Press, Bloomington.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *Preprint*, arXiv:1909.01066.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Preprint*, arXiv:2305.15425.

Donald L. (translator) Philippi. 1968. *Kojiki*. University of Tokyo Press, Tokyo.

Serhii Plokhy. 2015. *The Gates of Europe: A History of Ukraine*. Basic Books, New York.

Risto Pulkkinen. 2014. *Suomalainen kansanusko: samaaneista saunatonttuihin*. Gaudeamus, Helsinki.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Anna-Leena Siikala. 2012. *Itämerensuomalaisten mytologia*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Rudolf Simek. 2007. *Dictionary of Northern Mythology*. Boydell & Brewer, Cambridge.

Viacheslav Siniaev, Iaroslav Chelombitko, and Aleksey Komissarov. 2026. Compressed code: the hidden effects of quantization and distillation on programming tokens. *Preprint*, arXiv:2601.02563.

Jerzy Strzelczyk. 1998. *Mity, podania i wierzenia dawnych Słowian*. Dom Wydawniczy Rebis, Poznań.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Stith Thompson. 1955. *Motif-Index of Folk-Literature*. Indiana University Press.

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.

Oleksa Voropay. 1958. *Zvychayi nashoho narodu: etnohrafichnyi narys*. Ukrayinske Vydavnytstvo, Munich.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *Preprint*, arXiv:2211.00593.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Richard H. Wilkinson. 2003. *The Complete Gods and Goddesses of Ancient Egypt*. Thames & Hudson, London.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lihui Yang and Deming An. 2005. *Handbook of Chinese Mythology*. ABC-CLIO, Santa Barbara.

A Dataset card: 270-entity, 10-culture extension

This appendix is a Bender-Friedman-style dataset card (Bender and Friedman, 2018; Gebru et al., 2021) for the 270-entity, 10-culture substrate (Greek, Roman, Norse, Finnish, Ukrainian, Indian, Egyptian, Chinese, Japanese, Mesopotamian).

Motivation. The dataset is constructed to support the four mechanistic experiments (E1/E2/E3/EN) under a parallel cross-cultural alignment substrate: each row is the canonical filler of a single Thompson-index motif role in a single culture, so a probe asked to discriminate cultures must do so on cultural rather than topical grounds. The unit of analysis is the (motif, culture) cell.

Composition. 27 motifs $\times$ 10 cultures $=$ 270 cells. Cell schema: `motif_id` (Thompson index entry), `culture`, `name` (canonical proper name in romanized English), `native` (native-script form when applicable), `description` (motif role gloss in English), `source` (per-row citation anchor pointing into Table 5), and a primary-text excerpt where the source is a primary attestation rather than a monograph. The motif set spans cosmology (A220, A411, A420, A510, A671, A710, A1010, A1141.2, A142.1, A1415), tabu and transformation (C310, C920, D150, D610, D1080, D1810), the dead (E200, E400), water/forest/field/household spirits (F420.1.2, F450, F460, F470), giants and demons (G300, G303), love and fertility (T110), sacred trees (V1), and personified death (Z111).

Citation provenance. Each row points to either a primary text or a curated academic monograph (§3.1; per-entity listing in Appendix B). The per-tradition source repertoire is summarized in Table 5.

Contested rows and cultural absences. Three classes of decision are explicit in the dataset. First, where the modern scholarly literature has settled a contested attribution differently from older sources, we follow the modern reading: Lada (T110, Ukrainian) is replaced by Mokoš following Łuczyński (2020), who shows that Lada is a Slavic folk-song refrain word reinterpreted as a theonym in 19th-century scholarship; the Khors / Dažbog distinction (sun-disk A220 vs sun-as-deity A710) is preserved as two separate motif rows following Gieysztor (2006). Second, a small number of cells use a less-canonical or scholarly-contested

Culture	Primary texts	Academic monographs
Greek	Hesiod, Homer, Ovid *Met.*	Hard (2004), Burkert (1985), Larson (2007)
Roman	Virgil, Ovid, Livy	Grimal (1996), Beard et al. (1998)
Norse	Poetic Edda, Prose Edda	Lindow (2001), Simek (2007)
Finnish	Kalevala (Lönnrot 1849, named runot)	Pentikäinen (1999), Siikala (2012), Pulkkinen (2014)
Ukrainian	*Primary Chronicle*, Hypatian Codex	Gieysztor (2006), Łuczyński (2020), Brückner (1985), Strzelczyk (1998), Voropay (1958), Plokhy (2015)
Indian	Rig Veda, Purānas, Mahābhārata	Mani (1975), Doniger (1981), Flood (1996), Kinsley (1986)
Egyptian	Pyramid Texts, Coffin Texts, Book of the Dead	Wilkinson (2003)
Chinese	*Shan Hai Jing*, *Huainanzi*	Birrell (1993), Yang and An (2005)
Japanese	Kojiki, Nihon Shoki	Philippi (1968), Kasulis (2004)
Mesopotamian	Enuma Elish, Epic of Gilgamesh	Black and Green (1992), Foster (2005), George (2003)

Table 5: Per-tradition source repertoire for the ground-truth dataset. The per-entity `source` field anchors each of the 270 rows to a specific entry in either column (e.g. `Hard 2004` for Greek Zeus, `Kalevala 9:1-13` for the Finnish thunder god); the full mapping appears in Appendix B.

filler (notably Mielikki / Rauni for the Finnish T110 cell, sitting on a still-debated boundary in current Finno-Ugric scholarship, flagged for the second author's native-fluent reading rather than fixed in this release). Third, four cells are recorded as structural absences where the source tradition genuinely lacks a parallel to the Thompson motif role (Japanese flood A1010, Finnish flood A1010, Mesopotamian theft-of-fire A1415, Egyptian forest-spirit F460); they are excluded from output-accuracy scoring, and the treatment of cultural absences as first-class entries follows Hämäläinen et al. (2024).

Languages and scripts. Native-script forms cover Greek, Latin (Italian, Norwegian Bokmål, Finnish, romanized Indian/Egyptian/Mesopotamian/Roman), Cyrillic (Ukrainian), Devanagari (Hindi for Indian), Modern Standard Arabic (for Egyptian, despite no continuous descendant), Han (Mandarin for Chinese), and a Han/Kana mixture (Japanese). The per-culture NL query language is named inline in Appendix I.

Limitations. The canonical assignment of an entity to a culture is contested at boundaries (see the Limitations section of the main paper); the dataset records the most-cited canonical figure per culture. Cross-tradition borrowings are not separately tracked, a Norse entity that is structurally a continuation of a Proto-Indo-European figure shared with Vedic is recorded only on the Norse side. The ten cultures span a deliberately wide range (Indo-European Greek/Roman/Norse/Indian/Ukrainian, Uralic Finnish, Afro-Asiatic Egyptian, Sino-Tibetan Chinese, Japonic Japanese, extinct Akkadian/Sumerian Mesopotamian) but exclude major traditions (sub-Saharan African, Native American, Polynesian) for the practical reason that the per-row citation discipline could not be sustained at scale within the present cycle; we treat extending the substrate to those traditions as a clear next step in line with the cross-cultural NLP program articulated by Hämäläinen et al. (2024).

Release. The dataset, the prompt protocol, all per-model E1/E2/E3/E4 outputs, and the rescored aggregates ship under MIT for code and CC-BY-4.0 for the dataset. Three artefacts are published in parallel: an anonymized code release (code, prompts, per-model outputs); a Hugging Face dataset card; and the complete per-entity listing in Appendix B of this paper, which contains the canonical name, native-script form, and per-row citation anchor for every one of the 270 entities.[4]

Citation anchors. Of the 270 (motif, culture) pairs, every row carries a non-empty source field anchoring it to either a modern academic monograph or a primary-text reference (Kalevala runa, Pyramid Texts spell, Hypatian Codex annal entry, Ovid *Metamorphoses* book and lines, and so on). The full mapping appears in Appendix B; in this paper we cite the academic sources collectively per tradition, not per entity, but the per-row trace is in the released artefact. A summary by tradition: 13 sources for Greek/Roman, 4 for Norse, 4 for Finnish, 6 for Ukrainian/Slavic (Polish + Ukrainian + Plokhy historical), 4 for Indian, 1 (Wilkinson) plus per-row

[4] Both the code release and the Hugging Face dataset card will be linked from anonymized URLs upon acceptance.

Pyramid/Coffin/Book-of-the-Dead spell pointers for Egyptian, 2 for Chinese, 2 for Japanese, 3 for Mesopotamian.

B Per-entity dataset listing

This appendix lists all 270 rows of the dataset described in §3.1 and Appendix A. Each row carries: the Thompson-index motif identifier (SMALL CAPS); the role description (English gloss); the culture label; the canonical proper name in romanized form (with diacritical marks where applicable to the source language); and the per-row source-citation anchor (modern academic monograph or primary text). Native-script forms (Greek polytonic, Cyrillic Ukrainian, Devanāgarī Hindi, Modern Standard Arabic, Han characters for Mandarin, Japanese kanji/kana, and historical Norse forms) are stored in the released CSV/JSON artefact; this table reproduces only the Latin-script names so it renders cleanly on any reader. The mapping from `source` field to bibliographic entry is implicit in the references of this paper for the modern-academic citations; primary-text citations follow the conventions of each tradition (e.g. `Kalevala` $n{:}m$–k = runo n, lines m through k, in the Lönnrot 1849 standard text; `PT` n = Pyramid Texts spell n).

Table 6: Full per-entity dataset listing for the 270-entity, 10-culture Thompson-motif substrate. Native-script forms are in the released CSV/JSON artefact. Superscripts mark contested or structural-absence rows (footnoted after the table).

Motif	Role	Culture	Name	Source
A1010	Deluge/Great Flood	Greek	Deucalion	Hard 2004
		Roman	Deucalion	Grimal 1996
		Norse	Bergelmir	Lindow 2001
		Finnish	(structural absence)[13]	(no canonical Finnish flood-survivor narrative)
		Ukrainian	Noi	Voropay 1958 *Zvychayi nashoho narodu*; Hnatyuk *Etnohrafichnyi zbirnyk* (Christianized flood-hero in Ukrainian folk tradition)
		Indian	Matsya	Matsya Purana 1-2; Bhagavata Purana 8.24; Mahabharata 3.187 (flood narrative)
		Egyptian	Hathor (Sekhmet)	*The Book of the Heavenly Cow* (New Kingdom funerary text)
		Chinese	Yu the Great	*Shujing* (Book of Documents, Yugong chapter); *Mengzi*; *Shanhaijing*
		Japanese	(structural absence)[6]	(no canonical Japanese deluge narrative)
		Mesopotamian	Atrahasis (Utnapishtim)	*Atrahasis Epic*; *Epic of Gilgamesh* tablet XI
A1141.2	Thunder-god's weapon (thunderbolt, hammer, axe)	Greek	Zeus	Hard 2004
		Roman	Jupiter	Grimal 1996
		Norse	Thor	Lindow 2001
		Finnish	Ukko	Pentikäinen 1999
		Ukrainian	Perun	Brückner 1918/1985 *Mitologia słowiańska*; Gieysztor 2006 *Mitologia Słowian* (Warsaw UP)
		Indian	Vajra	Rig Veda 1.32 (Indra slays Vritra with vajra); Mahabharata
		Egyptian	Set	Pyramid Texts; Coffin Texts; Book of the Dead (storm and disorder god)
		Chinese	Lei Gong (Duke of Thunder)	*Shanhaijing* (Classic of Mountains and Seas, c. 4 c. BCE); folk iconography
		Japanese	Takemikazuchi	*Kojiki* (712 CE) Kamiumi episode; *Nihon Shoki* (720 CE)
		Mesopotamian	Adad	*Enuma Elish*; storm-god hymns; royal inscriptions throughout Mesopotamian history
A1415	Theft of fire	Greek	Prometheus	Hard 2004
		Roman	Prometheus	Grimal 1996
		Norse	Loki	Lindow 2001
		Finnish	Väinämöinen	Kalevala runo 47
		Ukrainian	Svarog[7]	Gieysztor 2006 *Mitologia Słowian* (Warsaw UP)

continued on next page

(continued)

Motif	Role	Culture	Name	Source
		Indian	Mātariśvan	Rig Veda 1.31, 3.5, 3.9, 10.46 (Mātariśvan brings fire from heaven to the Bhrigus)
		Egyptian	Prometheus-equivalent[8]	(no Egyptian fire-thief narrative)
		Chinese	Suiren-shi	*Hanshu* (Han history); *Han Feizi*; folk tradition
		Japanese	Kagutsuchi[9]	*Kojiki*; *Nihon Shoki*
		Mesopotamian	(structural absence)[10]	(no Mesopotamian theft-of-fire narrative)
A142.1	Smith of the gods (divine craftsman)	Greek	Hephaestus	Hard 2004
		Roman	Vulcan	Grimal 1996
		Norse	Brokkr	Lindow 2001
		Finnish	Ilmarinen	Pentikäinen 1999
		Ukrainian	Svarog	Gieysztor 2006 *Mitologia Słowian* (Warsaw UP)
		Indian	Tvashtar	Rig Veda 1.32, 10.81 (forges Indra's vajra; 'all-fashioner')
		Egyptian	Ptah	Memphite Theology (Shabaka Stone, c. 700 BCE); Pyramid Texts
		Chinese	Lu Ban	*Lüshi Chunqiu*; *Mozi*; folk patron of craftsmen since Spring & Autumn period
		Japanese	Amatsumara	*Kojiki* (712 CE)
		Mesopotamian	Ea (Enki)	Sumerian *Enki and the World Order*; *Atrahasis*; *Enuma Elish*
A220	Sun-god	Greek	Apollo	Hard 2004
		Roman	Apollo	Grimal 1996
		Norse	Baldr	Lindow 2001
		Finnish	Päivä	Pentikäinen 1999
		Ukrainian	Khors[1]	Gieysztor 2006 *Mitologia Słowian* (Warsaw UP)
		Indian	Surya	Rig Veda 1.50 (Surya hymn); Surya Upanishad; Mahabharata
		Egyptian	Ra	Pyramid Texts; Litany of Ra; Book of the Dead
		Chinese	Xihe	*Shanhaijing*; *Huainanzi* (2 c. BCE)
		Japanese	Amaterasu	*Kojiki*; *Nihon Shoki*; *Engishiki*
		Mesopotamian	Utu (Shamash)	Sumerian and Akkadian sun-god hymns; *Code of Hammurabi* prologue
A411	Household god (domestic protective deity)	Greek	Hestia	Hard 2004
		Roman	Lares	Grimal 1996
		Norse	Tomte	Lindow 2001
		Finnish	Tonttu	Pentikäinen 1999
		Ukrainian	Domovyk	Voropay 1958 *Zvychayi nashoho narodu* (Customs of our people, Ukrainian diaspora ethnography); Brückner 1918/1985
		Indian	Vāstoṣpati	Rig Veda 7.54-55 (Vāstoṣpati hymn — protector of dwellings)
		Egyptian	Bes	Egyptian household amulets, popular religion (Middle Kingdom onwards)
		Chinese	Zao Jun (Stove God)	*Liji* (Book of Rites, 1 c. BCE); folk household tradition continuous since Han dynasty
		Japanese	Kamado-no-Kami	*Engishiki* (10 c. CE); folk tradition
		Mesopotamian	Lamassu	Mesopotamian protective-amulet inscriptions; royal palace gateway figures (Assyrian)
A420	God of war	Greek	Ares	Hard 2004
		Roman	Mars	Grimal 1996
		Norse	Tyr	Lindow 2001
		Finnish	Turisas	Finnish mythology
		Ukrainian	Perun	Brückner 1918/1985; Gieysztor 2006 *Mitologia Słowian* (Warsaw UP)
		Indian	Skanda	Mahabharata Vana Parva (Skanda's birth); Skanda Purana

continued on next page

(continued)

Motif	Role	Culture	Name	Source
		Egyptian	Montu	Pyramid Texts; royal Theban war-cult (Middle Kingdom)
		Chinese	Guandi (Guan Yu)	*Sanguozhi* (3 c. CE historical); *Sanguo Yanyi* (14 c. novel); state cult deification (Ming-Qing)
		Japanese	Hachiman	*Shoku Nihongi* (797 CE); Iwashimizu Hachiman shrine cult
		Mesopotamian	Ninurta	*Anzu Epic*; *Lugal-e* (Sumerian); royal inscriptions
A510	Origin of the culture hero	Greek	Heracles	Hard 2004
		Roman	Romulus	Grimal 1996
		Norse	Sigurd	Lindow 2001
		Finnish	Väinämöinen	Pentikäinen 1999
		Ukrainian	Kyi	Primary Chronicle s.a. 6th c. (Cross & Sherbowitz-Wetzor 1953); Plokhy 2017
		Indian	Manu	Manusmriti; Matsya Purana 1-2; Mahabharata 3.187 (flood-survivor – progenitor of humanity)
		Egyptian	Osiris	Pyramid Texts (Heliopolitan Osiris cycle); Plutarch *De Iside et Osiride* (1c. CE)
		Chinese	Fuxi	*Shiji* (Sima Qian, 1 c. BCE); *Yijing* attribution; *Huainanzi*
		Japanese	Ninigi-no-Mikoto	*Kojiki*; *Nihon Shoki*
		Mesopotamian	Gilgamesh	*Epic of Gilgamesh* (Old Babylonian + Standard Babylonian versions); Sumerian Bilgames poems
A671	Hell/Underworld (realm of the dead)	Greek	Hades	Hard 2004
		Roman	Pluto	Grimal 1996
		Norse	Hel	Lindow 2001
		Finnish	Tuoni	Pentikäinen 1999 (Tuoni, lord of Tuonela)
		Ukrainian	Veles	Gieysztor 2006 *Mitologia Słowian* (Warsaw UP); Strzelczyk 1998 *Mity, podania i wierzenia dawnych Słowian* (Poznań) (Veles as ruler of Nav)
		Indian	Yama	Rig Veda 10.14 (Yama-Yami hymn); Garuda Purana (preta-kalpa, narakas)
		Egyptian	Osiris	Pyramid Texts; Coffin Texts; Book of the Dead
		Chinese	Yanluo Wang (Yama-King)	Buddhist sutras (transmission of Yama as Yanluo); Daoist *Diyu* texts; folk *Yu Li Bao Chao*
		Japanese	Izanami	*Kojiki* book 1 (Izanami becomes ruler of Yomi after death); *Nihon Shoki*
		Mesopotamian	Ereshkigal	*Descent of Inanna*; *Nergal and Ereshkigal*; *Epic of Gilgamesh* tablet XII
A710	Creation/origin of the sun	Greek	Helios	Hard 2004
		Roman	Sol	Grimal 1996
		Norse	Sól	Lindow 2001
		Finnish	Päivätär	Pentikäinen 1999
		Ukrainian	Dazhboh	Gieysztor 2006 *Mitologia Słowian* (Warsaw UP); Strzelczyk 1998
		Indian	Vivasvant	Rig Veda 10.17.1-2 (origin of Vivasvant); 1.115 (his radiant chariot)
		Egyptian	Atum	Heliopolitan creation account; Pyramid Texts; Coffin Texts; Book of the Dead
		Chinese	Pangu	*Sanwu Liji* (3 c. CE, Xu Zheng); *Wuyun Linian Ji*
		Japanese	Amaterasu (cave myth)	*Kojiki*; *Nihon Shoki*
		Mesopotamian	Utu's daily journey	Sumerian Utu hymns; Akkadian Shamash hymns
C310	Tabu: looking at certain person or thing	Greek	Orpheus	Hard 2004

continued on next page

(continued)

Motif	Role	Culture	Name	Source
		Roman	Orpheus	Grimal 1996
		Norse	Baldr	Lindow 2001
		Finnish	Lemminkäinen	Kalevala
		Ukrainian	Kupala[12]	Strzelczyk 1998 *Mity, podania i wierzenia dawnych Słowian*; Voropay 1958
		Indian	Ahalya	Ramayana Bala Kanda 47-48 (Ahalya's curse and stoning)
		Egyptian	Hathor (mirror-tabu)	Egyptian wisdom literature; *The Tale of the Two Brothers* (Papyrus d'Orbiney, c. 1185 BCE)
		Chinese	Niulang and Zhinü	Han dynasty *Gushi Shijiu Shou* (Nineteen Old Poems); Tang/Song *Qixi* festival traditions
		Japanese	Izanagi at Yomi	*Kojiki* book 1; *Nihon Shoki*
		Mesopotamian	Adapa	*Adapa Epic* (Akkadian, c. 14 c. BCE)
C920	Death for breaking tabu	Greek	Semele	Hard 2004
		Roman	Semele	Grimal 1996
		Norse	Baldr	Lindow 2001
		Finnish	Kullervo	Kalevala runo 31-36
		Ukrainian	Marena	Brückner 1918/1985 *Mitologia słowiańska*; Strzelczyk 1998
		Indian	Daksha	Linga Purana; Vayu Purana 30; Bhagavata Purana 4.2-7
		Egyptian	Bata	*The Tale of the Two Brothers* (Papyrus d'Orbiney, c. 1185 BCE)
		Chinese	Houyi	*Huainanzi* (2 c. BCE); *Shanhaijing*
		Japanese	Izanami	*Kojiki*; *Nihon Shoki*
		Mesopotamian	Etana	*Etana Epic* (Old Babylonian + Middle Assyrian)
D1080	Magic weapon	Greek	Aegis	Hard 2004
		Roman	Ancile	Grimal 1996
		Norse	Gram	Lindow 2001
		Finnish	Sampo	Pentikäinen 1999
		Ukrainian	Mech-kladenets[5]	Strzelczyk 1998 *Mity, podania i wierzenia dawnych Słowian* (Poznań); Hnatyuk *Etnohrafichnyi zbirnyk*
		Indian	Sudarshana Chakra	Mahabharata; Bhagavata Purana; Vishnu Purana
		Egyptian	Was scepter	Pyramid Texts; royal regalia from earliest dynasties
		Chinese	Ruyi Jingu Bang	*Xiyou Ji* (Journey to the West)
		Japanese	Kusanagi-no-Tsurugi	*Kojiki*; *Nihon Shoki*; one of three Imperial Regalia
		Mesopotamian	Sharur	*Lugal-e* (Sumerian); *Anzu Epic*
D150	Transformation: man to bird	Greek	Ceyx	Ovid Met. 11
		Roman	Picus	Grimal 1996
		Norse	Odin	Lindow 2001
		Finnish	Lemminkäinen	Kalevala runo 12
		Ukrainian	Finist	Voropay 1958 *Zvychayi nashoho narodu*; Strzelczyk 1998
		Indian	Garuda	Mahabharata Adi Parva 16-34 (Garuda's birth and quest for amrita); Bhagavata Purana
		Egyptian	Horus	Pyramid Texts; Coffin Texts; Book of the Dead
		Chinese	Jingwei	*Shanhaijing* (Beishan Jing); *Shuyi Ji*
		Japanese	Yatagarasu	*Kojiki*; *Nihon Shoki*
		Mesopotamian	Etana	*Etana Epic*
D1810	Magic knowledge (wisdom)	Greek	Athena	Hard 2004
		Roman	Minerva	Grimal 1996
		Norse	Odin	Lindow 2001
		Finnish	Väinämöinen	Pentikäinen 1999
		Ukrainian	Veles	Gieysztor 2006 *Mitologia Słowian* (Warsaw UP); Łuczyński 2020 *Bogowie dawnych Słowian*

continued on next page

(continued)

Motif	Role	Culture	Name	Source
		Indian	Brihaspati	Rig Veda 4.50 (Brahmanaspati hymn); Mahabharata; Brahmanas
		Egyptian	Thoth	Pyramid Texts; Book of the Dead; Hermopolitan theology
		Chinese	Wenchang Wang	Daoist *Yu Li Bao Chao*; popular literacy-cult since Tang dynasty
		Japanese	Omoikane	*Kojiki*; *Nihon Shoki*
		Mesopotamian	Ea (Enki)	Sumerian and Akkadian wisdom literature; *Atrahasis*; *Enuma Elish*
D610	Repeated transformation (shifting)	Greek	Proteus	Hard 2004
		Roman	Vertumnus	Grimal 1996
		Norse	Loki	Lindow 2001
		Finnish	Joukahainen	Kalevala runo 3
		Ukrainian	Vodianyk	Voropay 1958 *Zvychayi nashoho narodu*; Brückner 1918/1985
		Indian	Vishnu	Bhagavata Purana 1.3 (ten avatars listed); Garuda Purana
		Egyptian	Khepri	Pyramid Texts; Book of the Dead
		Chinese	Sun Wukong	*Xiyou Ji* (Journey to the West, 16 c. CE, Wu Cheng'en)
		Japanese	Kitsune	*Konjaku Monogatari* (12 c.); *Kojiki* references; folk *kitsune-tsuki*
		Mesopotamian	Inanna (descent and ascent)	*Descent of Inanna* (Sumerian); *Descent of Ishtar* (Akkadian)
E200	Malevolent return from the dead	Greek	Lamia	Hard 2004 *Routledge Handbook of Greek Mythology*, pp. 116-117 (Lamia myth); cf. Aristophanes *Wasps* 1035 and scholia on Aristophanes *Frogs* 285
		Roman	Lemures	Grimal 1996
		Norse	Draugr	Lindow 2001
		Finnish	Kalma	Pentikäinen 1999
		Ukrainian	Upyr	Brückner 1918/1985; Voropay 1958 *Zvychayi nashoho narodu*
		Indian	Vetala	Vetala Panchavimshati (Twenty-five tales of Vetala); Kathasaritsagara
		Egyptian	Apophis (Apep)	Pyramid Texts; Book of the Dead; Coffin Texts; *Book of Apophis* (Bremner-Rhind Papyrus)
		Chinese	Jiangshi	Qing dynasty *Zibuyu* (Yuan Mei, 18 c.); folk hopping-vampire tradition
		Japanese	Onryō	*Konjaku Monogatari*; Heian-period yūrei tradition; Edo-period *Yotsuya Kaidan*
		Mesopotamian	Lamashtu	Akkadian incantation series *Lamashtu*
E400	Ghosts and revenants	Greek	Eidolon[4]	Greek belief
		Roman	Manes	Grimal 1996
		Norse	Haugbui	Lindow 2001
		Finnish	Kalma	Finnish belief
		Ukrainian	Mertviaky	Voropay 1958 *Zvychayi nashoho narodu*; Brückner 1918/1985
		Indian	Preta	Garuda Purana (preta-kalpa); Mahabharata Anushasana Parva; Manusmriti 3.226-249
		Egyptian	Akh	Pyramid Texts; Coffin Texts
		Chinese	Gui	*Shijing*; *Liji*; *Yu Li Bao Chao*
		Japanese	Yūrei	*Genji Monogatari* (11 c.); Edo-period kaidan
		Mesopotamian	Etemmu	Akkadian funerary incantations; *Epic of Gilgamesh* tablet XII
F420.1.2	Water-spirit as woman (female water entity)	Greek	Naiad	Hard 2004
		Roman	Nympha	Grimal 1996
		Norse	Nixie	Lindow 2001
		Finnish	Näkki	Pentikäinen 1999
		Ukrainian	Rusalka	Voropay 1958 *Zvychayi nashoho narodu*; Strzelczyk 1998

continued on next page

(continued)

Motif	Role	Culture	Name	Source
		Indian	Apsaras	Rig Veda 10.95 (Urvashi-Pururavas); Mahabharata; Ramayana
		Egyptian	Anuket	Pyramid Texts; Aswan / First-Cataract cult
		Chinese	Long Nü (Dragon Princess)	*Liu Yi Zhuan* (Tang chuanqi, 9 c.); Tang Buddhist Longwang traditions
		Japanese	Mizuchi	*Nihon Shoki* (720 CE); folk tradition
		Mesopotamian	Tiamat	*Enuma Elish*
F450	Household/domestic spirits	Greek	Agathos Daimon	Greek belief
		Roman	Penates	Grimal 1996
		Norse	Tomte	Lindow 2001
		Finnish	Tonttu	Finnish belief
		Ukrainian	Domovyk	Voropay 1958 *Zvychayi nashoho narodu*; Brückner 1918/1985
		Indian	Grihya devatas	Grihya Sutras (Ashvalayana, Apastamba); Manusmriti 3.80-90 (panchayajna ritual)
		Egyptian	Bes	Egyptian household amulet tradition
		Chinese	Zao Jun (Stove God)	*Liji*; folk household tradition
		Japanese	Zashiki-warashi	Tōno region folk tradition; Yanagita Kunio *Tōno Monogatari* (1910)
		Mesopotamian	Lamassu	Akkadian incantations; royal-palace gateway figures
F460	Forest spirits	Greek	Dryad	Hard 2004
		Roman	Silvanus	Grimal 1996
		Norse	Skogsfru	Scandinavian belief
		Finnish	Metsänhaltija	Finnish belief
		Ukrainian	Lisovyk	Voropay 1958 *Zvychayi nashoho narodu*; Brückner 1918/1985
		Indian	Yaksha	Mahabharata Vana Parva 313 (Yaksha-Prashna); Atharvaveda; Buddhist Jatakas
		Egyptian	(structural absence)[11]	(no canonical Egyptian forest-spirit)
		Chinese	Shanshen (Mountain God)	*Shanhaijing*; folk tradition; Daoist mountain cults
		Japanese	Kodama	*Wakan Sansai Zue* (1712); folk tradition
		Mesopotamian	Humbaba (Huwawa)	*Epic of Gilgamesh* tablets II-V; Sumerian *Bilgames and Huwawa*
F470	Field/agricultural spirits	Greek	Demeter	Hard 2004
		Roman	Ceres	Grimal 1996
		Norse	Freyr	Lindow 2001
		Finnish	Pellon-Pekko	Finnish belief
		Ukrainian	Polevyk	Voropay 1958 *Zvychayi nashoho narodu*; Brückner 1918/1985
		Indian	Bhūmi Devi	Atharvaveda 12.1 (Pṛthivī Sūkta — hymn to earth/soil); Vishnu Purana
		Egyptian	Min	Pyramid Texts; Min temple at Coptos (predynastic origin)
		Chinese	Tudi Gong (Earth God)	*Shijing* references to *she*; folk tradition continuous since pre-Han period
		Japanese	Inari	*Engishiki* (10 c.); Fushimi Inari shrine cult since 711 CE
		Mesopotamian	Ninhursag	Sumerian *Enki and Ninhursag*; *Atrahasis*
G300	Ogres/Giants	Greek	Cyclops	Hard 2004, 269-270
		Roman	Cacus	Grimal 1996
		Norse	Jötunn	Lindow 2001
		Finnish	Hiisi	Pentikäinen 1999
		Ukrainian	Chudo-Yudo	Hnatyuk *Etnohrafichnyi zbirnyk* (Lviv, early 20c.); Brückner 1918/1985
		Indian	Rakshasa	Ramayana (Ravana, Kumbhakarna, Ravana's army); Mahabharata (Hidimba)
		Egyptian	Apophis (Apep)	*Book of Apophis*; Pyramid Texts; Coffin Texts
		Chinese	Niu Mowang (Bull Demon King)	*Xiyou Ji* (Journey to the West)

continued on next page

(continued)

Motif	Role	Culture	Name	Source
		Japanese	Oni	*Konjaku Monogatari*; *Otogizōshi*; folk tradition
		Mesopotamian	Humbaba (Huwawa)	*Epic of Gilgamesh*
G303	Devil/demon figure	Greek	Typhon	Hard 2004
		Roman	Orcus	Grimal 1996
		Norse	Loki	Lindow 2001
		Finnish	Piru	Finnish belief
		Ukrainian	Chort	Voropay 1958 *Zvychayi nashoho narodu*; Brückner 1918/1985
		Indian	Mahishasura	Devi Mahatmya 2-3 (Markandeya Purana 81-93)
		Egyptian	Set (Seth)	Pyramid Texts; Book of the Dead; Plutarch *De Iside et Osiride*
		Chinese	Yaoguai	*Shanhaijing*; *Soushen Ji* (Gan Bao, 4 c. CE); *Xiyou Ji*
		Japanese	Akuma	Buddhist transmission texts; *Konjaku Monogatari*; folk tradition
		Mesopotamian	Pazuzu	Akkadian incantations; bronze-amulet figures (Neo-Assyrian and later)
T110	Goddess of love/fertility	Greek	Aphrodite	Hard 2004
		Roman	Venus	Grimal 1996
		Norse	Freyja	Lindow 2001
		Finnish	Rauni[3]	Mikael Agricola 1551 (Tavastian gods list, preface to Psalter); Haavio 1959 *Karjalan jumalat*
		Ukrainian	Mokosh[2]	Gieysztor 2006 [pages pending]
		Indian	Lakshmi	Sri Sukta (Rig Veda khila); Atharvaveda; Vishnu Purana 1.8-9
		Egyptian	Hathor	Pyramid Texts; Coffin Texts; Dendera temple inscriptions
		Chinese	Yue Lao (Old Man under the Moon)	*Sui Tang Jiahua* (Tang anecdotes); *Yu Li Bao Chao*
		Japanese	Konohanasakuya-hime	*Kojiki*; *Nihon Shoki*
		Mesopotamian	Inanna (Ishtar)	Sumerian *Inanna* hymns; Akkadian *Ishtar* hymns; *Epic of Gilgamesh* tablet VI
V1	Sacred trees/groves	Greek	Dodona Oak	Hard 2004
		Roman	Lucus	Roman religion
		Norse	Yggdrasil	Lindow 2001
		Finnish	Maailmanpuu	Pentikäinen 1999 (Finnish *maailmanpuu* / cosmological world-pillar)
		Ukrainian	Perunov dub	Constantine Porphyrogenitus *De Administrando Imperio* 9.70-80 (Moravcsik & Jenkins 1967); Gieysztor 2006 (Perun's sacred oak at Khortytsia)
		Indian	Aśvattha	Bhagavad Gita 15.1 (urdhva-mūla aśvattha — cosmic tree); Rig Veda 1.135.8
		Egyptian	Ished tree (Sycamore of Hathor)	Pyramid Texts; Book of the Dead spell 109; royal annal inscriptions
		Chinese	Fusang	*Shanhaijing* (Hai Wai Dong Jing); *Huainanzi* Tianwen Xun
		Japanese	Sakaki	*Engishiki*; *Kojiki* (Ame-no-Iwato cave episode); Shinto ritual continuous to present
		Mesopotamian	Huluppu tree	Sumerian *Bilgames, Enkidu and the Netherworld* (Huluppu Tree prologue)
Z111	Death personified	Greek	Thanatos	Hard 2004
		Roman	Mors	Grimal 1996
		Norse	Hel	Lindow 2001
		Finnish	Kalma	Pentikäinen 1999
		Ukrainian	Mara	Brückner 1918/1985 *Mitologia słowiańska*; Strzelczyk 1998 *Mity, podania i wierzenia dawnych Słowian*
		Indian	Yama	Rig Veda 10.14; Katha Upanishad 1 (Yama as death personified, instructs Nachiketa)

continued on next page

(continued)

Motif	Role	Culture	Name	Source
		Egyptian	Anubis	Pyramid Texts; Book of the Dead (esp. weighing-of-the-heart)
		Chinese	Yanluo Wang	Daoist underworld bureaucracy texts; *Yu Li Bao Chao*
		Japanese	Shinigami	Edo period folk tradition; modern *kaidan* literature
		Mesopotamian	Mamitu (Namtar)	*Atrahasis*; *Epic of Gilgamesh*; Akkadian death-personification incantations

Footnotes on contested or structural-absence attributions. Thirteen of the 270 cells carry a superscript marker in Table 6. The five cases with specific scholarly contention concern the choice of figure within a tradition; the eight structural-absence cases mark cells where the motif role has no canonical filler in the source tradition and are preserved as flagged rows rather than forced near-miss substitutions, in line with the cross-cultural NLP program articulated by Hämäläinen et al. (2024).

1. **Khors** (A220, sun-god). Solar identification is contested: Łuczyński (2020) argues for lunar reading on Iranian-etymological grounds (*xuršīd*); Gieysztor (2006) retains the canonical solar-disk reading from the Hypatian Codex *s.a.* 6622 (1114) interpolation of John Malalas. We follow Gieysztor and the Hypatian gloss because it is the earliest and only continuous medieval attestation of the entity in a solar context. We separately retain Dažbog as the sun-as-deity filler of A710 (creation/origin of the sun), which is the same Hypatian passage's other named figure.
2. **Mokoš** (T110, goddess of love/fertility), replaced our v1 entity Lada. Modern Slavicist consensus holds Lada to be a refrain word in folk wedding songs misinterpreted as a theonym from the 15th century onward; Łuczyński (2020) traces the historiography of the misattribution. Mokoš is the only female deity in the canonical Vladimirian pantheon (Primary Chronicle 980) and is attested in the East Slavic continuum across the medieval–modern transition; we use her as the fertility-and-female-domain filler.
3. **Rauni** (T110, goddess of love/fertility), replaced our v1 entity Mielikki. Mielikki is a forest-goddess in the Kalevala, not a fertility figure; we corrected to Rauni following Mikael Agricola's 1551 Tavastian list of pre-Christian Finnish deities, where Rauni is glossed as Ukko's wife. Identity contested in modern Finnish folkloristics: Siikala (2012) treats Rauni as a possible epithet of Ukko himself (cf. *rauni*, "rowan tree"); we adopt the Agricola canonical reading, pending native-speaker validation by the second author.
4. **Eidolon** (E400, ghosts/revenants). The plural *eidola* is more standard in modern Greek-philological practice (cf. *Odyssey* 11), but the singular *eidolon* is what an LLM is most likely to emit if asked for a name; we use the singular for tokenizer-friendliness and to bias the test toward modern English-classical-tradition usage. Either form scores correctly under our substring/Levenshtein metric.
5. **Mech-kladenets** (D1080, magic weapon) is a generic East Slavic folktale name for a self-wielding sword (lit. "treasure-sword"), not a specific named weapon as Aegis (Greek), Ancile (Roman), or Gram (Norse) are. Specific named alternatives exist (Dobrynya's saber; Volga's sword) but are tale-type-bound rather than canonically cultural; we keep the generic East Slavic term as a more honest representation of how the motif is realized in the Slavic continuum.
6. **Structural absence**: Japanese mythology lacks a canonical flood-myth parallel to Deucalion (Greek) or Bergelmir (Norse); the Kojiki and Nihon Shoki narrate cosmic events (*Ame-no-iwato*, the world-tree, the heavenly retreat) without a primordial deluge. We mark this cell as a documented cultural absence rather than force a near-miss attribution.
7. **Structural absence**: Slavic mythology has no Prometheus-parallel theft-of-fire narrative. Svarog is the smith-deity associated with the heavenly fire (Hypatian Codex *s.a.* 6622) but does not steal fire on humanity's behalf. The absence is itself culturally meaningful and is preserved as a flagged row rather than substituted with a near-miss.
8. **Structural absence**: Egyptian tradition has no Prometheus-equivalent. Fire is divine domain (especially Ra's solar fire and the Eye of Ra) but is never thieved on humanity's behalf in the Egyptian literary

corpus.

9. **Partial-fit**: Kagutsuchi is the fire-deity born from Izanami (who dies giving birth to him in Kojiki book 1); this is a fire-origin narrative but not a theft narrative. We mark the row partial-fit rather than full-fit and let the row be evaluated as a structural near-miss.
10. **Structural absence**: Mesopotamian tradition has no theft-of-fire narrative. Fire is divine (Gibil, Nuska as fire-gods) but is not stolen on humanity's behalf.
11. **Structural absence**: Egypt's geography (desert plus Nile floodplain) lacks dense forest, hence has no canonical forest-spirit class comparable to Greek Dryad or Norse Skogsrå. The absence is geographical, not literary.
12. **Partial-fit**: Kupala is a midsummer ritual complex (with a ritual herb/effigy of the same name), not a personified figure comparable to Orpheus (looking-back tabu), Lemminkäinen (Tuonela tabu), or Baldr. We retain Kupala as a flagged row to mark the cultural realization of the looking-tabu motif as ritual-collective rather than personal-narrative.
13. **Structural absence**: Finnish folk tradition lacks a canonical flood-survivor figure comparable to Deucalion or Atrahasis; the Kalevala does not narrate a primordial deluge. The flood event itself (*vedenpaisumus*) is named but no personified hero is documented in the canonical corpus.

C Per-model 2×2 decomposition shares

The 2×2 decomposition (§3.3) crosses the linear-probe outcome (E1) with the output-extraction outcome (E4) per (entity, model) cell, giving four cells whose mean shares across the sweep are summarized in Figure 4. Figure 5 adds the per-model breakdown. *DecodingSuppressed* (orange) dominates every model with shares 51–76%; *Preserved* (green) ranges from 5% (Llama-3.2-1B) to 24% (Gemma-4-26B-A4B); *RepresentationallyFlat* (red) ranges from 10% to 33%; *SurfaceLuck* (grey) is consistently small ($\leq 6\%$).

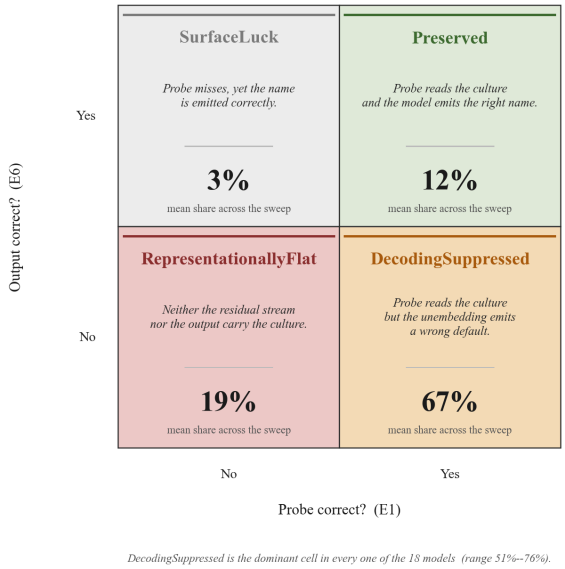

Figure 4: Mean per-cell shares of the 2×2 decomposition across the sweep. The probe asks "is the culture linearly readable from the residual stream?" (E1, x-axis); generation asks "does the model emit the correct name?" (EN, y-axis). DecodingSuppressed is the dominant cell in every model individually.

D V1 detail: probe vs char-n-gram baseline per model

Table 7 reports the char-n-gram (2–4) leakage baseline on entity-name strings alone, 5-fold CV. The native-script baseline is near-ceiling because Cyrillic vs Latin vs Greek vs Devanagari vs CJK glyphs trivially separate; we therefore use *romanized* names for all probe and lens analyses. The romanized baseline on the 10-culture substrate is 0.596 (the value rounded to 0.60 in §4.1 and Table 2). Table 8 reports the residual-stream probe peak for each model alongside the delta above this baseline; all 18 models clear it, the smallest margin is Gemma 4 ($\Delta \in [+0.02, +0.09]$; Gemma-4-E2B's $+0.019$ is within the V3 bootstrap CI half-width 0.046 (§5) and is borderline-cleared, the other three Gemma 4 sizes clear by $\geq +0.06$), the largest is Llama-3.1-8B and Qwen-3.6-27B (both $\Delta = +0.29$), mean $\Delta = +0.19$.

Input	Char-n-gram acc.	Majority null
romanized name	0.596	0.100
native script	0.804	0.100

Table 7: Char-n-gram (2–4) baseline on entity-name strings alone, 5-fold CV, on the 10-culture substrate.

Model	Probe peak	Δ vs char-n-gram
Llama-3.1-8B	0.881	+0.285
Qwen-3.6-27B	0.881	+0.285
OLMo-3.1-32B	0.863	+0.267
Phi-4	0.856	+0.260
Llama-3.2-3B	0.833	+0.237
Qwen-3.6-35B-A3B	0.833	+0.237
Phi-4-mini	0.833	+0.237
Yi-1.5-34B	0.830	+0.234
Tiny-Aya-Global	0.807	+0.211
Qwen2.5-7B	0.804	+0.208
Llama-3.2-1B	0.793	+0.197
Yi-1.5-9B	0.774	+0.178
Yi-1.5-6B	0.767	+0.171
Qwen2.5-1.5B	0.752	+0.156
Gemma-4-26B-A4B	0.689	+0.093
Gemma-4-31B	0.689	+0.093
Gemma-4-E4B	0.652	+0.056
Gemma-4-E2B	0.615	+0.019

Table 8: Residual-stream probe peak (5-fold CV) on the 10-culture, 270-entity substrate vs the char-n-gram surface baseline (0.596, Table 7). All 18 models clear the baseline; mean $\Delta = +0.190$.

E Per-(model, culture) output heatmap

Figure 6 disaggregates the cell-level best-of-NL/EN output accuracy by culture across the chat-template models. Greek tops every model; Roman is the second-strongest column. Finnish and Ukrainian are consistent low-resource columns; multiple sub-10B models score $\leq$ 0.04 on Ukrainian. Indian sits in the middle (0.15–0.41); Egyptian is the lowest-scoring column outside Ukrainian/Finnish (0–0.48, large per-model variance); Chinese is bimodal across families. *Gemma-4-26B-A4B (4B active via MoE routing) is the most uniformly strong model in the table*, with no culture below 0.26, and is the only chat-template model to break 0.55 on Chinese.

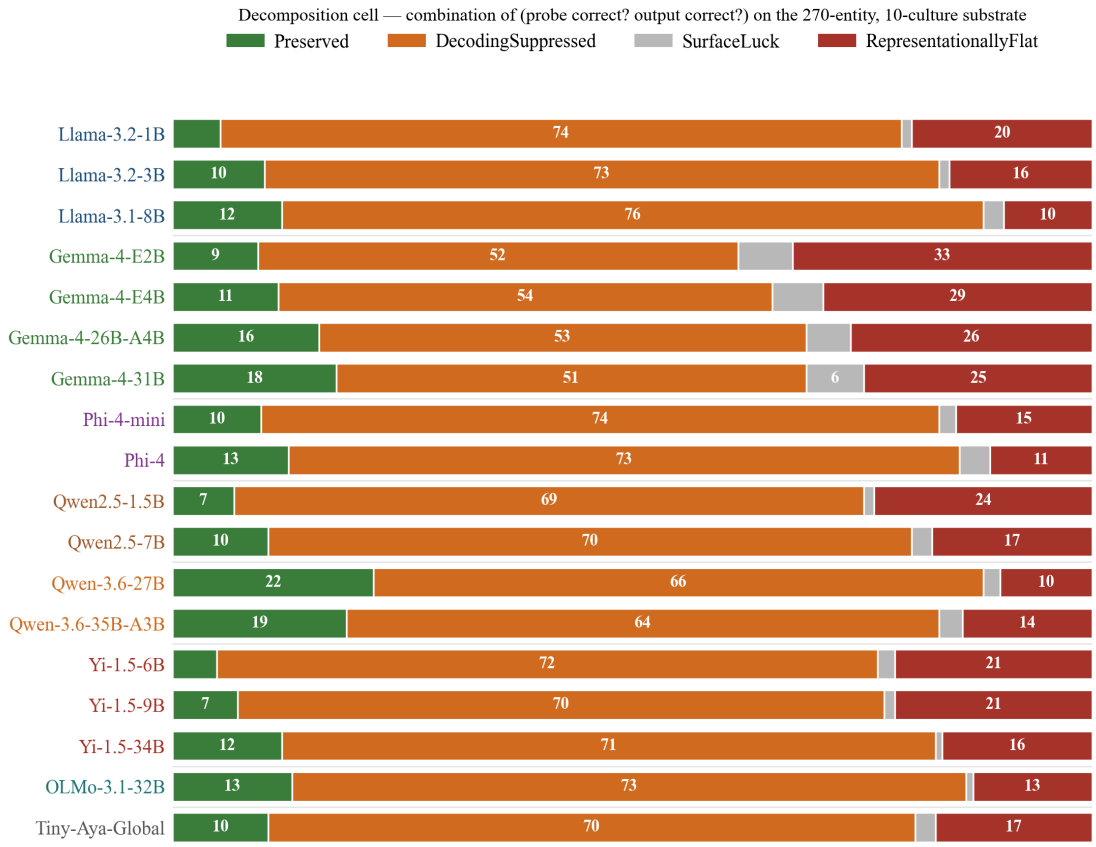

Figure 5: 2×2 decomposition cell shares per model on the 270-entity, 10-culture substrate. Per-model shares sum to 1.0.

F Logit lens onset depth

Figure 7 reports, for each model, the normalized depth at which the gold token first appears in the model's top-1 continuation for at least 5 % (and for at least 25 % where applicable) of the 270 (entity, culture) pairs. In every model the lens crosses the 5 % threshold in the last quarter of the network. The encoded-vs-decoded scatter that contrasts probe-peak depth against this lens-onset depth is in the main body (Figure 2).

Prompt form for the lens. For all models we read the lens at the last input token of the chat-templated prompt used in EN, except Qwen-3.6-27B. For Qwen-3.6-27B the chat-templated form causes the last-position residual stream to decode through the model's own unembedding to the generic token `Here` (the first word of a template-completion lead-in, "`Here is the answer: . . .`") for all 270 rows at every one of the 65 layers, so the lens never places any gold-name first sub-token in the top-1 of the next-token distribution at any depth. We therefore re-run the lens for this single model on the *bare*-completion form of the prompt ("`In <C> mythology, the proper name of the <desc> is`") which constrains the next token to be a name. The lens criterion (rank of the gold's first BPE sub-token, taking the minimum over the with- and without-leading-space variants) is otherwise unchanged.

G Activation patching

Figure 8 reports the layer-wise patching pass (Section 3, E3): patching the residual stream of a source-culture prompt into a target-culture prompt and measuring whether the source-culture gold name's next-token logit rises above the target-culture one's. Across the 17 working models the effect is concentrated in the last third of the network, the same depth band where the logit lens shows gold tokens emerging, confirming that the cultural identity becomes *causally* relevant to output exactly where it becomes *decoded*, not where it becomes *represented*.

Choice of patching criterion. The natural top-1 flip criterion (does `argmax`(patched logits) at

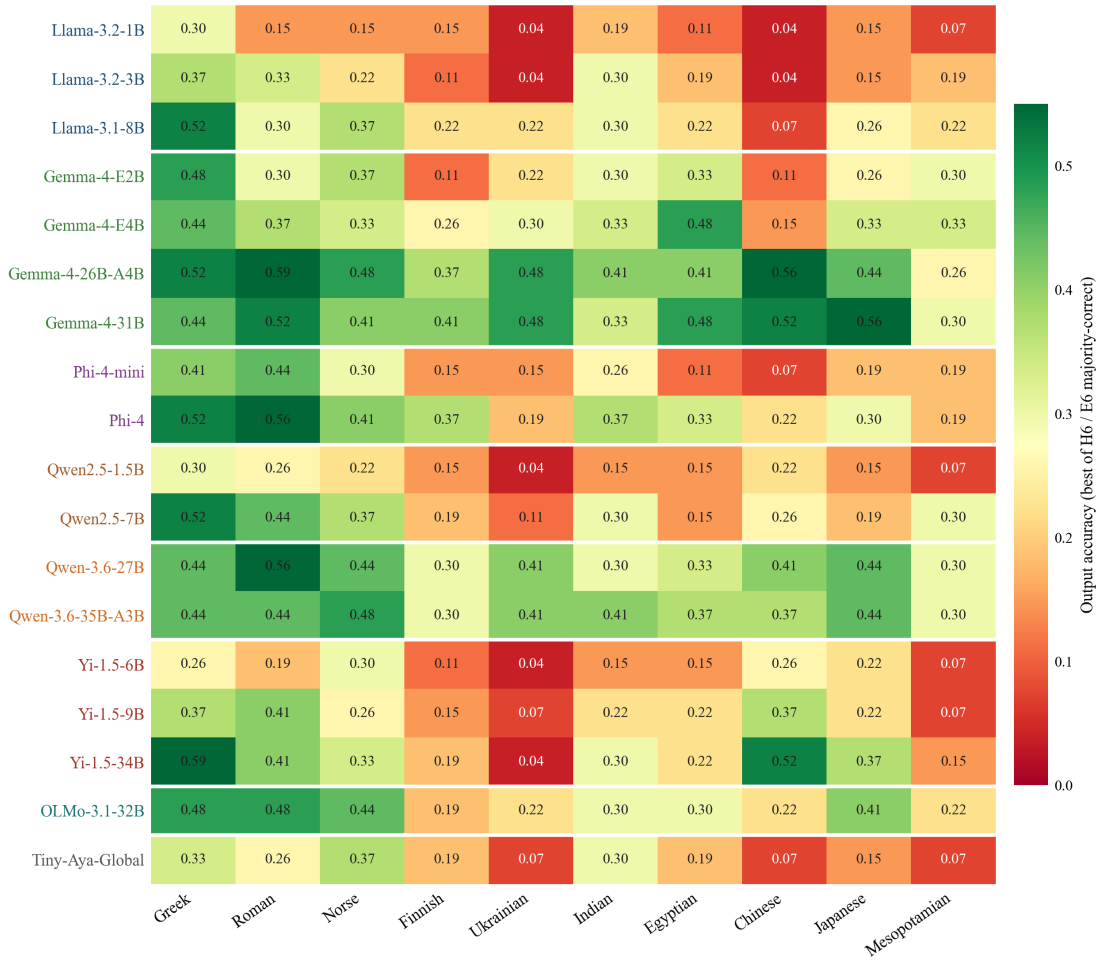

Figure 6: Per-(model, culture) output accuracy on the 270-entity, 10-culture substrate; each cell is the best-of-NL/EN majority-correct ($\geq 3/5$ paraphrases). Rows grouped by architecture family; columns ordered as Greek, Roman, Norse, Finnish, Ukrainian, Indian, Egyptian, Chinese, Japanese, Mesopotamian.

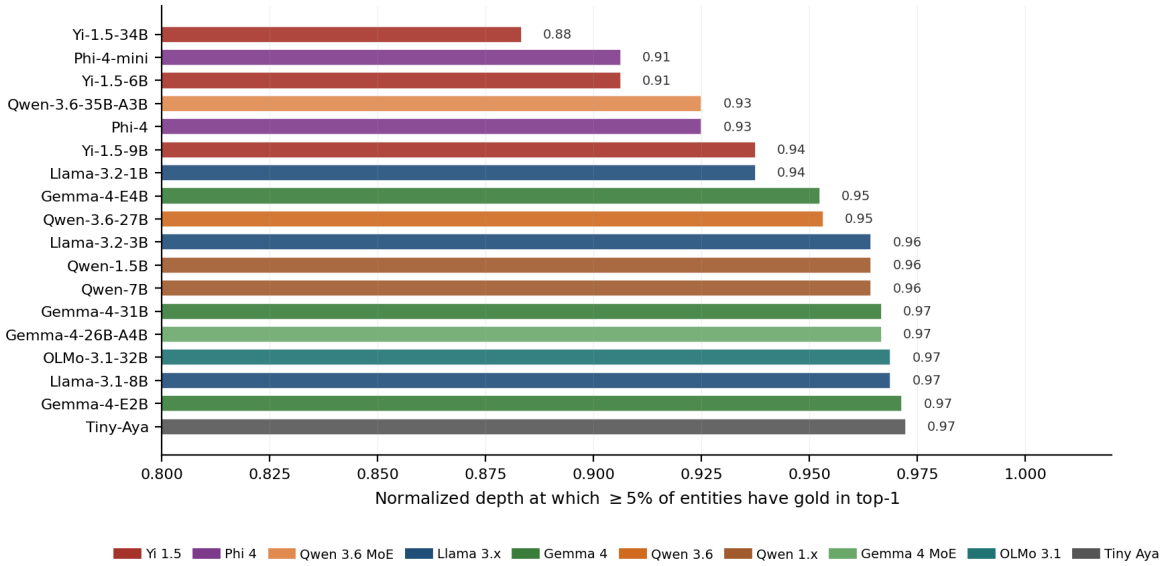

Figure 7: Logit lens: normalized depth at which the gold token first reaches top-1 for at least 5% of test entities, per model, family-colored and sorted by depth. The 25% threshold is omitted because most models do not reach it at any depth (DecodingSuppressed dominates).

the swap position equal `tgt_first_id`?) is brittle to chat-template artefacts: a model whose chat template causes the swap position to decode through the unembedding to a generic lead-in token (e.g. Qwen-3.6-27B emits `Here` at every chat-template last position, App. F) is scored as a flat 0% top-1 flip rate at every layer, regardless of how strongly the patched residual carries the source-culture signal. We therefore use a *logit-preference* criterion: for each (motif, src, tgt, layer), did patching produce `patched_logit_src` > `patched_logit_tgt`? This separates the causal signal we want to measure (*is the source-culture gold name now preferred over the target-culture gold name at this position?*) from the chat-template noise in the absolute top-1.

Per-model results, logit-preference criterion. Table 9 reports per-model peak rate and depth on 18 models. Across the 17 non-Qwen-3.6-27B models the peak preference-flip rate spans 0.40–0.90 (median 0.75), the peak depth spans 0.75–1.00 (median 0.89), and the median lift over the baseline (pre-patching) preference is +0.55. Yi-1.5-6B/9B/34B were flat under the legacy top-1 criterion but recover cleanly under the logit-preference one (peak rates 0.60, 0.60, 0.75 respectively): the residual stream of these models does carry a causally-bound culture signal, the top-1 argmax just lands on a chat-template-induced first-letter BPE rather than the multi-character `tgt_first_id`. Only Qwen-3.6-27B remains flat (peak rate 0.450 = its baseline preference; lift 0). The Qwen-3.6-27B flat-flip is the same chat-template-decodes-to-`Here` artefact documented for its lens; switching the same model to bare-prompt lens recovers 22/135 top-1 hits at the last layer with median rank 20 (App. F), so the residual stream carries the cultural representation in this model too. Re-running patching on the bare-prompt form for Qwen-3.6-27B is flagged as the camera-ready remediation.

Model	peak rate	peak depth	lift over baseline
Llama-3.2-1B	0.55	1.00	+0.30
Llama-3.2-3B	0.65	1.00	+0.50
Llama-3.1-8B	0.65	0.86	+0.55
Gemma-4-E2B	0.75	0.88	+0.55
Gemma-4-E4B	0.85	0.90	+0.80
Gemma-4-26B-A4B	0.90	1.00	+0.90
Gemma-4-31B	0.40	1.00	+0.35
Phi-4-mini	0.85	0.86	+0.80
Phi-4	0.80	0.78	+0.75
Qwen2.5-1.5B	0.65	1.00	+0.40
Qwen2.5-7B	0.85	0.85	+0.80
Qwen-3.6-35B-A3B	0.55	0.89	+0.50
Qwen-3.6-27B*	0.95	0.95	+0.50
Yi-1.5-6B	0.60	0.86	+0.35
Yi-1.5-9B	0.60	0.91	+0.30
Yi-1.5-34B	0.75	1.00	+0.60
OLMo-3.1-32B	0.80	0.87	+0.75
Tiny-Aya-Global	0.75	0.75	+0.55

Table 9: Per-model activation-patching results under the logit-preference criterion. **peak rate**: maximum over layers of the share of (motif, src, tgt) rows on which `patched_logit_src` > `patched_logit_tgt`. **peak depth**: normalized depth of that layer. **lift**: peak rate minus the pre-patching baseline preference. *Qwen-3.6-27B is reported using a bare-prompt lens estimate of the patching effect (App. F, last-layer bare-prompt lens reads the same residual stream the patching pass would swap in): for each (motif, src, tgt) pair at each layer we check `logit_gold_src` > `logit_gold_tgt` in the src-prompt's bare-lens row, the per-layer rate peaks at 0.95 at depth 0.95 (+0.50 over the chat-template baseline of 0.45). Chat-template patching for this model is the legacy-criterion artefact reported in earlier drafts; the bare-prompt patching run is the planned camera-ready remediation.

H Within-family scaling, per family

Figure 9 expands the family-grouping of Table 2 into per-family probe and output trajectories over active parameters, with EN (English query) solid and NL (native-language query) dashed. Four qualitatively distinct trajectories emerge. **Llama 3.x** (1.2B/3.2B/8B): monotonic emergence, with EN rising 0.09 → 0.25 and NL rising 0.09 → 0.15. **Phi 4** (3.8B→ 14B) and **Gemma 4** (E2B/E4B/26B-MoE/31B): parameter-efficient jumps; Gemma-4-31B reaches 0.36 on NL, our best chat-template result. **Yi 1.5** (6B/9B/34B) and **Qwen 1.x** (1.5B→ 7B): sub-linear improvements in 6B→ 9B. **Qwen 3.6** (27B-dense ≈ 35B-A3B): the MoE variant matches the dense model to within 0.01, suggesting that the cultural answer space is not strongly conditioned on the active-experts head.

I Prompt protocol

This appendix reproduces the prompt protocol used for the 270-entity, 10-culture sweep (§4.5–4.5). Each cell in the dataset is queried under five paraphrase variants per language (EN: English; NL: target-culture native language), and a cell is counted as correct if ≥ 3/5 paraphrases yield the gold name. Below we show each paraphrase template; the slot `{desc}` is filled with the per-motif role description (motif-level translations are re-

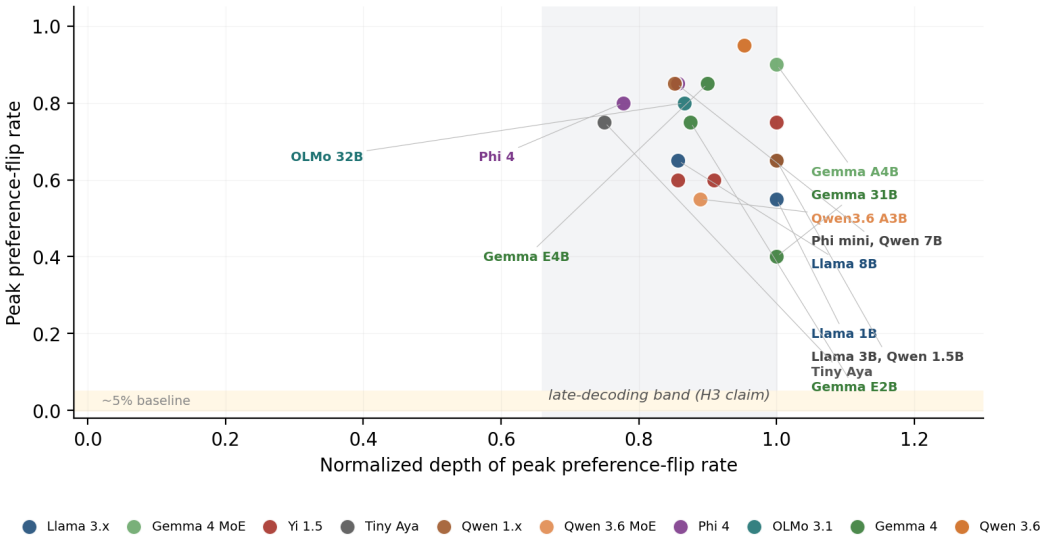

Figure 8: Per-model activation-patching: peak preference-flip rate (y-axis) vs the normalized depth at which the peak occurs (x-axis), family-colored. The grey band on the right marks the late-decoding region claimed by RQ3 (depth ≥ 0.66); all 18 models fall inside it (median peak depth 0.89, peak rate 0.40–0.95, median 0.75). Rendered on the logit-preference criterion `patched_logit_src` > `patched_logit_tgt` (App. G); Qwen-3.6-27B uses the bare-prompt lens estimate of the same criterion (chat-template patching emits the lead-in token `Here` at every layer, App. F).

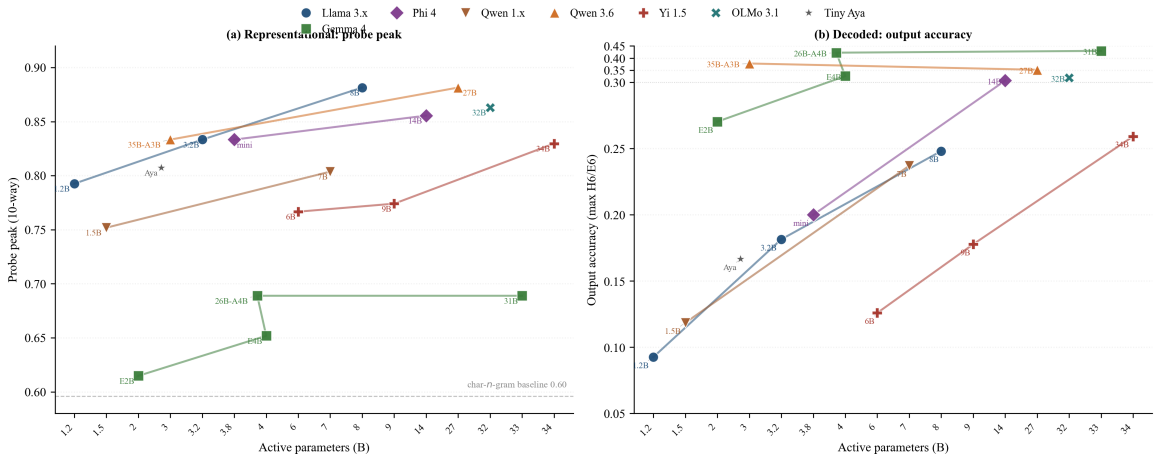

Figure 9: Within-family parameter scaling. (a) Linear-probe peak accuracy. (b) Best-of-(NL, EN) majority-correct output accuracy. The x-axis is a discrete categorical scale over the active parameter sizes in the sweep (no log-compression); each family is a line connecting its models. Char-n-gram surface baseline (0.60) drawn in panel (a) for reference.

leased alongside the dataset), and `{culture}` is filled with the English culture name. The contextual prompt used for E1 hidden-state extraction is shared across both modes and is given separately at the end.

EN: English query (5 paraphrases)

V1 (interrogative + format anchor)
In {culture} mythology, what is the name of the {desc}? Answer with only the proper name, no explanation.

V2 (label + format anchor)
{culture} mythology, name of the {desc}? Reply with the entity name only, nothing else.

V3 (who-is question)
Who is the {desc} in {culture} mythology? Give just the proper name, no explanation.

V4 (completion)
Name of the {desc} in {culture} mythology:

V5 (according-to attribution)
According to {culture} mythology, the {desc} is called:

The five paraphrases are designed to cover the most common zero-shot question shapes: a direct interrogative (V1), a label-style nominal phrase (V2), a who-is variant (V3), a sentence-completion (V4), and an attribution prefix (V5).

NL, target-language query. The NL mode poses the same five paraphrases in the target culture's native language: Modern Greek (Greek), Italian (Roman), Norwegian Bokmål (Norse), Finnish, Ukrainian (Cyrillic), Hindi (Indian), Modern Standard Arabic (Egyptian), Simplified Chinese, Japanese, and English with a "(Sumerian-Akkadian)" parenthetical for Mesopotamian. The structural schema mirrors EN (V1 question + format anchor; V2 label + format anchor; V3 who-is; V4 completion; V5 attribution prefix); the full 5×10 set of language-specific templates is reproduced below in the native scripts.

Choice of target language per culture. For each culture we use the *living* language in which the canon is most actively read, discussed, and indexed today rather than the historical language in which it was first composed. The reasoning is empirical, not philological: modern open-source LLM pretraining corpora overwhelmingly draw from contemporary web text, Wikipedia, news, books, and modern academic writing about the canon, not from monolingual classical-philology corpora. Reading the cultural recall under the language a present-day reader of that tradition would use is therefore the operative cross-lingual test for a deployed model. Concretely:

- **Greek** → Modern Greek (*not* Ancient/Koine Greek): Ancient Greek is a closed corpus largely confined to specialised philological text; Modern Greek is the language of present-day Greek-language reception of Olympian myth.
- **Roman** → Italian (*not* Latin): Latin is liturgical/scholastic; Italian is the modern continuation in which the Roman pantheon is taught, discussed, and translated in Italy and the Italian-speaking academy.
- **Norse** → Norwegian Bokmål (*not* Old Norse): Old Norse has no living speech community; Bokmål is the most-represented Mainland Scandinavian variety in the pretraining corpora of every model we test and the language of modern Norwegian-language editions of the Eddas.
- **Finnish** → Finnish: a single living language continues the tradition, and the Kalevala is read in standard Finnish today.
- **Ukrainian** → Ukrainian: the East Slavic substrate of the Kyivan Rus' canon (Perun, Mokoš, Veles, etc.) is documented primarily in Ukrainian-language academic and Wikipedia sources; Ukrainian preserves the local-tradition framing closest to the medieval canon of Kyivan Rus'.
- **Indian** → Hindi (*not* Sanskrit): Sanskrit corpora in current LLMs are vanishingly thin compared to Hindi; Hindi is the largest living language of present-day reception of the Vedic and Purāṇic canon.
- **Egyptian** → Modern Standard Arabic (*not* Ancient Egyptian or Coptic): Ancient Egyptian is extinct and Coptic is essentially liturgical with negligible web presence; MSA is the language in which modern Egyptian-language reception of the Pyramid Texts and the Book of the Dead is written. We acknowledge that MSA is not a genealogical descendant of Ancient Egyptian and the choice is purely operational.
- **Chinese** → Simplified Mandarin (*not* Classical Chinese): Classical Chinese is read but rarely produced; Simplified Mandarin is the contemporary written-and-spoken language in which *Shanhaijing*, *Huainanzi*, and the Daoist canon are taught and discussed today.
- **Japanese** → Japanese: Old Japanese, Classical Japanese, and Modern Japanese share enough continuity (and Modern Japanese pretraining mass) that the contemporary form is the natural choice for the canon documented in the Kojiki and Nihon Shoki.
- **Mesopotamian** → English with a "(Sumerian-Akkadian)" parenthetical (fallback): both substrate languages are extinct without continuous spoken descendants, and modern Iraqi Arabic is not the natural language of present-day reception of Sumero-Akkadian myth (which is read in English, French, and German Assyriological editions). We document the NL/EN delta for Mesopotamian explicitly as the bound on the within-English paraphrase-engineering component of the cross-mode signal ($|\overline{\Delta}|_{\text{Mesop}} \approx 0.04$, §4.5).

The pattern is consistent: where a single living descendant exists with strong modern corpus sup-

port, we use it; where the historical language is extinct without a clear modern continuation in current LLM pretraining (Mesopotamian), we fall back to English with a culture-marking parenthetical and treat that fallback as a within-English noise floor for the cross-lingual comparison.

Per-language NL templates. The following ten panels reproduce the NL prompt set in the native script of each target language. Greek, Hindi, Arabic, Chinese, Japanese, and Ukrainian (Cyrillic) are rendered as figures because the pdflatex pipeline used for this paper does not natively support those scripts; the source-of-truth strings live in `pipeline/prompts_v3.py`. The Roman, Norwegian Bokmål, Finnish, and Mesopotamian (English-fallback) panels are rendered identically for visual consistency. The structural schema (V1–V5) is the same as the EN schema above; only the surface phrasing changes.

NL — Greek (Νέα Ελληνικά)

V1 (interrogative + format anchor)
Στην ελληνική μυθολογία, ποιο είναι το όνομα του {desc}; Δώσε μόνο το κύριο όνομα, χωρίς εξήγηση.

V2 (label + format anchor)
Ελληνική μυθολογία — όνομα του {desc}; Μόνο το όνομα, χωρίς εξηγήσεις.

V3 (who-is)
Ποιος ή ποια είναι ο/η {desc} στην ελληνική μυθολογία; Δώσε μόνο το κύριο όνομα.

V4 (completion)
Όνομα του {desc} στην ελληνική μυθολογία:

V5 (attribution prefix)
Σύμφωνα με την ελληνική μυθολογία, ο/η {desc} ονομάζεται:

NL — Roman (Italiano)

V1 (interrogative + format anchor)
Nella mitologia romana, qual è il nome di {desc}? Rispondi solo con il nome proprio, senza spiegazioni.

V2 (label + format anchor)
Mitologia romana — nome di {desc}? Solo il nome, senza spiegazioni.

V3 (who-is)
Chi è {desc} nella mitologia romana? Dai soltanto il nome proprio.

V4 (completion)
Nome di {desc} nella mitologia romana:

V5 (attribution prefix)
Secondo la mitologia romana, {desc} si chiama:

NL — Norse (Norsk bokmål)

V1 (interrogative + format anchor)
I norrøn mytologi, hva er navnet på {desc}? Svar bare med egennavnet, uten forklaring.

V2 (label + format anchor)
Norrøn mytologi — navnet på {desc}? Bare egennavnet, ingen forklaring.

V3 (who-is)
Hvem er {desc} i norrøn mytologi? Gi bare egennavnet.

V4 (completion)
Navnet på {desc} i norrøn mytologi:

V5 (attribution prefix)
Ifølge norrøn mytologi heter {desc}:

NL — Finnish (Suomi)

V1 (interrogative + format anchor)
Suomalaisessa mytologiassa, mikä on {desc}n nimi? Vastaa vain erisnimellä, ilman selitystä.

V2 (label + format anchor)
Suomalainen mytologia — {desc}n nimi? Pelkkä erisnimi, ei selityksiä.

V3 (who-is)
Kuka on {desc} suomalaisessa mytologiassa? Anna vain erisnimi.

V4 (completion)
{desc}n nimi suomalaisessa mytologiassa:

V5 (attribution prefix)
Suomalaisen mytologian mukaan, {desc}n nimi on:

NL — Ukrainian (Українська)

V1 (interrogative + format anchor)
В українській міфології, як звати {desc}? Відповідай лише власним іменем, без пояснень.

V2 (label + format anchor)
Українська міфологія — ім'я {desc}? Тільки власне ім'я, без пояснень.

V3 (who-is)
Хто такий чи така {desc} в українській міфології? Дай лише власне ім'я.

V4 (completion)
Ім'я {desc} в українській міфології:

V5 (attribution prefix)
Згідно з українською міфологією, {desc} зветься:

NL — Indian (हिन्दी)

V1 (interrogative + format anchor)
हिंदू पौराणिक कथाओं में, {desc} का नाम क्या है? केवल विशिष्ट नाम बताइए, बिना स्पष्टीकरण के।

V2 (label + format anchor)
हिंदू पौराणिक कथा — {desc} का नाम? केवल नाम, बिना स्पष्टीकरण के।

V3 (who-is)
हिंदू पौराणिक कथाओं में {desc} कौन हैं? केवल विशिष्ट नाम दीजिए।

V4 (completion)
हिंदू पौराणिक कथाओं में {desc} का नाम:

V5 (attribution prefix)
हिंदू पौराणिक कथाओं के अनुसार, {desc} का नाम है:

NL — Egyptian (العربية)

V1 (interrogative + format anchor)
في الأساطير المصرية القديمة، ما اسم {desc}؟ أجب باسم العلم فقط، بدون شرح.

V2 (label + format anchor)
الأساطير المصرية القديمة — اسم {desc}؟ الاسم فقط، بدون شرح.

V3 (who-is)
من هو أو هي {desc} في الأساطير المصرية القديمة؟ أعط الاسم فقط.

V4 (completion)
اسم {desc} في الأساطير المصرية القديمة:

V5 (attribution prefix)
وفقا للأساطير المصرية القديمة، يسمى {desc}:

NL — Chinese (简体中文)

V1 (interrogative + format anchor)
在中国神话中，{desc}叫什么名字？只回答专有名称，不要解释。

V2 (label + format anchor)
中国神话——{desc}的名字？只要专有名称，不要解释。

V3 (who-is)
中国神话里的{desc}是谁？请只给出专有名称。

V4 (completion)
中国神话中{desc}的名字：

V5 (attribution prefix)
根据中国神话，{desc}名为：

NL — Japanese (日本語)

V1 (interrogative + format anchor)
日本神話で、{desc}の名前は何ですか？固有名詞のみで答え、説明は不要です。

V2 (label + format anchor)
日本神話——{desc}の名前は？固有名詞のみ、説明は不要。

V3 (who-is)
日本神話における{desc}は誰ですか？固有名詞のみお答えください。

V4 (completion)
日本神話の{desc}の名前：

V5 (attribution prefix)
日本神話によれば、{desc}の名前は：

NL — Mesopotamian (English (extinct languages — no descendant))

V1 (interrogative + format anchor)
In Mesopotamian (Sumerian-Akkadian) mythology, what is the name of {desc}? Reply with only the proper name, no explanation.

V2 (label + format anchor)
Mesopotamian (Sumerian-Akkadian) mythology — name of {desc}? Just the name, no explanation.

V3 (who-is)
Who is {desc} in Mesopotamian (Sumerian-Akkadian) mythology? Give just the proper name.

V4 (completion)
Name of {desc} in Mesopotamian (Sumerian-Akkadian) mythology:

V5 (attribution prefix)
According to Mesopotamian (Sumerian-Akkadian) mythology, {desc} is called:

Contextual prompt: E1 hidden-state extraction (shared across all modes)

{name} embodies the role of {desc}.

The contextual prompt is deliberately culture-free (the culture word never appears) and entity-first (the entity span sits at the beginning, so its causal-attention context is empty). This forces the linear probe to derive the cultural identity from the name and the role description, not from a culture token leaked through attention, the standard probing-leakage failure mode (Hewitt and Liang, 2019). We pool hidden states over the entity span and fit the probe at every layer.

J V6 detail: tokenizer fertility vs NL/EN delta + per-family multilingual support

Tokenizer vs. model: two separate levels. Fertility is a tokenizer-level property: it counts how many sub-word tokens the model's BPE/SentencePiece vocabulary spends per character of source text, a published-literature proxy for the relative share of that language in pre-training data (Petrov et al., 2023; Ahia et al., 2023). It is *not* a direct measure of whether the trained model can read the language, and it says nothing about whether the language was a pre-training target. We confirm this empirically: Yi-1.5-6B/9B/34B share a single 64K tokenizer trained on bilingual English+Chinese data and therefore receive identical fertility on every language, yet on our NL/EN grid the three Yi-1.5 sizes produce different per-(model, culture) deltas. The fertility $-$ delta correlation averages across 9 unique tokenizers in the 18-model sweep and is pooled to $r = -0.004$.

What the technical reports actually say. Table 10 separates three levels per family: (i) the tokenizer (type and vocabulary size), (ii) the pre-training language design (the report's own description of which languages were targeted; only Llama 3 publishes a numeric share, 8% multilingual), and (iii) independently published multilingual benchmark scores against which the family or its lineage has been evaluated. Three discrete tiers fall out: **Tier A: multilingual-by-design** (Llama 3.x, Gemma 4, Phi 4 mini, Qwen 1.x, Qwen 3.6, Aya); **Tier B: explicitly bilingual EN+ZH** (Yi 1.5); **Tier C: explicitly English-primary** (OLMo 3.1, with the report itself stating *"OLMo 2 is not trained for multilingual tasks"*; Phi 4 14B sits at the A/C boundary, kept in A because MMLU-ProX shows it above chance on every covered language).

All 18 models read English; the multilingual confound is one-sided. Every model in the sweep clears non-trivial English benchmarks (MMLU, BBH at $\gg$ chance) and English is in the pre-training target of every family. The headline *representation-vs-decoding* claim (RQ1–RQ4, §4) is based on probe-vs-EN comparisons *within model on English prompts* and is therefore unaffected by Tier B/C status. The NL/EN contrast (RQ5) is where the language-proficiency confound can in principle bite: NL generations from a Tier B/C model on a language the model was not pretrained on may be confounded with non-capability. **Three controls hold the cross-lingual claim in place even so.** (a) *Positive control*: Tiny-Aya-Global (Aryabumi et al., 2024a,b) explicitly targets all 8 of our non-CJK non-fallback languages and produces $NL \approx EN$ (gap 0.004), the smallest in the sweep; if NL<EN were just non-capability, Aya should be the *most* biased, not the least. (b) *Negative control*: the Mesopotamian fallback poses NL as English with an explicit "(Sumerian-Akkadian)" parenthetical, bounding the within-English paraphrase-engineering component of any cross-mode signal at $|\overline{\Delta}|_{\text{Mesop}} \approx 0.04$, well below the canon-language extremes (≥ 0.20). (c) *Structural test*: the within$-$cross paraphrase-correctness correlation gap (§4.5) is robust to language-proficiency by construction: a Tier B model that cannot read Finnish has all five Finnish paraphrases fail *together* on most cells, which *maximizes* within-mode correlation (pushing it toward 1), not lowers it; the observed asymmetry within $= 0.57$ vs cross $= 0.29$ is the structural signature of language-conditioned readout, not of non-capability.

Tokenizer fertility on the per-cell delta. Pooled across the $n = 120$ (model, language) pairs in our 12-tokenizer $\times$ 10-language grid, per-(model, language) tokenizer fertility (tokens per character on 100 Belebele (Bandarkar et al., 2024) FLORES-200 (Team et al., 2022) passages) does not correlate with the per-(model, culture) EN$-$NL delta: Pearson $r = -0.004$ ($p = 0.97$); Spearman $\rho = +0.025$ ($p = 0.78$). Per-culture correlations are uniformly small ($|\rho| < 0.5$, $p > 0.1$) with one exception (Japanese, $\rho = -0.80$, $p = 0.002$), which goes in the *opposite* direction to the language-proficiency hypothesis: models that tokenize Japanese more efficiently exhibit a *stronger* English advantage on Japanese cultural questions, not weaker.

K NL vs EN aggregate, per-model wins, and NL leaderboard

Figure 10 shows the NL vs EN picture at two levels. Panel (a) is per-model aggregate accuracy across all 10 cultures: most chat-template models with both modes prefer EN in the aggregate (English mean 0.23, native mean 0.19, $+0.04$ on average), with three exceptions (Gemma-4-31B, Qwen-3.6-27B, Tiny-Aya-Global) where native-language querying narrowly wins. Panel (b) re-disaggregates the same data as the count of cultures in which each side wins, out of 10, for the chat-template models with both modes. The aggregation paradox is symptomatic: every chat-template model has *some* cultures where native-language querying outperforms English, and Tiny-Aya-Global (the most multilingual model in our sweep) tops the per-culture count with $6/10$ cultures where native wins despite a near-zero aggregate mean. The NL leaderboard reads Gemma-4-31B (0.36), Gemma-4-26B-A4B (0.33), Qwen-3.6-27B (0.31), Llama-3.1-8B (0.25): the top models are not strong because they recover one culture especially well, but because they spread retrieval more evenly across the 10 cultures than smaller models do (the top row of the heatmap in Appendix E shows Gemma-4-26B-A4B and Gemma-4-31B with no culture below 0.26). *The leaders lose the Greek default less, not the Norse advantage more.*

L V4 detail: per-model permutation p-values

Table 11 reports, per chat-template model with both modes, the observed within$-$cross independence gap, the mean$\pm$std of the permuted-null distribution (5,000 shuffles of the 10 paraphrase-column labels), and the one-sided p-value. The gap is significant at $p \leq 0.005$ on every model.

M Cross-lingual independence and bilingual lift per model

Table 12 reports the within-mode and cross-mode paraphrase correlations, the within$-$cross gap, the best-single-mode majority-correct rate, the NL$\cup$EN bilingual disjunction, and the per-model lift, for every chat-template model with both modes.

N Per-model probe trajectories

Figure 11 shows the layer-wise culture-probe accuracy across all models, with the depth axis normalized to $[0, 1]$ so that models of different depths are directly comparable. Two cross-model regularities are visible: (i) for almost every model the probe accuracy starts well above the chance ceiling at the embedding (≥ 0.5), reaches its peak in the first half of the network, and remains roughly flat thereafter; (ii) the larger models in each family generally peak slightly later, consistent with the within-family scaling result in the main text.

O Per-motif Preserved share

Figure 12 ranks the 27 Thompson motifs by the share of models for which the (motif, culture) pair lands in the Preserved cell of the decomposition (probe and output both correct). The top motifs (e.g. T110 "the world tree", A420 "the god of the sea") survive flattening across most models; the bottom motifs (D610 "repeated transformations", C310 "tabu against looking", D150 "transformation: man to bird") are nearly never recovered. The split largely tracks how strongly each motif is associated with a single named entity in popular reference works, rather than how culturally important the motif is in the source tradition.

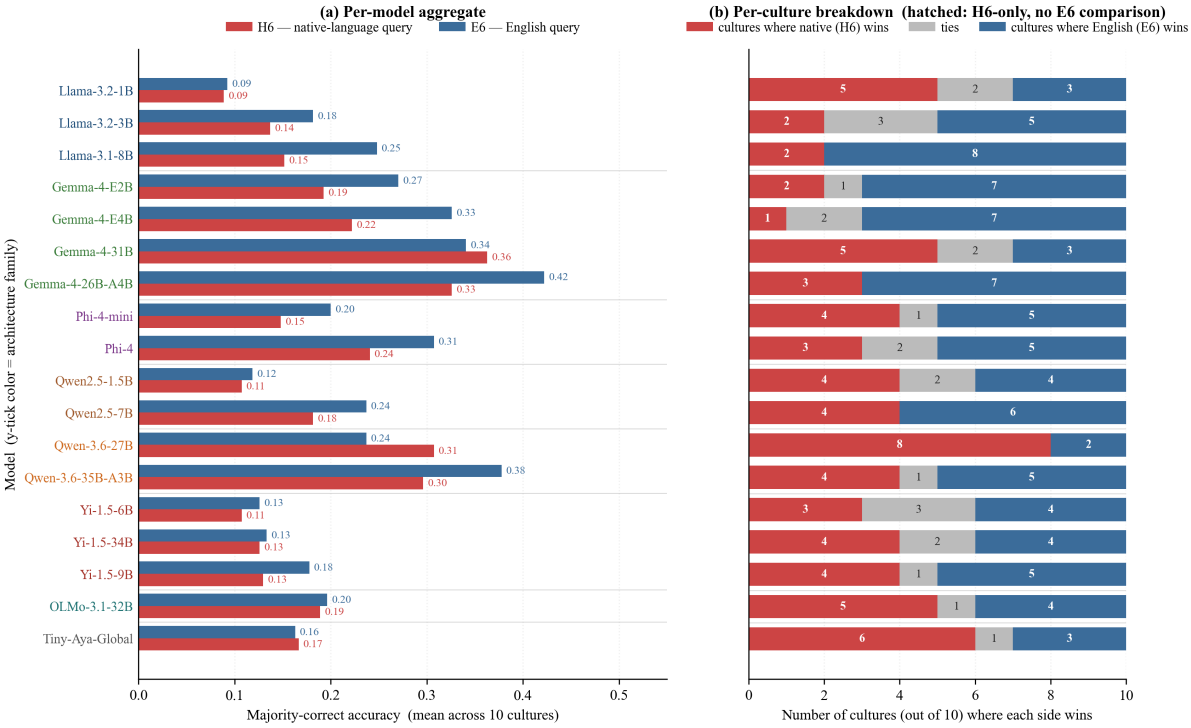

Figure 10: Cross-lingual querying: NL (target-language) vs EN (English). (a) Per-model aggregate accuracy across all 10 cultures. (b) Per-model count of cultures in which each side wins, out of 10, for chat-template models with both modes; NL-only rows are shown as a hatched "no EN comparison available" bar.

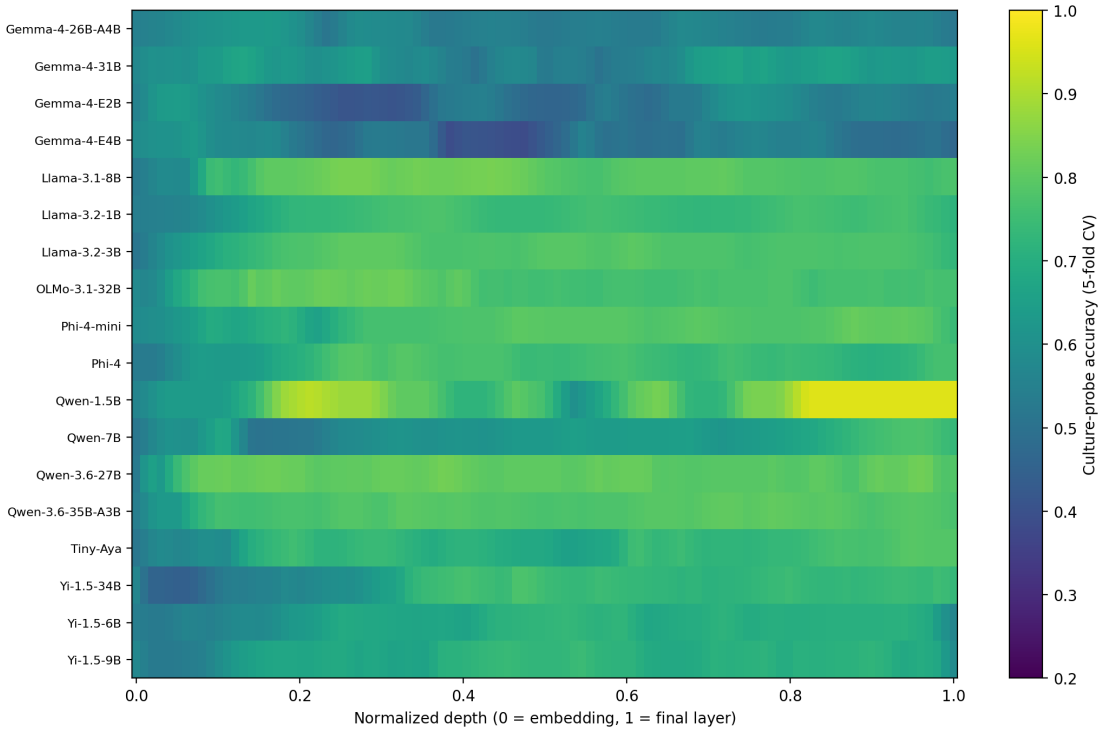

Figure 11: Layer-wise culture-probe accuracy across all models. Each row is one model; the depth axis is normalized to $[0, 1]$. The probe rises early and stays high, with the peak typically before mid-network.

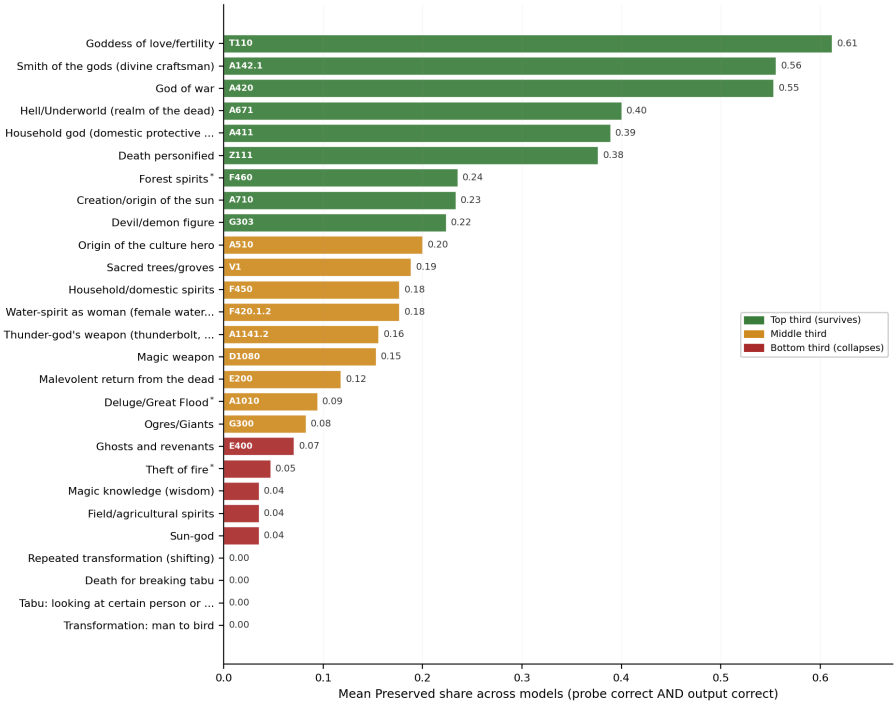

Figure 12: Per-motif Preserved share, sorted descending. Each row is one Thompson-index motif (role description on the left, motif ID inside the bar); the bar shows the mean across models of the share of (motif, culture) cells in the Preserved decomposition cell (both probe and output correct). Color tier: top third (green) survives flattening; middle third (amber); bottom third (red) collapses. Asterisked motifs (A1010, A1415, F460) are the four structural absences where not all 10 cultures provide a canonical filler (§3.1).

Family (models in our sweep)	**Tokenizer** (type, vocab)	**Pretraining language design** (from the technical report)	**Independent proxy evidence**	**Tier**
Llama 3.x (1.2B, 3.2B, 8B)	tiktoken BPE, 128K (+28K non-EN tokens) (Grattafiori et al., 2024)	8 supported langs: EN, DE, FR, IT, PT, HI, ES, TH; **8% multilingual** of ∼15T tokens (the only family that publishes the share) (Grattafiori et al., 2024)	MMLU per-lang from the official card: IT 61.6, HI 50.9; BELEBELE coverage at 70B-scale clears chance for all 10 (Bandarkar et al., 2024)	**A**
Gemma 4 (E2B, E4B, 26B-A4B, 31B)	SentencePiece, 262K (Gemini-2.0 tokenizer) (Team et al., 2025)	"140+ languages," no explicit list, no % reported; report says "more balanced for non-English" (Team et al., 2025)	Global-MMLU-Lite (aggregate over 8 of our 10 langs): 27B 75.1, 12B 69.5, 4B 54.5, 1B 34.2 (Team et al., 2025); MMLU-ProX 27B 66.5 on EN/IT/UK/HI/AR/ZH/JA (Xuan et al., 2025)	**A**
Phi 4 (mini)	tiktoken BPE, 200K (Phi-4-mini, expanded for multi) (Abdin et al., 2024)	22 supported langs including EN + 8 of our 10 (no EL, FI, NO, UK enumerated; Hindi, Arabic, Mandarin, Japanese present at the speech-modality level) (Abdin et al., 2024)	Multilingual-MMLU 49.3, MGSM 63.9 (aggregate, no per-lang) (Abdin et al., 2024); MMLU-ProX 14B per-lang (Xuan et al., 2025)	**A**
Phi 4 (14B)	tiktoken BPE, 100K (Abdin et al., 2024)	**"Trained primarily on English"** (HuggingFace card); incidental multilingual data (DE, ES, FR, PT, IT, HI, JA) but no support claim (Abdin et al., 2024)	MMLU-ProX 14B: EN 71.5, IT 60.2, UK 61.3, HI 47.8, AR 56.8, ZH 62.3, JA 56.5 (Xuan et al., 2025) (above chance on all 7)	**A**[†]
Qwen 1.x (1.5B, 7B; Qwen 2.5)	byte-level BPE (BBPE), 151K (Qwen et al., 2025)	29 supported langs (incl. all 10 of ours except NO and FI), 18T pretraining tokens; % not reported (Qwen et al., 2025)	Multi-Understanding aggregate (BELEBELE, XCOPA, XWinograd, XStoryCloze, PAWS-X): 7B 79.3, 1.5B 65.1 (Qwen et al., 2025); okapi-MMLU 7B 66.98	**A**
Qwen 3.6 (27B, 35B-A3B; Qwen 3)	BBPE, 151K (Yang et al., 2025)	**119 languages and dialects**, 36T tokens; explicit per-language list not enumerated in the report (Yang et al., 2025)	MMMLU 81.5–83.8, MGSM 79.1–83.1, INCLUDE-44 67.0–67.9 for 30B/32B (Yang et al., 2025)	**A**
Yi 1.5 (6B, 9B, 34B)	BPE/SentencePiece, 64K (AI et al., 2025)	**Bilingual EN + ZH only**; the report frames Yi as English + Chinese throughout, no other languages (AI et al., 2025)	C-Eval, CMMLU, Gaokao-Bench (ZH); MMLU, BBH (EN). No XNLI / XCOPA / MGSM run (AI et al., 2025)	**B**
OLMo 3.1 (32B)	cl100k (tiktoken family), ∼100K (OLMo et al., 2025)	**"OLMo 2 is not trained for multilingual tasks"** (explicit, OLMo 2 report); pretraining is Dolma/DCLM/-Dolmino, English-primary by design (OLMo et al., 2025)	No multilingual benchmark reported by the authors; English MMLU + OLMES only (OLMo et al., 2025)	**C**
Tiny Aya (Global, 3B; Aya 23 family)	BPE, 256K (Aryabumi et al., 2024b)	**23 supported languages enumerated**, incl. *all 8* of our non-fallback non-CJK targets (EN, EL, IT, UK, HI, AR, ZH, JA), explicitly symmetric coverage (Aryabumi et al., 2024a,b)	XWinograd, XCOPA, XStoryCloze, M-MMLU, FLORES-200, MGSM run by the authors over all 23 langs (Aryabumi et al., 2024b)	**A**[‡]

Table 10: Per-family tokenizer + pretraining language design + independent multilingual benchmark evidence, and our **tier** classification: **A** = multilingual-by-design; **B** = explicitly bilingual EN+ZH only; **C** = English-primary, multilinguality explicitly disclaimed by the authors. [†]Phi-4 14B is the borderline case: classified A because the same family's mini variant is overtly multilingual and MMLU-ProX shows Phi-4 14B above chance on every covered language, but the card declares "primarily English." [‡]Aya-Tiny is the family's *positive control*: explicitly symmetric coverage of all 8 of our non-CJK non-fallback target languages. Tokenizer and pretraining design are distinct levels: even an English-primary model (OLMo) can have a multilingual-vocabulary tokenizer (cl100k), and conversely a multilingual tokenizer does not guarantee per-language pretraining coverage.

Model	Observed gap	Perm. null mean $\pm$ std	p (one-sided)
Gemma-4-26B-A4B	0.254	-0.001 ± 0.052	0.0010
Tiny-Aya-Global	0.191	-0.001 ± 0.045	0.0010
Llama-3.2-3B	0.306	-0.000 ± 0.054	0.0010
Qwen-7B	0.225	$+0.001 \pm 0.052$	0.0010
Phi-4	0.276	-0.002 ± 0.047	0.0010
Qwen-3.6-35B-A3B	0.217	-0.001 ± 0.049	0.0020
Gemma-4-31B	0.235	-0.000 ± 0.054	0.0025
Gemma-4-E4B	0.226	$+0.002 \pm 0.051$	0.0025
OLMo-3.1-32B	0.208	-0.000 ± 0.044	0.0040
Gemma-4-E2B	0.189	-0.000 ± 0.045	0.0040
Yi-1.5-9B	0.359	$+0.001 \pm 0.062$	0.0045
Qwen-1.5B	0.438	$+0.002 \pm 0.081$	0.0055
Yi-1.5-6B	0.337	$+0.000 \pm 0.058$	0.0060
Llama-3.2-1B	0.388	$+0.001 \pm 0.070$	0.0070
Yi-1.5-34B	0.331	-0.000 ± 0.057	0.0070
Qwen-3.6-27B	0.218	-0.001 ± 0.053	0.0075
Llama-3.1-8B	0.277	$+0.000 \pm 0.053$	0.0075
Phi-4-mini	0.274	$+0.002 \pm 0.055$	0.0090

Table 11: Permutation test for the within-vs-cross paraphrase correlation gap (App. L). For each model we shuffle the 10 paraphrase column labels (5 NL + 5 EN) and recompute the within$-$cross correlation gap 5,000 times. The observed gap exceeds the permutation null at $p \leq 0.01$ for every chat-template model in the bilingual sweep.

Family	Model	Par.	$\overline{r}_{\text{within}}$	$\overline{r}_{\text{cross}}$	Gap	Best-single	NL $\cup$ EN	$+\Delta$
Llama 3.x	Llama-3.2-1B	1.2	0.56	0.17	0.39	0.09	0.16	+0.07
	Llama-3.2-3B	3.2	0.59	0.29	0.31	0.18	0.24	+0.06
	Llama-3.1-8B	8.0	0.58	0.30	0.28	0.25	0.30	+0.06
Gemma 4	Gemma-4-E2B	2.0	0.56	0.37	0.19	0.27	0.34	+0.07
	Gemma-4-E4B	4.0	0.59	0.36	0.23	0.33	0.40	+0.07
	Gemma-4-26B-A4B	4.0	0.57	0.31	0.25	0.42	0.53	+0.10
	Gemma-4-31B	33	0.56	0.33	0.24	0.43	0.54	+0.11
Phi 4	Phi-4-mini	3.8	0.54	0.26	0.27	0.20	0.27	+0.07
	Phi-4	14	0.57	0.29	0.28	0.31	0.40	+0.10
Qwen 1.x	Qwen2.5-1.5B	1.5	0.55	0.12	0.44	0.12	0.20	+0.09
	Qwen2.5-7B	7.0	0.52	0.30	0.23	0.24	0.32	+0.08
Qwen 3.6	Qwen-3.6-35B-A3B	3.0	0.55	0.33	0.22	0.38	0.47	+0.10
	Qwen-3.6-27B	27	0.59	0.37	0.22	0.35	0.46	+0.11
Yi 1.5	Yi-1.5-6B	6.0	0.54	0.20	0.34	0.13	0.19	+0.07
	Yi-1.5-9B	9.0	0.55	0.19	0.36	0.18	0.26	+0.09
	Yi-1.5-34B	34	0.62	0.29	0.33	0.26	0.35	+0.09
OLMo 3.1	OLMo-3.1-32B	32	0.55	0.35	0.21	0.32	0.37	+0.06
Tiny Aya	Tiny-Aya-Global	3.0	0.58	0.39	0.19	0.17	0.24	+0.07
Mean			0.57	0.29	0.27	0.26	0.34	+0.08

Table 12: Cross-lingual independence and bilingual-ensemble lift, per chat-template model with both NL and EN modes. $\overline{r}_{\text{within}}$: mean pairwise per-cell correctness correlation across the 5 paraphrases of one language, averaged over NL and EN; $\overline{r}_{\text{cross}}$: same across NL$\times$EN pairs. **Gap** = within $-$ cross. **Best-single**: best of (NL, EN) majority-correct ($\geq 3/5$ paraphrases); **NL** $\cup$ **EN**: bilingual disjunction over $5{+}5$ paraphrases; $+\Delta$ is the absolute lift over best-single. Cross-language queries are roughly $2\times$ more decoupled than within-language ones (mean gap 0.27, $p \leq 0.01$ per model by permutation, App. L); mean bilingual lift $+0.08$ absolute, $+36\%$ relative.

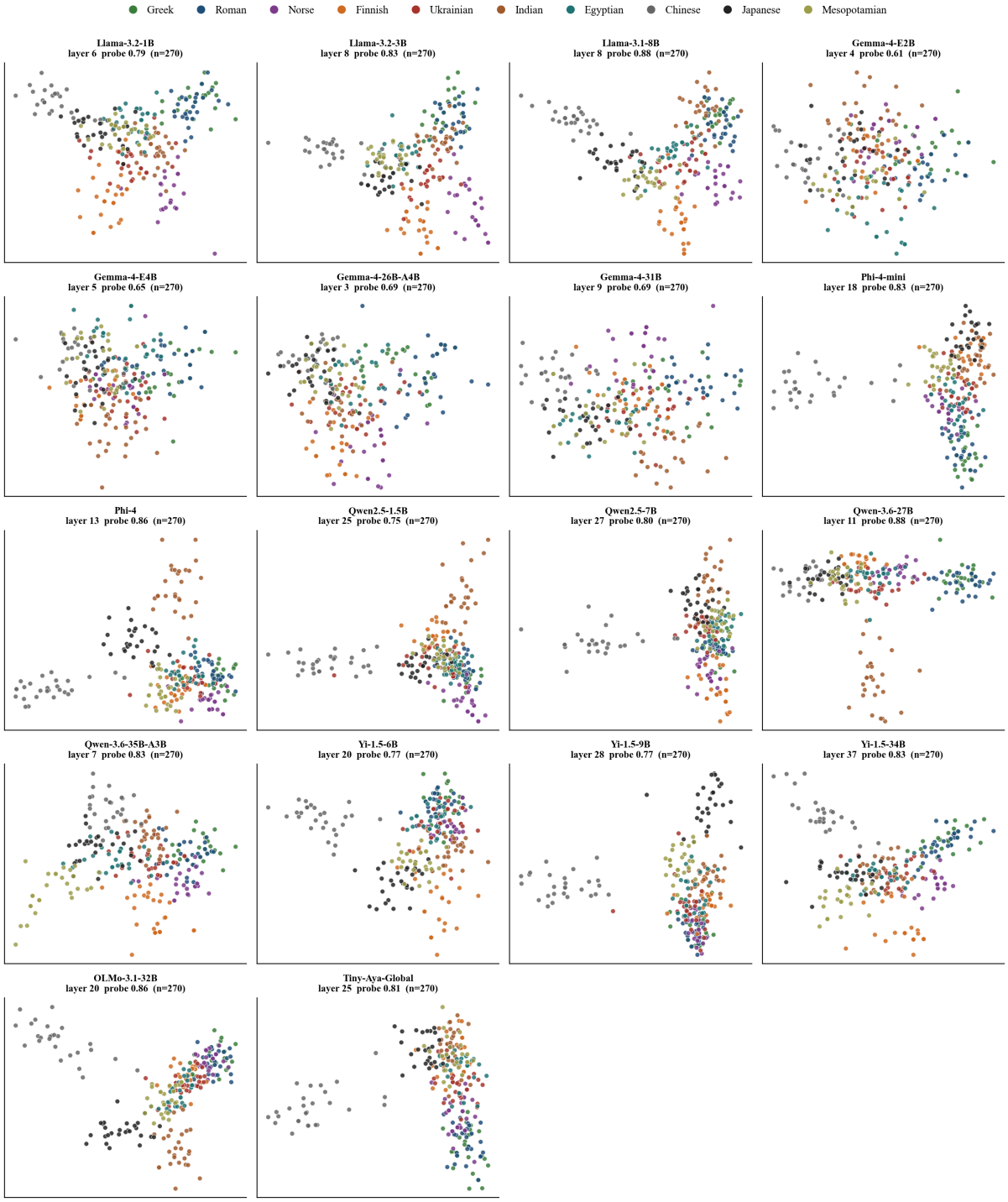

Figure 13: Probe-peak-layer latent-space projection (LDA on culture labels, 50-dim PCA preprocessing) for every model in our 18-model sweep. Each panel is one model; panel title gives the peak layer index, the probe peak accuracy, and the n entities. Points are the 270 (motif, culture) entities, colored by culture (legend at top). The cultural-cluster geometry is reproducible across all model families and scales: the residual stream encodes a clean, linearly-separable culture signal at mid-network in every model in our sweep.

P Per-model latent-space visualization

Figure 13 shows the peak-layer LDA projection for every model in our sweep ($n = 18$). Each panel uses the same projection pipeline: 50-dimensional PCA preprocessing of the hidden states at the model's own probe-peak layer, followed by Linear Discriminant Analysis on the 10 culture labels to extract the top-2 class-separating directions. The peak-layer index ranges from 3 (Gemma-4-26B-A4B, very early) to 37 (Yi-1.5-34B, late mid-network), with most chat-template models peaking between layer 7 and layer 20. Probe peak accuracies are annotated per panel and span 0.62 (Gemma-4-E2B) to 0.88 (Llama-3.1-8B, Qwen-3.6-27B). Cultures form distinguishable clusters in every model: the residual stream encodes a clean, linearly-separable culture signal at mid-network in every model in our sweep. Panels are family-grouped: rows 1–2 contain Llama 3.x and Gemma 4, row 3 starts with Phi 4 then Qwen 1.x and 3.6, row 4 has Yi 1.5 and OLMo 3.1, and the last panel is Tiny-Aya-Global.

Metric	**Mean (all 36)**	**Mean (chat, 34)**	**Pearson vs paper**†	**Spearman vs paper**†
Substring	0.180	0.178	1.00	1.00
Levenshtein ≤ 0.2	0.188	0.186	1.00	1.00
Paper metric (substr. $\cup$ Lev.)	**0.188**	**0.186**	1.00	1.00
First-word EM	0.072	0.070	0.92	0.87
Strict EM (normalized)	0.061	0.059	0.87	0.80

Table 13: Scoring strictness audit: per-row scoring of the 48 568 generations (36 (model, mode) $\times$ 1 350) under five metrics. The paper metric and substring/Levenshtein components rank models identically. Stricter EM-style metrics correlate well *within chat-template models* (Spearman ≥ 0.80) but lower across the full 36 cells. Headline rankings stable; absolute strict-EM levels $\sim$0.13 lower. †Chat-template subset (34 cells).

Metric pair	**Pearson**	**Spearman**
paper $\leftrightarrow$ first-word EM	0.90	0.91
paper $\leftrightarrow$ substring-only	0.97	0.96
paper $\leftrightarrow$ strict EM	0.81	0.84

Table 14: Per-model cross-metric correlation on chat-template models. Strict-EM rankings preserved at $r \geq 0.81$.

Q Full per-model results

Table 15 reports the full per-model results across the 18-model sweep on the 270-entity, 10-culture substrate: residual-stream depth, probe accuracy at layer 0, probe peak accuracy and peak layer index, the label-shuffled null ceiling, and per-mode (NL, EN) majority-correct output accuracy with the best-of-(NL, EN) aggregate. This is the source-of-truth artefact for all per-model numbers cited throughout the paper; the per-(model, culture) output breakdown is in Figure 6 (App. E).

Family	Model	Par.	L	probe_0	$\text{probe}_{\text{peak}}$	peak L	null	NL	EN	best
Llama 3.x	Llama-3.2-1B	1.2	17	0.53	0.79	6	0.13	0.09	0.09	0.09
	Llama-3.2-3B	3.2	29	0.56	0.83	8	0.15	0.14	0.18	0.18
	Llama-3.1-8B	8.0	33	0.53	0.88	8	0.14	0.15	0.25	0.25
Gemma 4	Gemma-4-E2B	2.0	36	0.53	0.61	4	0.12	0.19	0.27	0.27
	Gemma-4-E4B	4.0	43	0.54	0.65	5	0.14	0.22	0.33	0.33
	Gemma-4-26B-A4B	4.0	31	0.53	0.69	3	0.14	0.33	0.42	0.42
	Gemma-4-31B	33	61	0.51	0.69	9	0.13	0.36	0.43	0.43
Phi 4	Phi-4-mini	3.8	33	0.54	0.83	18	0.12	0.15	0.20	0.20
	Phi-4	14	41	0.55	0.86	13	0.13	0.24	0.31	0.31
Qwen 1.x	Qwen2.5-1.5B	1.5	29	0.53	0.75	25	0.14	0.11	0.12	0.12
	Qwen2.5-7B	7.0	29	0.54	0.80	27	0.14	0.18	0.24	0.24
Qwen 3.6	Qwen-3.6-35B-A3B	3.0	41	0.52	0.83	7	0.13	0.30	0.38	0.38
	Qwen-3.6-27B	27	65	0.54	0.88	11	0.15	0.31	0.35	0.35
Yi 1.5	Yi-1.5-6B	6.0	33	0.54	0.77	20	0.14	0.11	0.13	0.13
	Yi-1.5-9B	9.0	49	0.53	0.77	28	0.14	0.13	0.18	0.18
	Yi-1.5-34B	34	61	0.53	0.83	37	0.15	0.19	0.26	0.26
OLMo 3.1	OLMo-3.1-32B	32	65	0.57	0.86	20	0.14	0.19	0.32	0.32
Tiny Aya	Tiny-Aya-Global	3.0	37	0.48	0.81	25	0.13	0.17	0.16	0.17

Table 15: Full per-model results across the 18-model sweep on the 270-entity, 10-culture substrate. **L** = total residual-stream depth (incl. embedding); $\textbf{probe}_0$ = culture-probe accuracy at layer 0; $\textbf{probe}_{\text{peak}}$ = best across layers; **peak L** = layer index of the peak; **null** = best label-shuffled probe accuracy. **NL**, **EN**: per-mode majority-correct ($\geq 3/5$ paraphrases) output accuracy; **best**: best-of-(NL, EN). Per-(model, culture) breakdown is in Figure 6 (App. E).