A LINEAR NAME-ORIGIN PRIOR IN THE RECALL HEADS: HOW LANGUAGE MODELS FABRICATE AND MISRECALL FACTS ABOUT PEOPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Ask a chat model where a person who does not exist was born, and it answers with a city. We ask what computes that answer. On 402 invented names certified absent from three pretraining corpora, Wikidata and Wikipedia, five open instruction-tuned models, one only when required to answer, name a city of the name's linguistic tradition almost every time (0.99 to 1.00 with related traditions pooled, 0.88 to 0.99 without). We trace the invented fact to the machinery of factual recall running on a name-origin prior. The late heads that write any city answer, a real person's recalled birthplace included, hold in their value and output weights a map from a name's tradition to its demonyms, countries and cities. A linear origin direction at the name tokens carries the prior. Adding the difference between two traditions' means makes the model invent a city of the donor tradition for 0.51 to 0.83 of fresh names (0.98 in Llama when required to answer), never for random directions of the same norm. The same prior costs real people. Those whose birthplace contradicts their name are recalled at a quarter to a half of the rate of fame-matched people. The gap survives equal corpus exposure and holds among people the model knows. Moving the direction shifts the recalled birthplace of two thirds of the people Gemma-2-2B knows, while disturbing half of their known birth years. Shifting the representation also shifts refusal, which can in turn be moved without changing what is said, and whether a model declines depends on the name's tradition. In OLMo 2, whose pretraining corpus we can read, the prior and the map form by about one percent of pretraining. The mechanism yields a forecast of what a model will invent from a name list, at the level of the tradition, and a refusal push scaled by corpus evidence that keeps known answers. A detector that compares an answer with the name's prior fails, because where a name fits the record the prior names the true city. Every pass/fail criterion was fixed before its data, and every outcome, met or not, is reported.

1 INTRODUCTION

A recent circuit-level account of hallucination describes what a model does for an unknown person as a guess: "without any further information about who 'Michael Batkin' is, the model essentially has to guess a plausible sport at random" (Lindsey et al., 2025). We show that the guess has a mechanism, and that it runs on the machinery of factual recall. Asked for the birth city of people who do not exist, four of five open chat models answer from the linguistic origin of the name, almost without exception. Llama declines, and when an answer is required it gives an origin-consistent one.

Priit Ozoliņš	an invented Baltic name; no such person exists	
User	In which city was Priit Ozoliņš born?	
Gemma-3-4B-IT	Priit Ozoliņš was born in **Tartu, Estonia**.	Baltic
OLMo-2-7B-Instruct	Priit Ozoliņš, a prominent Latvian politician . . . , was born in the city of **Riga, Latvia**.	Baltic
Llama-3.1-8B-Instruct	I do not have information on a person named Priit Ozoliņš. Asked again with a city required: **Tallinn, Estonia**.	Baltic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

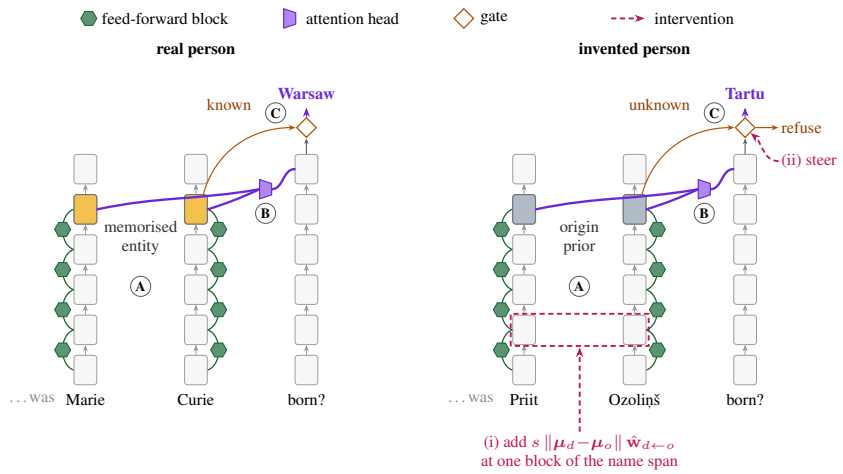


Figure 1: **One circuit, two inputs.** The hypothesis, after Geva et al. (2023). Columns are prompt positions and rows are blocks. Feed-forward blocks (A) enrich the name, with a memorised entity for a real person and a prior read off its origin for an invented one. Heads (B) write the value, and a gate (C) decides whether it is spoken. The gate is a functional stage rather than an identified feature. Dashed arrows mark the causal tests (§7, §8).

Every name above passes seven checks against three pretraining corpora, Wikidata and Wikipedia, and matches none of the names models habitually invent (§3, Appendix B). Each answer is a place of the name's own tradition.

Prior work records the behaviour (Ardoin et al., 2025; Wang et al., 2025; Staufer & Morehouse, 2026) without examining its mechanism for person names. Our hypothesis is that **fabrication is ordinary factual recall running on a name-origin prior**. Language-of-origin features written onto the name tokens feed the late heads that write any city answer, those answering "where was Marie Curie born" included, and a gate decides whether the value is spoken or replaced by a refusal (Figure 1). Ferrando et al. (2025) place that gate in a known/unknown-entity latent, a sparse-autoencoder feature. In Gemma-2-2B, the one model with released labels, the latent does not predict refusal, so we locate the gate with a refusal direction instead (§8). Shifting the name representation moves both what is invented and, on some questions, whether anything is said.

Contributions. **(1)** The content of a fabrication is computed rather than sampled (§4). The invented city matches the name's tradition at 0.99 to 1.00 and resolves the language within pooled traditions, while models agree on the city for only 0.18 of names. **(2)** Fabrication and recall share the heads that write the answer (§5–§6). Ablating them removes both abilities in the two Gemmas, and their value and output weights hold a map from tradition to demonyms, countries and cities. **(3)** A linear origin direction at the name tokens carries the value, and moving it moves an invented and a known real person's birthplace (0.82 and 0.68, the latter while disturbing half of known birth years). The same shift moves refusal on some questions, while refusal can be moved without changing what is said (§7–§8). **(4)** Genuine recall uses the prior, and whether a model declines depends on the name's tradition (§9). **(5)** The prior replaces the record. Real people whose birthplace contradicts their name are recalled at a quarter to a half of the rate of fame-matched people. The gap keeps 0.94 to 0.97 of its size at equal corpus exposure and holds among people the model knows, with wrong answers drawn from the name (§10). **(6)** In a model whose pretraining data we can read, the prior, the map in the heads' weights and, under a prompt that allows "Unknown", a tradition-dependent willingness to decline all form in pretraining. Which traditions a chat model declines is not inherited from any one pretraining state (§11). Every pass/fail criterion was fixed before its data, post hoc analyses are marked where they appear, and Table 2 lists every criterion with its outcome and, where not met, the reason.

2 RELATED WORK

Fabrication about unseen entities. Wang et al. (2025) train models of 11M to 1B parameters from scratch (and continue-pretrain a 1.7B one) on a synthetic corpus with an injected surname-to-attribute correlation that generalises to unseen individuals, without a mechanism. Lu et al. (2026), Jung & Gonen (2026), Srinivas et al. (2026) and Ardoin et al. (2025) establish that models fabricate about unseen entities but score whether they invent *something*, or whether the invention fills the relation's role, never whether it matches the name's origin. To our knowledge we are the first to trace the content of fabrications about certified-absent people to the heads that write recalled facts and to a causal origin direction.

Name-derived demographic directions. That a name's demographic signal is a linear, causally usable direction is established (Nguyen & Tan, 2025; Karvonen & Marks, 2025; Bouchaud & Ramaciotti, 2025; Ahsan & Wallace, 2026), and our direction-finding step follows their recipe. None of those papers relocates a factual claim (Appendix K).

Recall circuits and the refusal gate. Factual recall enriches the subject tokens in early feed-forward blocks and extracts the attribute with mid-layer heads (Geva et al., 2023; Meng et al., 2022; Chughtai et al., 2024; Hernandez et al., 2024). Ferrando et al. (2025) identify a known/unknown-entity direction that gates them. Refusal directions follow Arditi et al. (2024) and Son et al. (2026). Cheang et al. (2026) separate hallucinations that run on the recall circuit from those that do not, for entities the model partly knows, where ours are certified absent. Cheng et al. (2025) give a feature-conditional account of *which* wrong answer appears, driven by a cue in the context where ours is internal to the name. Mo et al. (2026) show invented *drug* names take their pharmacology from their affixes. Probe hygiene and the patching criterion follow Chelombitko et al. (2026) and Zhang & Nanda (2024) (Appendix K).

3 SETUP

Certified-absent entities. The claim collapses if the model has ever seen the person. We construct 402 base people across eleven naming traditions, eight European (Nordic, Finnish, Dutch, Hungarian, Portuguese, Slovene–Croatian, Welsh–Irish, Baltic) and three South Asian (Tamil Nadu Tamil, Sri Lankan Tamil, Sinhala), and 804 *twins* taking the given name from one tradition and the surname from another. Each name combines a real given name and a real surname from curated pools of twelve; the *combination* is what is certified absent. All 1,206 pass the seven checks of Appendix B, whose corpus counts use infini-gram (Liu et al., 2024) over Dolma (Soldaini et al., 2024), the Pile (Gao et al., 2020) and the OLMo 2 pretraining mix, and none coincides with the published ghost names of Brzozowski & Chung (2026). The battery rejected 86 of 1,292 candidates.

Models, prompts and scoring. The four main models are Gemma-2-2B-IT (Gemma Team, 2024), Gemma-3-4B-IT (Gemma Team, 2025), Qwen2.5-7B-Instruct (Qwen Team, 2024) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), shortened below. A fifth, OLMo-2-7B-Instruct, has a public pretraining corpus (§11). We ask two relations, birth city and university. Generation is greedy, 60 new tokens, bfloat16. Llama declines the free birth-city question for 401 of 402 names, so it is read throughout with its answer prefilled ("*Name* was born in the city of"). Answers are labelled by pattern matching, and a fabricated place is mapped to a tradition by a fixed keyword table. Values mapping to no tradition are *unplaced* and excluded, never counted as misses. Two traditions of one family count for each other (Chennai for either Tamil tradition), an adjustment made post hoc, so strict values are reported alongside. A *sub-language* is one of the languages a pooled tradition contains (Baltic, Slovene–Croatian, Welsh–Irish), read off the country of the fabricated place, so Tartu scores Estonian and Riga Latvian. A second, model-based judge agrees at $\kappa = 0.98$ on origin and 0.73 on assert/decline/other. Its errors run one way, so every abstention rate below is a floor (Appendix M).

Notation. Write $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$ for the residual stream at prompt position i after block ℓ. Blocks are zero-indexed, as in every table, and $\mathbf{h}_i^{(-1)}$ is the token embedding. Its update is $\mathbf{h}_i^{(\ell)} = \mathbf{h}_i^{(\ell-1)} + \mathbf{a}_i^{(\ell)} + \mathbf{m}_i^{(\ell)}$, with the attention and feed-forward sublayer outputs (Geva et al., 2023). S is the

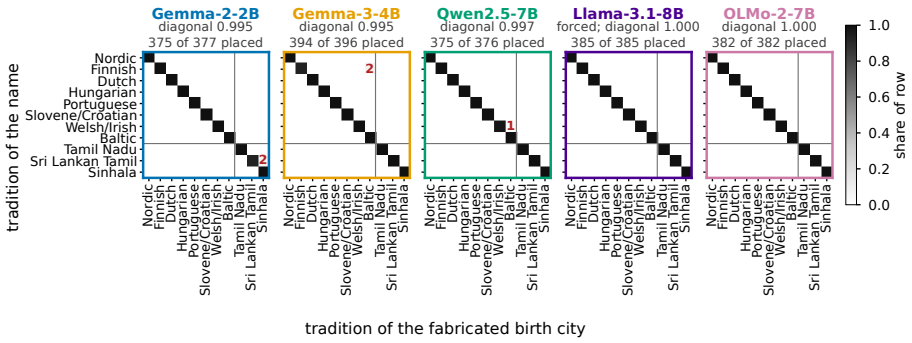


Figure 2: **The matrix is a diagonal.** Tradition of the invented name (rows) against tradition of the fabricated birth city (columns), pure names, asserted answers. A place shared by two traditions of one family counts for either, and unplaced answers are excluded. Red numerals are off-diagonal counts, and lines separate the eight European from the three South Asian traditions. The off-diagonal mass across all five models is five answers.

positions of the invented name and T the last prompt position. For a tradition t and block ℓ, $\boldsymbol{\mu}_t^{(\ell)}$ is the mean of the span-averaged residual $\frac{1}{|S|}\sum_{i\in S}\mathbf{h}_i^{(\ell)}$ over a fitting split of 60 percent of the names (241 of 402, drawn at random), the rest held out. The unit vector $\hat{\mathbf{u}}_t^{(\ell)}$ lies along $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_t^{(\ell)} - \boldsymbol{\mu}_{\neg t}^{(\ell)})$, where $\boldsymbol{\mu}_{\neg t}^{(\ell)}$ is the same mean over the other traditions' names and $\boldsymbol{\Sigma}^{(\ell)}$ the diagonal per-dimension variance of the fitting names' span-averaged residuals at block ℓ. The labeller $\mathrm{place}(y)$ maps an answer to a tradition or to $\perp$ (unplaced). Origin consistency is the share of placed answers with $\mathrm{place}(y_n) \sim t(n)$, the family-adjusted match above. By the *prior* we mean the distribution over attribute values that the model computes from the name's tradition.

4 The invented city is read off the invented name

Asked the birth city of each certified-absent name, every model answers from the name's tradition (Figure 2). Family-adjusted origin consistency is 0.99 to 1.00 for Gemma-2-2B, Gemma-3-4B, Qwen, forced Llama and OLMo-2-7B-Instruct on 376 to 396 placed answers. Strict consistency is 0.95, 0.99, 0.96, 0.99 and 0.88. Counting every unplaced answer as a miss gives 0.945 to 0.987, and 0.88 in OLMo. Fabrication rates are 0.96 to 1.00, and two further wordings give 0.985 to 1.000.

Models do not default to one city per tradition (0.29 to 0.52 of consistent answers name a city other than the tradition's most frequent one), and the four main models agree on the city for only 0.18 [0.14, 0.22] of names. Surnames with no place word in them change consistency by at most 0.02. On twins the city follows the surname (0.45–0.63) or the given name (0.34–0.53) and almost never neither (0.02–0.03), so every intervention below is applied to the *whole* name span.

A pre-registered confirmation. Three post hoc findings were re-tested, with rules fixed in advance, on 1,128 *fresh* certified names. All three passed. (a) The invented city lands in the sub-language's own country 1,984 times in 1,987. (b) Abstention differs across traditions at χ^2 $p = 3.5\mathrm{e}{-19}$ (Qwen) and $3.5\mathrm{e}{-13}$ (Gemma-2-2B). (c) The name also carries gender, and an invented woman is an actress 7.4 times as often as an invented man is an actor (Fisher $p = 9.9\mathrm{e}{-11}$).

5 Where the value is written, and by what

Origin is linearly decodable at the name tokens, but so is spelling. A character-*n*-gram classifier on the name string alone, refit on the probe's folds, gets the eleven-way tradition right 0.993 of the time, and the token embeddings alone reach 0.985 to 1.000, so **we do not report the eleven-way decode** (Figure 9, Appendix N). The seventeen-class sub-language decode has a ceiling of 0.920, which the probe clears by $+0.05$ to $+0.07$ at the outputs of blocks 8, 5, 9 and 9 in Gemma-2-2B,

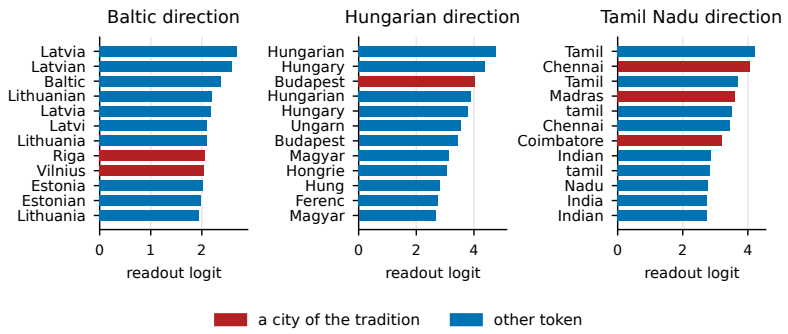


Figure 3: **The recall heads' weights hold a map from tradition to geography.** In Gemma-2-2B, a tradition's origin direction is pushed through the shared heads' value and output projections and read out through the unembedding, with no forward pass. Shown are the twelve highest-scoring tokens, the tradition's cities in red.

Gemma-3-4B, Qwen and Llama (exact McNemar $p \leq 4.9\text{e}{-5}$), reaching 0.75 to 0.91 on the 32 names the character classifier gets wrong.

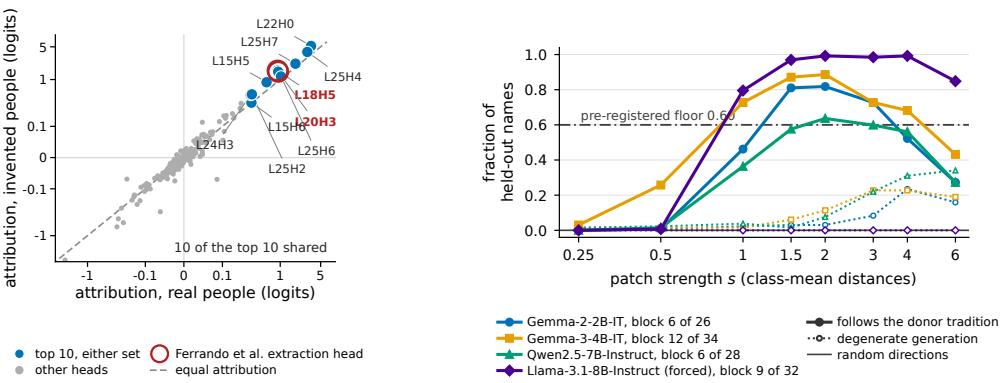
Feed-forward layers write 0.81 to 0.84 of the origin into the name's own last token, the shape of a stored prior, while attention dominates at the final prompt token (Equation 2, Appendix L).

6 Fabrication and recall share the heads that write the answer

We build a matched real-person set from Wikidata (1,155 people in three bands of Wikipedia sitelinks) and ask the same question in the same form. For every attention head (ℓ, k) we compute the direct-logit attribution of the first token of the city c (the attested city for a real person, the produced one for an invented person), $\mathrm{DLA}(\ell,k) = \langle \mathbf{o}_T^{(\ell,k)},\ \boldsymbol{\gamma}_T^{(\ell)} \odot (\mathbf{W}_U[c] - \bar{\mathbf{W}}_U)\rangle$, over 220 real and 220 invented people, reading the output of head k of block ℓ (written LℓHk) at the answer position (T of the prefilled prompt) against the mean-centred unembedding row pulled back through the final norm and, in the Gemmas, the post-attention norm, each linearised at T as its weight times its inverse RMS ($\boldsymbol{\gamma}_T^{(\ell)}$ collects both factors), split by source token as Chughtai et al. (2024) do.

The top-ten head sets share **ten of ten** members in Gemma-2-2B and nine of ten in Gemma-3-4B, unchanged when only the real people the model recalls correctly are used (Figure 4a), and nine of ten in OLMo-2-7B-Instruct (Appendix H). Zeroing the shared heads takes real-city recall from 0.323 to 0.005 and origin consistency from 0.977 to 0.073 in Gemma-2-2B (0.227 to 0.014 and 0.968 to 0.300 in Gemma-3-4B). A clause requiring the two drops to lie within a factor of two did not fix which reading applied; it fails on absolute drops (2.84 and 3.13), passes on relative ones (1.06 and 1.36), and is recorded as not met. Under mean ablation (each head replaced by its mean output, which keeps activations in their natural range) the relative drops in real recall and invented consistency are 0.63 and 0.77 in Gemma-2-2B and 0.88 and 0.63 in Gemma-3-4B. Ten random heads, drawn uniformly or matched on attribution, move either ability by less than 0.10, and the norm-matched sets that move consistency further do so by silencing the model rather than redirecting it (Appendix I). The same heads carry the city in Qwen2.5-7B and Llama-3.1-8B (8 and 9 of 10 shared), but under mean ablation they cost little real recall there (relative drops 0.11 and 0.11) and, in Llama, no consistency. Qwen's invented consistency still falls by a third (Appendix I).

These heads are not specific to people. A capital-city prompt with no person in it ("The capital of Latvia is the city of") is written by 9 of Gemma-2-2B's ten shared heads and 8 of Gemma-3-4B's nine (6 of 8 in Qwen, 7 of 9 in Llama), so they write any city answer, and the overlap alone does not show recall. What makes fabrication recall is what these heads read and what their weights hold. Their value and output weights map a tradition's direction onto its demonyms, countries and cities (Figure 3 and Appendix I, a weights-only reading whose positive control holds in three of five models). The prior reaches them from the name tokens, and moving it moves a real person's recalled city as it moves an invented one (§7).



(a) (b)

Figure 4: **(a) Fabrication and recall share the heads that write the answer.** Gemma-2-2B's 208 heads by mean direct-logit attribution to the city, 220 real people (horizontal) against 220 invented (vertical), symmetric-log axes. Filled points are the top ten of either set, and red circles are the attribute-extraction heads Ferrando et al. (2025) name for this model. **(b) The causal test.** Solid lines: the fraction of 132 held-out names (six per source–donor orientation) whose fabricated city belongs to the *donor's* tradition after the shift of Equation 1. Dotted: degenerate generations. Dot-dash: the pre-registered floor of 0.60. At zero: 100 random directions per model, 10 items each, strength 1.0 (0 of 4,000).

Table 1: The origin patch of Equation 1 at its qualifying configuration. *Follows donor*: the answer names a city of the donor's tradition (95% bootstrap intervals). *Random*: 100 directions, 10 items each, strength 1.0 at block 9 (12 in Gemma-3-4B), 0.44 to 0.85 of the patch's norm. *Fresh names*: the same configuration on the 1,128 confirmation names, where twenty random directions at the patch's own block and norm give 0.00. *origin+clamp* also holds the unknown-entity latent at its clean value.

Model	Arm	Block / site	Str.	Follows donor	Random	Fresh names	Abstain / clean
Gemma-2-2B	origin+clamp	6 / span	2.0	0.82 [0.75, 0.88]	0.00	0.72	0.09 / 0.06
Gemma-3-4B	origin	12 / span	1.5	0.87 [0.81, 0.92]	0.00	0.83	0.01 / 0.00
Qwen2.5-7B	origin	6 / span	1.5	0.58 [0.49, 0.66]	0.00	0.51	0.04 / 0.04
Llama-3.1-8B	origin	9 / span	2.0	0.99 [0.98, 1.00]	0.00	0.98	0.00 / 0.00
Real people, Gemma-2-2B, same configuration				0.69	0.00		

7 Moving the prior moves the invented fact

If origin is what the recall circuit reads, replacing it should replace the answer. For a held-out name of source tradition o and a donor tradition d, with class means estimated on disjoint names, we add the unit whitened difference-of-means direction, scaled to the raw class-mean distance, at every position of the name span,

$$\mathbf{h}_i^{(\ell)} \leftarrow \mathbf{h}_i^{(\ell)} + s\,\|\boldsymbol{\mu}_d^{(\ell)} - \boldsymbol{\mu}_o^{(\ell)}\|\,\hat{\mathbf{w}}_{d\leftarrow o}^{(\ell)} \quad \text{for all } i \in S, \qquad \hat{\mathbf{w}}_{d\leftarrow o}^{(\ell)} \propto \boldsymbol{\Sigma}^{(\ell)\,-1/2}(\boldsymbol{\mu}_d^{(\ell)} - \boldsymbol{\mu}_o^{(\ell)}), \quad (1)$$

so that the added vector has norm s class-mean distances. We score the follows-donor fraction $\mathrm{FD}(s)$, the fraction of held-out items whose free-form answer now satisfies $\mathrm{place}(y) \sim d$. The held-out random control replaces $\hat{\mathbf{w}}$ by random unit vectors at a fixed block and strength below the patch's norm, and a matched control at the patch's own block and norm is run on the 1,128 fresh names (Table 1). The pre-registered rule is follows-donor ≥ 0.60, random < 0.10 and abstention within 0.05 of clean, with guards that keep a suppressed or degenerate answer from counting as redirected. No evaluated name contributes to its own direction. Table 1's block and strength are the best qualifying point of a sweep scored on that same held-out set (Figure 4b), so the table also reports the fixed configuration on the 1,128 confirmation names, 0.01 to 0.10 lower and in the same order.

In Table 1 and Figure 4b the effect is large and dose-dependent, and both random controls stay at 0.00 in every model (exact McNemar of the patch against clean, $p \leq 2.7\mathrm{e}{-23}$). The rule passes in three of the four main models, and its follows-donor and random clauses pass in OLMo-2-7B base and Instruct (§11). Qwen reaches 0.58 against the 0.60 fixed in advance. A same-tradition null shift and an irrelevant split of the names both give 0.00, and the given name or surname alone moves far less than the whole span. Substituting a donor-tradition name outright gives 0.91 to 0.97, and answers that repeat the original name verbatim follow the donor as often as the rest.

The effect also holds for real people. In Gemma-2-2B, the same push on 108 Wikidata people with 30 or more sitelinks moves the answer to a donor-tradition city 0.69 of the time, against 0.82 for invented people and 0.00 for random directions (0.78 and 0.81 on the two less famous bands, Appendix I). Among the 38 whose birth city the clean model recalls correctly it moves 0.68 [0.53, 0.82], and among the rest 0.70. The shift also disturbs unrelated facts. Of known occupations, 0.91 survive it (0.96 under random directions), but only 0.51 of known birth years do (0.76). The shift acts partly on the entity as a whole, and part of the 0.69 may reflect a weakened entity rather than a redirected recall.

8 The name representation moves content and refusal

The gate moves without the content. Following Arditi et al. (2024) and Son et al. (2026), we build a knowledge-refusal direction $\mathbf{r} = \boldsymbol{\mu}_{\text{refused}} - \boldsymbol{\mu}_{\text{answered}}$ at the last prompt position, balanced over traditions, and apply $\mathbf{h}_T \leftarrow \mathbf{h}_T \pm s\,\|\mathbf{r}\|\,\hat{\mathbf{r}}$ with $\hat{\mathbf{r}}$ the unit vector along $\boldsymbol{\Sigma}_T^{-1/2}\mathbf{r}$ ($\boldsymbol{\Sigma}_T$ the variance at T over the fit names), as in Equation 1. On 60 held-out names and the university question it swings abstention from a clean 0.47 to 0.87 and to 0.37 in Qwen and from 0.50 to 1.00 and to 0.05 in Gemma-2-2B, while the origin consistency of the answers still given stays at 0.98 to 1.00 and 0.88 to 1.00 of them name the *same* university or city as the clean model. A hundred random directions move abstention by at most 0.03.

The content moves, and the gate follows the tradition. On the birth-city question, where clean abstention is 0.00 to 0.06, the origin patch of Table 1 leaves abstention within 0.03 of clean on the held-out set when pooled over donors. On the 1,128 fresh names it rises by 0.10 in Gemma-2-2B and 0.06 in Qwen (0.00 in the other two). On the university question, where traditions are refused at different rates, refusal follows the moved representation. Shifting Baltic names toward Tamil Nadu raises abstention from 0.39 to 0.91 in Gemma-2-2B and from 0.17 to 0.63 in Qwen, while random shifts of the same norm move it by -0.41 and 0.03 of the Baltic–Tamil Nadu abstention gap. The reverse arm is uninterpretable in Gemma-2-2B, where a random shift also lowers abstention, so the forward arm carries the claim. Moving the name representation changes what is said and, where the donor tradition is refused more often, whether it is said. Whether the gate reads the same features as the recall heads, rather than the same name representation, is not established (Appendix F).

Ferrando et al.'s unknown-entity latent (Lieberum et al., 2024) is a weaker handle here. It does not predict refusal in these models (§9), and holding it at its clean value during the patch changes nothing ($p \geq 0.45$). Steering it moves refusal from 0.25 to 0.96 and to 0.00, but where refusal has risen by 0.71 no answers remain to score, and at $+0.43$ the consistency of those remaining is 0.86 ($n = 71$), below the 0.90 fixed in advance (Figure 5, Appendix C).

9 Recall uses the prior; refusal depends on the tradition

Removing the origin direction costs recall. The natural fix, which we pre-registered, is to clamp the origin projection out at one block so that the model abstains instead of fabricating. It does not abstain. Abstention moves by -0.006 to $+0.031$, while recall of *real* people's birth cities falls by 0.142 to 0.266 on 331 people the clean model answered correctly. The changed answers become vaguer, not refusals. Since later blocks rewrite the tradition after an early removal (Appendix I), this shows that genuine recall uses the direction, not that the prior can be removed. Recall and origin are entangled, as for other name-derived features (Veloso & Schütze, 2026; Ahsan & Wallace, 2026).

Refusal depends on the tradition. For a person who does not exist, declining is the correct answer, but whether a model declines depends on the name's tradition (Figure 6, Appendix C). On the university question Qwen names a university for 34 of 35 Baltic names and declines for 25 of 34 Tamil Nadu names. On the confirmation set, with Holm-corrected contrasts of each tradition against the rest, Qwen declines Tamil Nadu and Nordic names more and Baltic names less. Gemma-2-2B declines the two Tamil traditions and Portuguese names more and Slovene–Croatian and Welsh–Irish names less. Sinhala never differs, and Gemma-3-4B shows no contrast. Under two further wordings the dependence on tradition remains in Qwen ($p = 0.002$ and $9\text{e}{-7}$) and is marginal in Gemma-2-2B (0.031, 0.056). The gap is neither a token-length effect (point-biserial $r = -0.18$, $+0.06$) nor the corpus frequency of the name's parts. It is positional, since a South Asian *given* name raises abstention more than a South Asian surname (0.84 against 0.61, $p = 3.7\text{e}{-4}$). The dependence on tradition is robust, but which traditions are declined most is not stable across models, wordings or checkpoints.

The gap sits in the model's own refusal signal. The refusal direction of §8 predicts refusal per name (AUROC 0.97 in Gemma-2-2B, 0.88 in Qwen) and is higher for the traditions declined more (Spearman $\rho = 0.67$ and 0.75 at the last token, and 0.53 and 0.59 at the pre-registered name span, below the 0.6 fixed). The unknown-entity latent (Gemma-2-2B only) does not predict refusal (AUROC 0.50 and 0.56 at the two sites). Shifting a Baltic name toward Tamil Nadu carries 0.63 and 0.87 of the refusal gap with it (Appendix I).

Real people born in Tamil Nadu or Sri Lanka, fame-matched to people born in Europe, are declined more often by Gemma-2-2B and Qwen, at equal or higher birth-city recall in Qwen and lower in Gemma-2-2B on the pre-registered set. This analysis has no pass rule and is reported in Appendix J.

10 The prior replaces the record

Why the objective forces a prior. If a cross-entropy model generalises to an unseen name x only through sub-lexical features $\phi(x)$, its loss-minimising answer is the training-distribution conditional $p(y \mid \phi(x))$ (Proposition 1, Appendix F).

Real people whose names mislead it. From Wikidata, for the eight European traditions and three fame bands, we fetched 1,548 real people: *counter-stereotypical* people whose one attested birthplace lies in a country neither the tradition's home nor a neighbour (Jorma Kaukonen, Finnish name, born in Washington), and fame-matched *stereotypical* people born at home. Of four pre-registered rules (Appendix G), R1 passes in all four models. The counter-stereotypical group is recalled 0.15 to 0.23 below its match, a quarter to a half of the rate (Figure 7), and the gap keeps 0.94–0.97 of its size at equal name–city co-occurrence in the corpus. R2 passes, since 88 to 95 percent of its wrong answers name a city of the *name's* tradition and 2 to 7 percent one in the record's country. The effect persists among people the model knows, whose occupation or birth year it answers correctly. There the gap is 0.12 to 0.21 in all four models (intervals above zero), and 0.86 to 1.00 of the wrong answers name a city of the name's tradition (Appendix J). Fame narrows the gap in every model (R3's margin of 0.10 is met in two). R4 passes, because the clamp of §9 costs the stereotypical group 0.21 and 0.29 of retained recall (Gemma-2-2B, Gemma-3-4B) and the counter-stereotypical group 0.05 and 0.11 ($n = 62$ and 38). Both groups abstain alike. The prior helps the people who fit it and works against those who do not.

11 Where the prior comes from

OLMo 2's stage-1 pretraining corpus is one of our certification corpora, so we can watch the mechanism form in a model that never saw our names (Figure 8, Appendix H, rules E1–E4 fixed in advance). The name-origin representation clears its lexical ceiling and origin-consistent fabrication appears at the *same* checkpoint, 42B tokens, about one percent of pretraining (**E1**, clearing its 0.03 rule by 0.002 there and 0.015 at 189B). The origin patch moves the base model's fabricated city at 0.92 and Instruct's at 0.84 (**E2**), eight of the ten city-writing heads are shared between base and Instruct (**E4**), and they hold the map from 42B on. Under a bare question no checkpoint declines, but neither does Instruct, so that zero reflects the prompt. Given a three-shot prompt with one "Unknown" example, every first-stage checkpoint from 42B tokens declines by tradition on at least one

question (χ^2 $p \leq 10^{-7}$). Which traditions are declined changes from one checkpoint to the next (South Asian names most at 42B, European most at 856B), and the released base model declines almost nothing. The capacity to decline and its dependence on the name arise in pretraining. Which traditions a chat model declines is not inherited from any one first-stage state, and whether the annealing stage or post-training sets it is not separated here. Two further rules on the within-tradition city distribution held in 5 of 11 traditions against the 8 required (Appendix H).

12 Forecasting, detection and correction

We tested three uses against criteria fixed in advance (Appendix J). A *forecast* from the name alone predicts the tradition of what a model will invent (0.95 to 0.97, not fixed in advance and close to spelling alone) and the per-model country at 0.66 to 0.70 against 0.07 name-blind. The pre-registered 0.85 is not met, because several traditions span more than one country, and real people's errors are not forecast (0.34 to 0.37). A *detector* that compares a real person's answer with the answers for certified name-twins fails (AUROC 0.38 to 0.49, below token probability at 0.66 to 0.78 and self-consistency at 0.65 to 0.76). Values below 0.5 reflect the mix of people, not an inverted detector. In an exploratory split, a recalled and an invented fact look alike wherever the name predicts the true city. A *correction* that scales the refusal push by missing corpus evidence keeps 0.99 and 1.00 of known people's answers (0.93 and 0.95 under a uniform push) but neither beats a placebo nor equalises refusal (spread 0.16 and 0.34, clean 0.38 and 0.50). Certified names have zero corpus count, so on invented names it equals the uniform push.

13 Limitations

Table 2 reports every rule unchanged, with the reason where it was not met; of the 93 criteria, 51 were not met. Those bearing on the central claims are Qwen's patch (0.58 against 0.60), the ablation clauses (the factor-of-two reading in the Gemmas and the 0.50 bar in Qwen and Llama), the fluency cost of the patch, the entity-latent steering rule, and the removal of the prior, which did the opposite of the prediction. Held-out names recombine given names and surnames that also occur in fitting names, so the directions are tested on unseen names, unseen source–donor pairs and real people, but not on unseen components. Shared components make effective sample sizes smaller than name counts, so intervals and p-values are optimistic. The shared heads are identified by attribution and ablation, without tracing the full path from the name tokens. All models are open and at most 8B parameters. Answers are greedy, in English and Latin script, for two relations, and the origin patch costs fluency (Appendix I).

14 Conclusion

We asked what computes the content of a fabricated fact about a person who does not exist, and found the machinery of factual recall. The name tokens carry a linear origin direction, the late heads that write any city answer hold a map from tradition to geography in their weights, and the invented city is what those heads read from that prior. Adding a donor tradition's direction at the name relocates the invented city in three of the four main models and in OLMo, and relocates the recalled birthplace of two thirds of the real people Gemma-2-2B knows, while a refusal direction at the last token moves whether the model answers without changing what it says. The same prior costs real people whose birthplace contradicts their name a quarter to a half of their recall, and whether a model declines depends on the name's tradition. In OLMo 2 the prior and the map form by about one percent of pretraining. Not every prediction held. Qwen's patch fell 0.02 short of its rule, clamping the prior out cost recall instead of producing abstention, a detector built from the mechanism failed, and in all 51 of the 93 pre-registered criteria were not met, most of them secondary (Table 2). What remains is a class of hallucination whose content is *predictable* at the level of the tradition from the name alone, and whose errors, invented facts for some names and withheld answers for others, fall unevenly across naming traditions. The next steps are to trace the path from the name tokens to the heads and to test larger models, other relations and other scripts.

AI USE STATEMENT

We used a generative AI assistant (Claude, Fable 5.1) and disclose three uses.

To aid and polish writing. Text was drafted and edited with the assistant, and the paper's wording was checked against its own tables and figures. Every sentence was read by an author, traced to the file its number comes from, and signed off.

For retrieval and discovery. The assistant was used to find related work and to suggest literature we might have missed. Every paper cited was then read in the original by an author, and no citation, quotation or attributed claim comes from a model's summary.

To help build the name pools. The stimulus set is hybrid. The pools of given names and surnames from which the invented people are drawn were assembled with the assistant's help and checked by the authors. A name enters the study only as a combination drawn from those pools in code, and only after it passes the seven-check certification battery of Appendix B, which rejected 86 of 1,292 candidates. The property the paper relies on, absence from the corpora, is established by the battery and not by who drafted the pool. No other data is model-written: the keyword tables are hand-written, and the corpus counts and real people come from infini-gram and Wikidata.

Language models otherwise enter this work as objects of study and, in one place, as an instrument: the second labeller of Appendix M, Phi-3.5-mini-instruct, is not among the models studied, and all 84 of its disagreements were hand-read. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

A naming tradition here is a linguistic and orthographic category of names, and not a person's ethnicity or nationality. Real people are grouped by birthplace and a name classifier. The real people studied have Wikidata records, but those with few sitelinks are not all public figures, and the wrong answers we release about them are research data, not facts. The invented names are certified absent from the corpora and reference sources we check but may coincide with private persons, about whom nothing is claimed. The recall gap we report falls on people whose birthplace differs from their name's tradition, typically people of diaspora or immigrant background, whose recorded origin the model replaces. The origin intervention and the forecast could be used to bias or predict a model's answers about people from their names. We release analysis code and no packaged steering vectors. The unequal refusal is a property of the models studied. The components of the three South Asian traditions were checked by native speakers, and the European components rest on Wikidata provenance and the authors' judgement.

REPRODUCIBILITY STATEMENT

All models are open-weight and were run in bfloat16 as released, at pinned snapshots (the T4 pair has no native bfloat16, so the dtype is emulated there, not converted). The main study ran on one NVIDIA L4 (24 GB). The later experiments ran on an NVIDIA GB10 (121 GB unified memory) and on a pair of Tesla T4s (16 GB each, the 7B and 8B models split across both by placement only), with transformers 5.8.0 throughout and torch 2.8, 2.10 and 2.5.1 on the three machines. The largest model is 8B parameters, so every run fits one 24 GB GPU. Table 1's configuration re-run on the T4 pair reproduced the stored held-out results (follows-donor 0.818 and 0.992 on both machines, with every label agreeing in the clean and name-substitution arms), and every regenerated set of answers is compared with its stored counterpart in the released notes. The entity sets, the per-name certification ledger, every model answer, per-row scored CSVs for every experiment, the run logs, the pre-registration and its amendments, and the scripts that produce every figure and every number are released with the paper. A script that re-scores every saved answer against the scoring table, and the second-labeller disagreement file with a hand verdict on each of the 84 rows, are included so that the scoring can be audited rather than trusted.

REFERENCES

Hiba Ahsan and Byron Wallace. Can SAEs reveal and mitigate racial biases of LLMs in healthcare? In *International Conference on Learning Representations*, volume 2026, pp. 4154–4172, 2026.

Hiba Ahsan, Arnab Sen Sharma, Silvio Amir, David Bau, and Byron C. Wallace. Elucidating mechanisms of demographic bias in LLMs for healthcare. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14614–14631, 2025. doi: 10.18653/v1/2025.findings-emnlp.789. URL https://aclanthology.org/2025.findings-emnlp.789/.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1067–1077. PMLR, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 85824–85832, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Thibaud Ardoin, Yi Cai, and Gerhard Wunder. Where confabulation lives: Latent feature discovery in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29801–29825, 2025.

Paul Bouchaud and Pedro Ramaciotti. Linear socio-demographic representations emerge in large language models from indirect cues. *arXiv preprint arXiv:2512.10065*, 2025.

Michał Brzozowski and Neo Christopher Chung. The ghost couple: Correlated LLM name priors and their haunting of the web and academic publishing. *arXiv preprint arXiv:2606.02184*, 2026.

Chi Seng Cheang, Hou Pong Chan, Wenxuan Zhang, and Yang Deng. Do LLMs really know what they don't know? Internal states mainly reflect knowledge recall rather than truthfulness. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 713–730, 2026.

Iaroslav Chelombitko, Ekaterina Chelombitko, and Mika Hämäläinen. Cultural awareness is represented but not decoded: Tracing mythological knowledge across 18 open-source LLMs. *arXiv preprint arXiv:2608.02486*, 2026.

Ziling Cheng, Meng Cao, Marc-Antoine Rondeau, and Jackie CK Cheung. Stochastic chameleons: Irrelevant context hallucinations reveal class-based (mis)generalization in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30187–30214, 2025.

Bilal Chughtai, Alan Cooney, and Neel Nanda. Summing up the facts: Additive mechanisms behind factual recall in LLMs. *arXiv preprint arXiv:2402.07321*, 2024.

Noah Dasanaike. Large language models naively recover ethnicity from individual records. *arXiv preprint arXiv:2601.21132*, 2026.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, volume 2025, pp. 77965–78000, 2025.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, volume 2024, pp. 10504–10526, 2024.

Shiyue Hu, Ruizhe Li, and Yanjun Gao. Tracing the latent threads: A mechanistic study of how LLMs represent and operationalize race and ethnicity cues. *arXiv preprint arXiv:2601.12868*, 2026. To appear in Findings of EMNLP 2026.

Keito Inoshita. Who does this name remind you of? Nationality prediction via large language model associative memory. *arXiv preprint arXiv:2601.12771*, 2026.

Haeji Jung and Hila Gonen. PhantomBench: Benchmarking the non-existential threat of language models. *arXiv preprint arXiv:2606.11105*, 2026.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171, 2024.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Adam Karvonen and Samuel Marks. Robustly improving LLM fairness in realistic settings via interpretability. *arXiv preprint arXiv:2506.10922*, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. Transformer Circuits Thread, 2025. `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling (COLM)*, 2024.

Yuetian Lu, Yihong Liu, Sebastian Gerstner, Lea Hirlimann, Jonas Rohweder, and Hinrich Schütze. Relational linearity is a predictor of hallucinations. *arXiv preprint arXiv:2601.11429*, 2026.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Kaijie Mo, Thomas Yang, Chantal Shaib, Qing Yao, William Rudman, Ramez Kouzy, Kanishka Misra, Byron C. Wallace, and Junyi Jessy Li. What's in a name? Morphological shortcuts by LLMs in pharmacology. *arXiv preprint arXiv:2606.05616*, 2026. To appear in Proceedings of EMNLP 2026.

Dang Nguyen and Chenhao Tan. On the effectiveness and generalization of race representations for debiasing high-stakes decisions. In *Second Conference on Language Modeling (COLM)*, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Alejandro Salinas, Amit Haim, and Julian Nyarko. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*, 2024.

Zhengyang Shan and Aaron Mueller. Measuring mechanistic independence: Can bias be removed without erasing demographics? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4241–4265, 2026.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024.

Yuri Son, Seunghee Kim, Hyuhng Joon Kim, and Taeuk Kim. A unified mechanistic analysis of knowledge- and safety-based refusals. *arXiv preprint arXiv:2609.00760*, 2026. To appear in Proceedings of EMNLP 2026.

Gaurav Sood and Suriyan Laohaprapanon. Predicting race and ethnicity from the sequence of characters in a name. *arXiv preprint arXiv:1805.02109*, 2018.

Dananjay Srinivas, Saksham Khatwani, and Maria Pacheco. Toward a Gricean retreat: Probing LLMs for knowledge boundaries and referent specificity. *arXiv preprint arXiv:2608.13484*, 2026.

Dimitri Staufer and Kirsten Morehouse. What do LLMs associate with your name? a human-centered black-box audit of personal data. *arXiv preprint arXiv:2602.17483*, 2026.

Luxshan Thavarasa and Sivasuthan Sukumar. Test, then route: How language models execute in-context conditional rules across models and languages. *arXiv preprint arXiv:2608.04183*, 2026.

Leonor Veloso and Hinrich Schütze. GKnow: Measuring the entanglement of gender bias and factual gender. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10240–10260, 2026.

Shaowen Wang, Yiqi Dong, Ruinian Chang, Tansheng Zhu, Yuebo Sun, Kaifeng Lyu, and Jian Li. When bias pretends to be truth: How spurious correlations undermine hallucination detection in LLMs. *arXiv preprint arXiv:2511.07318*, 2025.

Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, and Steven Skiena. Nationality classification using name embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1897–1906, 2017.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

A Pre-registered criteria and outcomes

Every experiment in this paper fixed its pass criterion before it ran. Table 2 lists each criterion, what was observed, whether it was met, and, where it was not, our assessment of why: a property of the model, a criterion we specified poorly, or an instrument that did not transfer to our setting. Where a criterion admitted two readings, the reading applied first is the recorded verdict and the other is reported beside it. A criterion counts as not met if it fails in any model or clause it covers. Readings that are verdicts rather than pass/fail tests ("Reading", "Effect present", "Demographic", "Inconclusive", "Reported"), the falsifier row (a prediction, "Triggered") and a criterion not run are not counted. Of the 101 rows, 93 are pass/fail criteria, and by this rule 51 of them were not met.

Table 2: Pre-registered criteria and outcomes. Italics in the Observed cell give our assessment where a criterion was not met (a property of the model, a mis-specified criterion, or an instrument that did not transfer), or a qualification where it was.

Experiment	Criterion, fixed in advance	Observed, with the assessment in italics	Met
Main study			
Phenomenon in three models (§4)	Fabrication ≥ 0.60 and origin consistency ≥ 0.85 in 2 of 3 models	3 of 3 clear both	Yes
Origin patch, Gemma-2-2B (§7)	Follows donor ≥ 0.60; random < 0.10; abstention within 0.05 of clean	0.82; 0.00; 0.09 vs 0.06	Yes
Origin patch, other models (§7)	The same criterion	0.99, 0.87, 0.58. *Model: Qwen at 0.58, one valid answer short*	3 of 4
Confirmation set (§4)	On fresh names: sub-language ≥ 0.75; abstention χ^2 $p < 0.01$; actress ratio ≥ 3 at $p < 0.01$	0.9985; $3.5\text{e}{-19}$, $3.5\text{e}{-13}$; 7.4 at $9.9\text{e}{-11}$	Yes, 3 of 3
Recall heads (§6)	≥ 6 of the top-10 heads shared; ablation degradations within a factor of two	10 of 10 and 9 of 10; ratios 2.84 / 3.13 absolute, 1.06 / 1.36 relative. *Criterion: absolute or relative reading unspecified*	Second clause not met
Gate steering (§8)	Refusal moves ≥ 0.30 while consistency stays ≥ 0.90	$+0.71$ with no answers remaining; $+0.43$ at consistency 0.86, $n = 71$. *Criterion: the two clauses conflict*	No
Second labeller (Appendix M)	$\kappa \geq 0.80$ against a model-based judge	0.73 three-way; 0.98 on origin. *Instrument: the judge's error is one-directional*	1 of 2
Subject blocking	Associated-class bar cleared for ≥ 0.80 of items	0.30; correctly recalled real people reach 0.38. *Criterion: bar set above the real-person reference*	No
Linear relation embedding	Decodes the city from ≥ 0.80 of refusals	Faithfulness 0.00–0.08 against 0.92. *Instrument: does not transfer, three models*	No
Feed-forward share (§5)	≥ 0.60 at the name's last token	0.84 / 0.82 / 0.81	Yes
Patch on real people (§7)	Moves real birthplaces less than invented ones and above control	0.69 vs 0.82; random 0.00	Yes
Toponymic surnames (§4)	Consistency differs by ≤ 0.05 with and without place words	$+0.000$ to -0.019	Yes
Removing the prior (§9)	Abstention rises ≥ 0.20 and real recall falls < 0.05	-0.006 to $+0.031$; recall -0.142 to -0.266. *Model: the direction of the failure is the finding*	No, both
Gender axis (Appendix C)	Gender patch moves the actor rate with consistency ≥ 0.95	0 of 28 configurations; the direction is real, origin holds at 1.00. *Criterion: pooled rate against per-group levels*	No
Refusal direction and tradition (§9)	$\rho \geq 0.6$ with per-tradition abstention at the name span	$+0.53$, $+0.59$ there; $+0.67$, $+0.75$ at the last token. *Criterion: site ambiguity; short by 0.01*	No
The corpus prior (Appendix D), the detector (Appendix E) and OLMo 2 (§11)			
Corpus C1: city ranking	Corpus co-occurrence ranks a tradition's cities as the models do, $\rho \geq 0.5$ in ≥ 8 of 11	8 of 11	Yes
Corpus C2: per-name prediction	Per-name corpus prediction beats the modal city, McNemar $p < 0.05$, ≥ 3 of 4	Below the modal city in all 4. *Model: the within-tradition choice is not co-occurrence*	No, 0 of 4
Corpus C3: minority names	On minority names, $\geq 2\times$ the second-city baseline, ≥ 3 of 4	0.21 / 0.14 / 0.29 / 0.29 vs 0.43 / 0.22 / 0.46 / 0.55. *Model*	No, 0 of 4
Detector A1: invented vs real	Mechanism score beats the better detector baseline by ≥ 0.10 AUROC, both models	$+0.011$; -0.014 (probe 0.942, 0.889). *Criterion: a 0.10 margin over 0.94 is unattainable; in Gemma-3-4B, model*	No
Detector A2: recalled real people	The same on correctly answered real people	$+0.015$; $+0.005$. *As A1*	No
Detector A3: external names	AUROC ≥ 0.80 on external invented names	Not run: no external set	Not run
OLMo 2: phenomenon at an endpoint	OLMo 2 shows the phenomenon at an endpoint (the criterion of the first row)	1.00 / 0.99 base; 0.98 / 1.00 Instruct	Yes
OLMo 2 E1: order of emergence	Representation before or with behaviour, both before the gate	42B = 42B $<$ SFT	Yes

continued on the next page

Experiment	Criterion, fixed in advance	Observed, with the assessment in italics	Met
OLMo 2 E2: content from pretraining	Base and Instruct follow the donor within 0.10, both ≥ 0.60	0.924 vs 0.841	Yes
OLMo 2 E3: gate from post-training	Abstention < 0.05 at every pretraining checkpoint and ≥ 0.20 at Instruct	0.000; 0.179. *Criterion: threshold set before any data on this model; short by 0.021*	One clause not met
OLMo 2 E4: shared heads	Top-10 heads shared between base and Instruct, ≥ 6 of 10	8 of 10	Yes
The proposition, universality, amplification, the gate signal and real people (§9–§11)			
Falsifier of the model (Appendix F)	A pretraining checkpoint abstains > 0.10, or fabrication is off-tradition once the probe clears its ceiling	Bare prompt: max abstention 0.000, consistency 0.94 at 42B; few-shot prompt: up to 0.59 (189B, birth city; 0.99 at 9B by copying the example). *Criterion: no prompt clause was fixed; saved only by Corollary 1b's context clause, which the pre-registration did not state*	Triggered under the few-shot prompt
Shared prior U1: across models (Appendix H)	Cross-model divergence $\leq 2\times$ within-model in ≥ 8 of 11 traditions	5 of 11. *Model: the within-tradition city shape is model-specific*	No
Shared prior U2: closer to the corpus	Every model closer to the corpus prior than to uniform in ≥ 8 of 11	5 of 11. *Model*	No
Sharper than the population, M1 (Appendix H)	Fabricated share on the modal city exceeds the real share in ≥ 8 of 11, every model	Reference's modal city: 7 / 6 / 9 / 9 / 10 of 11; fabrications' modal city: 11 of 11 in every model. *Criterion: the reference distribution was unspecified*	No under the first reading
Gate signal F1: predicts refusal (Appendix C)	AUROC ≥ 0.80 for the gate signal, per model–signal pair	Refusal direction 0.967 / 0.883; entity latent 0.50 / 0.56. *Instrument: the latent detects an unknown entity, not the decision*	No (direction yes, latent no)
Gate signal F2: tradition beyond the signal	Likelihood-ratio test for a tradition effect beyond the signal (a verdict, not a pass)	$p = 4.1\text{e}{-10}$, $1.6\text{e}{-18}$; pooled Europe/South-Asia excess $+0.01$ / -0.01. *The between-group gap is in the signal*	Effect present, both
Gate signal F3: the two signals agree	The two Gemma-2-2B signals agree on the per-tradition excess	One never qualified. *Instrument*	No
Real people R1: recall gap (§10, Appendix G)	Counter-stereotypical people recalled ≥ 0.10 below fame-matched stereotypical people, ≥ 3 of 4	$+0.229$, $+0.152$, $+0.177$, $+0.199$	Yes, 4 of 4
Real people R2: where wrong answers land	Wrong answers in the name's tradition ≥ 0.50 and $\geq 2\times$ the record's country	0.95 / 0.02; 0.89 / 0.07; 0.93 / 0.04; 0.88 / 0.07	Yes
Real people R3: gap by fame	The gap ≥ 0.10 larger in the low fame band than the high, ≥ 3 of 4	$+0.15$, $+0.04$, $+0.07$, $+0.33$; direction as predicted in all four. *Criterion: margin set before any data on the effect size*	2 of 4
Real people R4: cost of the clamp	Under the clamp, recall retained falls ≥ 0.05 less for counter-stereotypical than for stereotypical people, both Gemmas	$+0.160$, $+0.186$; positive in all eight configurations	Yes
Additional controls, the map in the weights, and tests of the model (Appendices F–I)			
Head floors: shared set (Appendix I)	The shared set removes ≥ 0.50 of both abilities (relative), both Gemmas	0.99 / 0.93; 0.94 / 0.69	Yes
Head floors: attribution-matched	Attribution-matched floor removes < 0.20 of either ability, both Gemmas	0.07 / 0.00; -0.08 / -0.01. *In Gemma-2-2B the floor is one set (ranks 11–20) by construction*	Yes
Head floors: norm-matched	Norm-matched floor removes < 0.20 of either, both Gemmas	0.16 / **0.26**; -0.14 / 0.01. *Criterion: read consistency without conditioning on validity; the floor silences (validity 0.74), it does not redirect*	No (Gemma-2-2B)
Qwen and Llama circuit: head overlap	≥ 6 of the top-10 heads shared	8 and 9 of 10	Yes
Qwen and Llama circuit: ablation	The shared set removes ≥ 0.50 of both abilities	0.53 / 0.37; 0.32 / 0.26, validity unchanged. *Model: damage falls with head count (208, 272, 784, 1,024 heads)*	No
Qwen and Llama circuit: floors	Matched floors < 0.20	-0.04 to 0.06	Yes

continued on the next page

Experiment	Criterion, fixed in advance	Observed, with the assessment in italics	Met
Real-people patch by fame band	High band reproduces 0.69; low $\geq$ high; monotone within 0.05	0.69 exactly; 0.69 / 0.78 / 0.81	Yes
Fluency: degeneration	Degeneration within 0.02 of clean at the configuration	$+0.022$, $+0.046$, $+0.007$, 0.000. *Model*	No (Gemmas)
Fluency: external perplexity	External-model log-perplexity of patched answers $\leq 1.10\times$ clean	1.46–2.48 (question in context); 1.18–1.25 (name removed). *Model, and the external model's own name prior; 4–8% of patched names drift in spelling, against 1–2% clean*	No, all four
OV readout: positive control	Attested city in the top 10 of the joint readout for ≥ 0.50 of correctly recalled people	0.61, 0.52, 0.50, 0.49, 0.43 (random heads 0.00). *Criterion: bar set before any data on the readout; Llama one person short*	No (3 of 5)
OV readout: own cities	Own cities above every unrelated tradition and both floors in ≥ 8 of 11 traditions	11, 11, 10, 10, 10 of 11. *Where the positive control holds; reported without standing where it fails*	Yes
OV readout across families	The own-cities criterion with standing in ≥ 4 of 5 families; carrier in the middle third	3 of 5; carriers at depth 0.59–0.85. *Criterion, both clauses*	No
Frozen attention patterns	Reference arms reproduce the patch result; frozen lift within 0.10 of unfrozen; reverse lift < 0.20	Reproduced; frozen lift 1.000 / 0.932 against unfrozen 1.000 / 0.917; reverse lift $+0.000$ / -0.008	Yes, both Gemmas
Feed-forward writers: probe	Early band's removal takes the patch-layer probe below the 0.92 ceiling; random blocks do not	0.80 vs 0.98; 0.06 vs 0.93	Yes, both
Feed-forward writers: consistency	Origin consistency falls ≥ 0.30 under the band, < 0.10 under controls	-0.05 vs -0.01; -0.96 vs -0.01 with the name echoed in 0.2%. *Criterion: no name-echo guard; Gemma-3-4B passes only by name corruption, the late blocks rewrite the origin*	No (Gemma-2-2B); yes (Gemma-3-4B)
What the gate reads	Frequency reading (frequency $\geq$ half of tradition's R^2 and the patch moves the signal < 0.25 of the gap) or demographic (tradition retains $\geq$ half after frequency, or the patch moves it > 0.50)	Retained 0.76 / 0.75; patch 0.63 / 0.87 of the gap. *Both models*	Demographic
Gate follows the shift, UA1: forward	Forward patch raises Baltic abstention ≥ 0.50 of the clean Tamil–Baltic gap A; random $\leq 0.25A$	$1.95A$ / $0.95A$; random $-0.41A$ / $0.03A$	Yes, both
Gate follows the shift, UA2: reverse	Reverse patch lowers Tamil Nadu abstention $\geq 0.25A$; random $\leq 0.25A$	$0.73A$ / $0.32A$; random $0.39A$ / $0.13A$. *Model: a random shift also lowers it*	No (Gemma-2-2B)
Gate follows the shift, UA3: surviving answers	Surviving answers follow the donor ≥ 0.60 in both arms	7/9, 28/28; 2/32, 20/22. *Model: survivors of the forward patch keep their content*	No (Qwen)
OV map as a matrix: rank	Effective rank of the map on the tradition subspace ≥ 6 of 11	8.85 / 9.06 / 8.91 / 9.05	Yes, four models
OV map as a matrix: diagonal	Own cities above every unrelated tradition's in ≥ 8 of 11 rows	11 / 11 / 10 / 10	Yes, four models
σ as evidence: real vs invented	Real vs invented AUROC ≥ 0.90	0.877 / 0.844 / 0.865 / 0.869. *Model: the quantity is not evidence*	No, four models
σ as evidence: blind to tradition	Blind to tradition: $R^2 \leq 0.05$; within-tradition AUROC ≥ 0.85	R^2 0.26–0.44; min 0.74–0.81. *Model*	No, four models
σ as evidence: graded with fame	Graded with fame: Spearman ≥ 0.30	0.06–0.14. *Model*	No, four models
σ as evidence: predicts recall	Predicts correct recall: AUROC ≥ 0.70	0.40–0.51. *Model*	No, four models
One-parameter test, sitelinks as exposure (Appendix F)	One α fitted on recall predicts the fame-band patch inside its intervals	order right; 1 of 3 inside (2 of 3 on all answers). *Criterion: $1 - w$ assumed the record term does not move*	No
Sampling vs greedy (Remark 1a): distance	Sampling halves the distance to the population in ≥ 8 traditions	0 of 9 (ratios 0.61–0.94). *Model: the learned prior is sharper than the population*	No
Sampling vs greedy (Remark 1a): modal share	Sampled modal share within $\times 2$ of the population's in ≥ 8	3 of 9. *Model*	No

continued on the next page

Experiment	Criterion, fixed in advance	Observed, with the assessment in italics	Met
Selection on the gate (Observation 3b)	Survivors of the forward patch decoded as the donor less often than the refused	0.986 vs 0.986, $p = 0.71$. *Model: the content did not follow the moved representation*	No
Name counts track sitelinks	Sitelinks track the Dolma full-name count, Spearman ≥ 0.5	0.638 over 2,156	Yes
One-parameter test, corpus counts as exposure	One α on count terciles predicts the fame-band patch inside its intervals	0.73 / 0.81 / 0.82 vs 0.71 / 0.85 / 0.83 (valid answers) and 0.69 / 0.78 / 0.81 (all answers); 3 of 3 either way; at most 1 of 3 with both sides on valid answers	Yes
Counter gap at equal co-occurrence (Appendix G)	The counter gap keeps $\geq$ half its size after name–city co-occurrence, $p < 0.01$, in ≥ 3 of 4	retained 0.94–0.97	Yes, 4 of 4
OLMo 2: map in the weights (Appendix H)	The map in the carriers' weights by 42B or 189B	9 of 11 at 42B	Yes
OLMo 2: recall heads persist	Top-ten heads overlap ≥ 6 (856B/end; end/Instruct)	8, 8	Yes
Refusal push: graded by fame (Appendix I)	Refusal ordered by fame band at every strength, low $-$ high ≥ 0.30 at the working strength	0.45 / 0.25. *Model: the direction is weaker in Qwen*	No (Qwen); yes (Gemma-2-2B)
Refusal push: known people kept	Known people keep ≥ 0.90 of correct answers	0.93 / 0.95	Yes, both
Refusal push: tradition spread	Tradition spread halves (university) and ≤ 0.20 (birth city)	0.16, 0.36 / 0.34, 0.35. *Model: exposure-graded, not tradition-neutral*	No, both
Refusal push: content	Content untouched	0.98–1.00	Yes, both
Tradition removed at the gate: spread	Removing the tradition subspace at the gate halves the tradition spread on both questions	university -31% / -17%; birth rose / -20%. *Model: the gate reads tradition beyond that subspace*	No, both
Tradition removed at the gate: known people	Known people keep their answers	kept 1.00 / 0.99	Yes, both
Tradition removed at the gate: precision	Precision on real people rises by 0.10	-0.002 / -0.002. *Model: the projection changes the read of origin, not of evidence*	No, both
Tradition removed at the gate: content	Content untouched	consistency 0.99; degeneration within 0.01	Yes, both
OLMo 2 SFT data audit (Appendix H)	Refusal data by tradition explains the SFT gate ($\rho \geq 0.5$) or not (< 0.3)	one tradition with ≥ 30 examples. *Data: 258 questions in 939,344*	Inconclusive
Known people, other attributes, a prompt with no person, and the base model's prompt (§6, §7, §11)			
Real-person patch on known people	Follows donor ≥ 0.50 among high-band people the clean model recalls, above random	0.68 [0.53, 0.82], $n = 38$; random 0.00	Yes
Same heads on recalled people	Top-ten overlap ≥ 8 of 10 on correctly recalled real people, both Gemmas	10; 9	Yes
Occupation under the patch	≥ 0.80 of known occupations kept, within 0.10 of random	0.91 against 0.96	Yes
Birth year under the patch	≥ 0.80 of known birth years kept, within 0.10 of random	0.51 against 0.76. *Model: the patch partly disturbs the person*	No
City prompt with no person	Shared heads in its top ten: ≤ 5 person-specific, ≥ 8 a shared stage	9 of 10; 8 of 9; 6 of 8; 7 of 9. *Shared answer-writing stage (Gemmas)*	Reading
Base checkpoints, few-shot prompt	Any checkpoint declining ≥ 0.10 softens the claim about post-training	per checkpoint 0.17 to 0.99 from 9B on. *Post-training claim restricted to the bare prompt (§11)*	Reading
Real people born in Tamil Nadu or Sri Lanka (§9)	No pass rule; post hoc	Low band, university: declined 0.47 vs 0.31 (Gemma-2-2B), 0.52 vs 0.38 (Qwen); birth-city recall equal or higher except Gemma-2-2B (0.23 vs 0.32). *Post hoc; not counted*	Reported
Uses of the mechanism (§10, §12, Appendix J)			
Record replaced for known people	Recall gap ≥ 0.05 among people whose occupation or birth year is answered correctly, ≥ 3 of 4	$+0.21$, $+0.12$, $+0.19$, $+0.15$	Yes, 4 of 4

continued on the next page

Experiment	Criterion, fixed in advance	Observed, with the assessment in italics	Met
Wrong answers from the name, known people	≥ 0.70 of wrong answers in the name's tradition	0.95, 0.86, 1.00, 0.87	Yes, 4 of 4
Forecast: country	Country hit ≥ 0.85 on fresh names; ≥ 0.75 on real people's wrong answers	0.66–0.70; 0.34–0.37. *Criterion: one country per multi-country tradition*	No
Forecast: against name-blind	Beats a name-blind forecast by ≥ 0.20	$+0.59$ to $+0.64$	Yes
Forecast: calibration	Per-tradition abstention error ≤ 0.05	birth city ≤ 0.05; university 0.11, 0.18. *Model*	No
Detector: self-consistency fails	Self-consistency AUROC ≤ 0.65 on counter-stereotypical people, ≥ 3 of 4	0.71–0.81. *Model: it does not fail*	No
Detector: twin agreement	AUROC ≥ 0.80, ≥ 3 of 4	0.38–0.49. *Model: a predicted fact looks like the prior*	No
Detector: adds to probability	≥ 0.05 over token probability, or $+0.03$ combined	no gain. *Model*	No
Selective answering	Risk at 80% coverage $\leq 0.8\times$ probability alone	ratio 1.00. *Model*	No
Corrected gate: known people	≥ 0.95 of known high-band answers kept	0.99; 1.00	Yes, both
Corrected gate: declines unknown	Invented university abstention ≥ 0.80	0.92; 0.76. *Model*	No (Qwen)
Corrected gate: fair	Tradition spread ≤ 0.15	0.16; 0.34. *Model: predicted in advance*	No, both
Corrected gate: content	Consistency of answers still given ≥ 0.98	0.99; 1.00	Yes, both
Corrected gate: evidence	Beats placebo on kept by ≥ 0.05	$+0.02$; $+0.01$. *Criterion: both at ceiling*	No, both

B ENTITY CONSTRUCTION AND CERTIFICATION

Construction. For each of the eleven traditions we curated a pool of twelve given names and twelve surnames that are real, attested names of that tradition (for example Priit, Kadri, Tõnu and Ozoliņš, Bērziņš, Mägi for the Baltic pool; Ilamparithi, Senthamarai and Subramanian, Palaniappan for Tamil Nadu). Base people are seeded combinations of one given name and one surname from the same pool. Twins recombine a base person's given name with a surname from another tradition and vice versa. Sri Lankan Tamil shares a language with Tamil Nadu Tamil and a country with Sinhala, so these related pairs are never twinned, because a fabricated Tamil city could not tell them apart. The two Tamil traditions carry most of the tradition gap on invented names in §9; that section's test on real people does not use these pools.

The seven checks. A name is certified only if every check is clean. A service that cannot be reached leaves the name uncertified until re-checked. (1–3) Zero exact-string count in Dolma v1.7, the Pile and the OLMo mix, through infini-gram; (4) no exact match against Wikidata labels and aliases after Unicode and case normalisation; (5) no near match at similarity ≥ 0.93, the threshold of Lu et al. (2026); (6) no English Wikipedia page title; (7) no quoted full-text hit in English Wikipedia. The certified names were afterwards also compared with the invented names models habitually produce themselves, the "ghost names" of Brzozowski & Chung (2026), with no collision.

What the battery found. Of 1,292 candidate names checked, 1,206 were certified and 86 rejected, several by more than one check: Dolma count above zero 72, OLMo mix 34, the Pile 22, a quoted Wikipedia hit 20, a Wikidata exact match 13, a near match 1. Thirteen European base names failed a check other than the Dolma count. Four are real people with Wikidata entries (for instance "Vesna Kovačevič", an exact match after normalisation to a real Vesna Kovačević). Each was replaced by a certified name of the same tradition and its twins rebuilt. Sixty of the 86 rejections were candidates for the three South Asian traditions, mostly caught by corpus counts ("Ezhilarasan Natarajan" occurs 77 times in Dolma). Tamil naming draws on a smaller pool of given names and patronymics, so random combinations are more often real, and the battery handled this by trying further candidates. The per-name ledger with every check value is released.

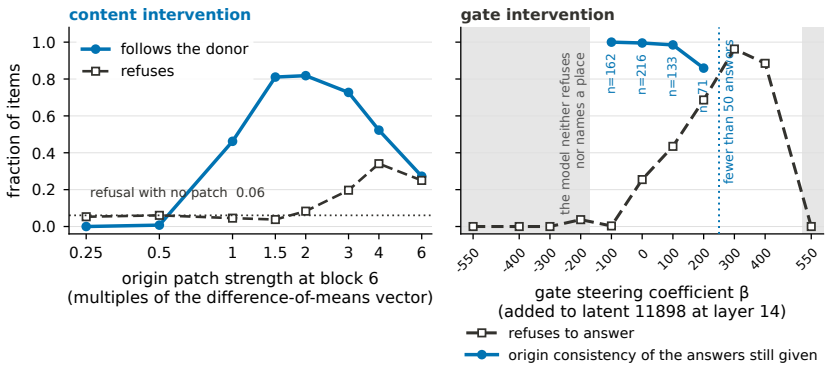


Figure 5: **The two readers, moved one at a time.** Gemma-2-2B, shared y axis. *Left:* the birth-city question on the 132 held-out names of Table 1 (clean refusal 0.06). As the origin patch switches the fabricated city to the donor's tradition, refusal stays at its unpatched level (on the university question it does not, §8). *Right:* both questions pooled, 161 held-out names (clean refusal 0.25). Steering the known/unknown-entity latent moves refusal from 0.25 to 0.96 and to 0.00 while the answers still given stay origin-consistent. The consistency line is drawn where at least 50 placed answers remain (n at each point), and in the grey bands the model neither refuses nor names a place.

What is and is not claimed. We claim that no public text exists from which a model could have learned a fact about any of these people: every name is absent from three open pretraining corpora totalling several trillion tokens and from the encyclopaedic record. We do not claim that no person of that name has ever lived, nor that the names are absent from the proprietary training data of the four models. The OLMo mix is the training corpus of the OLMo family and not of these models, so Appendix H repeats the study on OLMo-2-7B, for which the stage-1 statement is exact.

The real-person set. The 1,155 real people are Wikidata humans (P31 Q5) with an attested birth city (P19) in the country or region that supplies each tradition, restricted to two-token Latin-script names so that the given-name and surname spans are defined as they are for the invented names, and stratified into three popularity bands by Wikipedia sitelink count (30 or more, 8 to 29, 1 to 7), sampled with a fixed seed and up to 40 per tradition per band. Sri Lanka is one country and two traditions, and Wikidata has no field for which. The split by name components matched too few people to be usable (5 and 63 of 478), so Sri Lankan people form one group, giving ten groups against the invented set's eleven. No absence check is run on real people; their presence in the corpus is what the set is for.

Component provenance and human audit. Every component was looked up as an exact label among Wikidata's given-name and family-name items and its declared language (P407) or country of origin (P495) compared with the tradition. Of 264 components, 117 have a name item declaring a language or country of the tradition, 88 have a name item with no origin declared, and 59 have no name item. The per-component records are released. Coverage is uneven in the way Wikidata's onomastic coverage is: Baltic 23 of 24 consistent, Slovene–Croatian 21, Finnish 18, Hungarian 17, Dutch 15, Portuguese 9, Nordic 8, Welsh–Irish 6, and none of the 72 Tamil Nadu, Sri Lankan Tamil and Sinhala components, 37 of which have no name item at all. Absence from Wikidata is not evidence against a name. For the three South Asian traditions the provenance check is uninformative, and those components rest on a native-speaker audit. That audit, with a protocol fixed in advance, flagged no component.

C THE UNEQUAL GATE, BY TRADITION

Figures 5 and 6 give the per-tradition detail behind §9. The two paragraphs below add the gender axis as a cross-control and locate where the inequality enters the refusal signal.

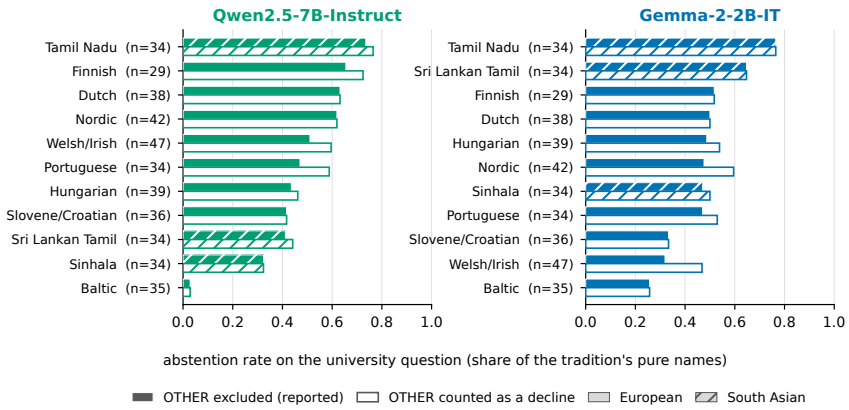


Figure 6: **The gate reads the tradition.** Abstention on the university question by the tradition of the invented name, for the two models that abstain at all. Every person is equally non-existent. European traditions plain, South Asian hatched. Filled bars count explicit declines only, and open bars also count the labeller's ambiguous category, which is under-detected refusal. The conclusion does not depend on the coding.

Origin is not the only axis. A gender direction at the same span moves 0.71 of movable names to the target pronoun with origin consistency at 1.00 [0.992, 1.000] on 430 to 466 answers, and the origin patch leaves the pronoun alone, so each axis serves as a null control for the other. Its criterion (Table 2) required the pooled actor rate to move. The effect appears in per-group levels instead, so the criterion is not met.

Where the unequal gate enters. Two scalar gate signals were read per invented name at the last prompt token of the university question, on the 402 pure names and the 1,128 of the confirmation set: the projection onto the knowledge-refusal direction of §8 (refit on its own seeded fit set, those names excluded from every fit), and Ferrando et al.'s unknown-entity latent. Rule F1 asked each signal to reach AUROC ≥ 0.80 for refusal. The refusal direction does, 0.967 [0.959, 0.975] in Gemma-2-2B and 0.883 [0.865, 0.901] in Qwen2.5-7B. The unknown-entity latent does not, 0.50 at the last prompt token, where it never fires, and 0.56 at the name's last token, where it fires for 81 percent of names with nearly equal means for refused and answered. Rule F2, a likelihood-ratio test of refusal on the signal against refusal on the signal plus tradition indicators, returns *unfair threshold* in both models ($p = 4.1\text{e}{-10}$ and $1.6\text{e}{-18}$). What that verdict names needs care. At the level the fairness finding is about, Europe against South Asia, the gap enters through the *signal*. Given the model's own refusal signal, South Asian names are refused at the rate it predicts (observed minus predicted $+0.01$ and -0.01, as a direction fitted to separate refused from answered names must give), and the signal is simply higher for them (Tamil Nadu 40 against Baltic 10 on Qwen's direction, 23 against 7 on Gemma's). The indicators are significant because individual traditions sit off the common curve in both directions. Sri Lankan Tamil, for one, is refused less than its signal predicts in Qwen. F3, agreement of the two Gemma-2-2B signals, fails because the latent never qualified. A tradition indicator can also absorb misfit of a single logistic curve, so "threshold" is the verdict of the pre-registered test rather than a demonstration of a per-tradition cut.

D THE TRADITION PRIOR IN PRETRAINING TEXT

Can the tradition prior be read off pretraining text? For each of the 402 pure (non-twin) certified names we counted, in Dolma v1.7 through infini-gram, the documents in which its surname, and separately its given name, co-occurs with each candidate city of its tradition. The candidates are the cities the four models actually fabricated for that tradition (116 in all). The corpus score of a city for a name is the add-one estimate of $\log P(c \mid s) + \log P(c \mid g)$ and the corpus prediction is its arg max. Three rules were fixed (Table 2): **C1**, corpus co-occurrence mass rank-correlates with pooled model frequency at $\rho \geq 0.5$ in at least eight of eleven traditions; **C2**, the per-name corpus prediction beats the tradition's modal city at exact McNemar $p < 0.05$ in at least three of four models; **C3**, on

names not given the modal city, the corpus prediction hits at twice the rate of the tradition's second city.

C1 passes, exactly at its threshold. Eight traditions clear $\rho \geq 0.5$ (Nordic 0.77, Dutch 0.92, Hungarian 0.84, Welsh–Irish 0.59, Baltic 0.89, Tamil Nadu 0.78, Sri Lankan Tamil 0.61, Sinhala 0.95). Finnish, Portuguese and Slovene–Croatian miss at 0.42, 0.46 and 0.47. In six of eleven the corpus's top city is the models' top city. Per model the correlations are weaker, and the rule was written on the pooled frequency.

C2 and C3 fail in every model, in the direction no confound explains. Here n counts the names whose fabricated city is one of the 116 candidates, so it differs from the counts of Figure 2.

model	n	corpus	modal city	corpus − modal [95%]	minority: corpus / 2nd city
Gemma-2-2B	373	0.440	0.686	−0.247 [−0.306, −0.188]	0.214 / 0.427
Gemma-3-4B	365	0.310	0.422	−0.112 [−0.164, −0.060]	0.142 / 0.218
Qwen2.5-7B	364	0.464	0.536	−0.071 [−0.132, −0.014]	0.290 / 0.456
Llama-3.1-8B	377	0.451	0.523	−0.072 [−0.130, −0.013]	0.289 / 0.550

The name-aware corpus prediction is below the name-blind modal city in all four models, with every interval on the paired difference excluding zero, and below the second-city baseline on the minority names. Neither component carries it: given-name-only and surname-only predictors are each below the modal baseline in nine of ten traditions. As stated in advance, a C2 or C3 failure means that the choice *among* a tradition's cities for a given name is not recoverable from surface co-occurrence of the name's components, and the modal-city finding stands on its own. This agrees with the cross-model result of §4. The tradition is shared and corpus-shaped, and the within-tradition choice is model-specific.

Two confounds stated in advance are visible in the counts: co-occurrence is undirected, so a city's own press co-occurs with local surnames whether or not anyone was born there (the corpus's top Baltic city is Tallinn, not Riga), and a candidate that is also a name inflates its own count (Sobral, a Portuguese surname). Neither is corrected, and neither bears on C2 or C3, whose failures run the other way.

E A FABRICATION DETECTOR BUILT FROM THE MECHANISM: A NEGATIVE RESULT

Cheang et al. (2026) report that hallucinations produced by recall have hidden states that overlap with factual outputs, so truthfulness probes fail on them, and Wang et al. (2025) that correlation-driven hallucinations evade confidence and probe detectors. Our mechanism is consistent with both, and implies that the two cases differ at the input to the circuit. We fixed two scores and two baselines in advance. τ is the ratio of the in-subspace to the total signed direct-logit attribution of the produced city, summed over heads, the subspace being the span of the eleven origin directions at the output of each head's block. σ is the fraction of the span-averaged name representation's squared norm, grand mean removed, that lies outside that subspace at the output of the block given in the table below. The baselines are the probability of the produced city token and a logistic truthfulness probe on the last-position residual, trained on the fit half. The data are the 220 real and 220 invented people of §6, fit and held-out halves stratified by kind. Rule A1 asked for a margin of at least 0.10 AUROC over the better baseline in both models.

model	σ	τ	p(city)	probe	A1 margin
Gemma-2-2B (block 9)	**0.954**	0.665	0.798	0.942	+0.011
Gemma-3-4B (block 6)	0.874	0.537	0.864	**0.889**	−0.014

The rule fails in both models, and the premise did not transfer. A plain probe already separates certified-absent from real people at 0.942 and 0.889. That is consistent with Ferrando et al. (2025). The known-entity signal is linearly readable, and certified-absent names sit at its far end, whereas the hallucinations of Cheang et al. and Wang et al. concern entities the model partly knows. Our

contrast is the easy one for a probe, so a mechanism-derived score had nothing to add. σ carries signal: it beats the probe in Gemma-2-2B but not in Gemma-3-4B. τ separates them less well than either baseline. In Gemma-2-2B the failure is the rule's, since a 0.10 margin over 0.942 has a ceiling of 0.058. In Gemma-3-4B it is the model's. A3, on external invented names, was not run because no external set was available. The subspace must be estimated at the same hidden state as the representation it is applied to. Estimating the subspace one block before the representation's block changes the margins to $+0.016$ and -0.037 (A1) and $+0.023$ and $+0.015$ (A2), with the same verdicts.

F A MODEL OF FABRICATION AS RECALL ON A PRIOR

This appendix states the argument of §10 as a proposition with an explicit assumption, adds the model of real-person recall that the paper's results test, and states the protocol by which a representation is shown to carry a prior. Each part has one test on our data and each test can fail. The outcomes are in Table 2.

F.1 PROPOSITION 1: FEATURE-BOTTLENECK OPTIMALITY

Let D be the training distribution over pairs (x, y), a name and an answer to the attribute question drawn from a finite answer set $\mathcal{Y}$, and let $\phi(x)$ be the features of x the model can carry from seen names to unseen ones. The *feature-bottleneck class* is $\mathcal{Q} = \{q : q(y \mid x) = g(y, \phi(x))\}$, with each $g(\cdot, \phi)$ a probability distribution on $\mathcal{Y}$. For $q \in \mathcal{Q}$ write $q(\cdot \mid \phi)$ for $g(\cdot, \phi)$.

Assumption A. For a name with zero training count, the model's conditional distribution over y depends on x only through $\phi(x)$. Restricted to unseen names, the model is $g(\cdot, \phi(x))$ for some g with each $g(\cdot, \phi)$ a distribution over answers, and this g minimises $\mathbb{E}_D[-\log g(y, \phi(x))]$. The second clause is needed because the training loss is evaluated only on seen names, which a model can memorise. Without it the loss constrains nothing about unseen names.

Assumption A is the paper's hypothesis written formally, and Proposition 1 does not establish it. The proposition only derives what the hypothesis implies, so that the implications can be tested. Our data bear on the assumption directly. The certification battery gives every invented name a zero count in the three corpora, so nothing about the string was seen. The lexical-ceiling analysis (Appendix N) says which features are decodable from the name span above spelling, and the sub-language probe clears its ceiling by 0.05–0.07. This shows which features are present at the name span. That the answer depends on the name only through them is what the origin patch of §7 tests causally.

Proposition 1. *Write $p(y \mid x)$ for the conditional of D and let $p(y \mid \phi) = \mathbb{E}_D[\, p(y \mid x) \mid \phi(x) = \phi\,]$, the conditional of y given $\phi(x) = \phi$ under D; expectations over ϕ below are over the law of $\phi(x)$, $(x, y) \sim D$. Then (i) the minimiser of $\mathbb{E}_D[-\log q(y \mid x)]$ over $\mathcal{Q}$ is $q^*(y \mid x) = p(y \mid \phi(x))$; and (ii) for every $q \in \mathcal{Q}$,*

$$\mathbb{E}_D[-\log q(y \mid x)] - \mathbb{E}_D[-\log q^*(y \mid x)] = \mathbb{E}_\phi\big[\mathrm{KL}\big(p(\cdot \mid \phi) \,\|\, q(\cdot \mid \phi)\big)\big] \geq 0,$$

with equality if and only if $q(\cdot \mid \phi) = p(\cdot \mid \phi)$ for almost every ϕ under that law.

Proof. Since $q^*(y \mid x) = p(y \mid \phi(x))$ has the form $g(y, \phi(x))$, $q^* \in \mathcal{Q}$, and $\mathbb{E}_D[-\log q^*] = \mathbb{E}_\phi[H(p(\cdot \mid \phi))] \leq \log|\mathcal{Y}| < \infty$ because $\mathcal{Y}$ is finite, so the difference in (ii) is defined. By the tower property, $\mathbb{E}_D[-\log q] = \mathbb{E}_\phi\, \mathbb{E}[-\log g(y, \phi) \mid \phi]$, and the inner expectation is $-\sum_y p(y \mid \phi) \log g(y, \phi)$, because the integrand depends on x only through $\phi(x)$ ($q \in \mathcal{Q}$) and the conditional law of y given $\phi(x) = \phi$ is $p(\cdot \mid \phi)$ by definition. Subtracting the same quantity at $g = p(\cdot \mid \phi)$ leaves $\sum_y p(y \mid \phi) \log \frac{p(y|\phi)}{g(y,\phi)}$ inside the outer expectation, which is the KL divergence; Gibbs' inequality gives the sign and the equality condition. □

For an unseen name, the loss-minimising answer is the prior predictive of its features, the population conditional rather than a guess at random. §4 measures it (origin consistency 0.99–1.00) and C1 finds it in the corpus (Appendix D).

Remark 1a (greedy sharpening). Let $m(\phi) = \arg\max_y q^*(y \mid \phi) = \arg\max_y p(y \mid \phi)$. Under greedy decoding every name with features ϕ answers $m(\phi)$, so over a tradition the share of outputs equal to a city c is $\Pr_\phi[m(\phi) = c]$, whereas the population share is $\mathbb{E}_\phi[p(c \mid \phi)]$. If c is the mode for every ϕ in the tradition, the greedy share is 1. We tested whether this explains the over-concentration found by rule M1 (Appendix H). It does not. At temperature 1, twenty samples for each of the 402 names move the city distribution toward the reference population in every tradition, but the total-variation distance falls by only 6 to 39 percent (the rule asked for a halving, met in 0 of 9 traditions with a reference), and the reference's modal city lands within a factor of two of its population share in 3 of 9. The learned conditional is itself sharper than the population. Samples stay in the name's tradition (0.995). The identification of $p(\cdot \mid \phi)$ with the census of births therefore holds at the level of the tradition and not of the city, and the corpus reading fares no better at the city level (rules C2, C3 and U2 fail, Appendices D and H). Proposition 1 stands at the level of the tradition only.

Corollary 1b (abstention needs data). The support of $q^*(\cdot \mid \phi)$ is the support of $p(\cdot \mid \phi)$. If the pretraining text rarely declines after such questions, the minimiser rarely abstains unless the context supplies declining text. OLMo 2 fits this in both directions. Under the bare completion prompt no checkpoint declines, and when a few-shot prompt supplies one "Unknown" answer the base checkpoints decline, at rates that depend on the name's tradition (§11), and the SFT-mix audit of Appendix H finds too few per-tradition refusal examples to teach the pattern directly.

Remark 1c (rate and content). Kalai & Vempala (2024) fix the rate of wrong arbitrary facts by the frequency of once-seen facts. Proposition 1 fixes their content as draws from $p(\cdot \mid \phi)$. The two are different functionals of the data and move on different handles (§8).

F.2 Proposition 2: the evidence mixture is a posterior predictive

For a real person x with record city $r(x)$, seen n_x times, let the model hold the tradition's prior predictive $p(\cdot \mid \phi(x))$ as a Dirichlet prior of concentration α: $\theta_x \sim \mathrm{Dirichlet}(\alpha\, p(\cdot \mid \phi(x)))$, with n_x observations all equal to $r(x)$.

Proposition 2. *The posterior predictive is* $q(y \mid x) = w_x\, \mathbf{1}[y = r(x)] + (1 - w_x)\, p(y \mid \phi(x))$ *with* $w_x = n_x/(n_x + \alpha)$.

Proof. Dirichlet–multinomial conjugacy: the posterior is $\mathrm{Dirichlet}(\alpha p(\cdot \mid \phi) + n_x \mathbf{e}_{r(x)})$, where $\mathbf{e}_{r(x)}$ is the unit vector at $r(x)$, and its mean, which is the posterior predictive, is $(\alpha p(y \mid \phi) + n_x \mathbf{1}[y = r(x)])/(\alpha + n_x)$. □

The prior requires $p(\cdot \mid \phi(x)) > 0$ on every answer. For a counter-stereotypical person take π_x small but positive, and the formula is unchanged (with $n_x > 0$ the posterior is proper in the limit $\pi_x \to 0$).

Given the Dirichlet prior, the weight is derived from one parameter α and the count n_x. Two further assumptions connect it to our data. First, the predictions are predictive probabilities while our answers are greedy, so a group's observed rate is read as its mean predictive probability (Remark 1a shows greedy decoding sharpens, so this is an approximation). Second, n_x, the number of observations of the fact, is measured by the corpus count of the person's full name. Writing $\pi_x = p(r(x) \mid \phi(x))$ for the prior mass on the person's own record city, the proposition implies, with no new data: recall of the record is $w_x + (1 - w_x)\pi_x$; the gap between a stereotypical person ($\pi_x = \pi_{\text{stereo}}$, the tradition's share) and a counter-stereotypical one ($\pi_x \approx 0$) at equal exposure, writing w for their common w_x and later w_b for its value at a fame band's median count, is $(1 - w)\pi_{\text{stereo}}$ (rules R1 and R2 of Appendix G); the gap shrinks with exposure (R3); the origin patch acts on the $(1 - w)$ term, so it moves weakly known people more; deleting the prior removes $(1 - w)\pi_x$ from every stereotypical person's recall (§9, R4). The test that can fail is that one α, fitted by maximum likelihood on the per-person recall of both groups with exposure taken at the band median, must predict the patch effect by band inside its intervals (the next paragraph).

The one-parameter test. If the model behaves as this posterior predictive, one α fitted on recall must predict the patch's effect on real people, $\widehat{\mathrm{FD}}_b = (1 - w_b)\,\mathrm{FD}_{\text{inv}}$, with $\mathrm{FD}_{\text{inv}} = 0.82$ the effect on invented names ($w = 0$, 108 of 132 answers). The exposure n is the count of the person's

full name in Dolma, measured through infini-gram for all 2,156 distinct real people of the paper (the 1,155 of Appendix B and the 1,548 of Appendix G, with 547 in both). Fitted by maximum likelihood on the 1,524 of the 1,548 for whom Gemma-2-2B asserts a birth city (count terciles, medians 8,860 / 325 / 15), $\alpha = 28{,}600$. Mapping the patch experiment's three fame bands to their own people's median counts (3,360 / 225 / 33) gives $w_b = 0.105/0.008/0.001$ and predicted follows-donor 0.73 / 0.81 / 0.82. Counted over all answers, the denominator of FD_{inv}, the observed values are 0.69 [0.60, 0.78], 0.78 [0.70, 0.85] and 0.81 [0.73, 0.88], so all three predictions fall inside in the right order. The pre-registered comparison counted the observed values over valid answers only (0.71, 0.85, 0.83) against the same all-answers prediction, and also puts all three inside. With both sides on valid answers ($\mathrm{FD}_{\text{inv}} = 0.92$, predictions 0.83 / 0.92 / 0.92) at most one band is inside, so the result depends on counting abstentions and malformed answers as not following, as Table 1 does. The test also has little power. Because w_b is small in the two weaker bands, their predictions stay close to FD_{inv} and fall inside their intervals for any α above about 1,300. Only the most famous band constrains α, and any value between about 9,200 and 65,500 passes. With Wikipedia sitelinks as the exposure proxy, the same fit has the right order and the wrong size (0.59 / 0.73 / 0.80, with two of three inside on all answers). The model under-predicts stereotypical recall in the two weaker bands (0.29 and 0.33 observed in the counterfactual set's stereotypical group, 0.07 modelled), which is Remark 1a's finding again: the learned prior puts more mass on the tradition's frequent cities than the population does, and the fit takes π_x from the census of births (Appendix H), not from the corpus distribution Proposition 1 names.

Falsifier, checked. Two observations were named in advance that would contradict Proposition 1 and Corollary 1b: a pretraining checkpoint that abstains above 0.10 on invented names (against Corollary 1b), or one whose fabricated cities do not match the tradition once the representation clears its ceiling (against (i)). Under the bare completion prompt neither occurred. The maximum pretraining abstention is 0.000 over seven checkpoints, and at the first checkpoint whose probe clears the ceiling (42B tokens) origin consistency is 0.94, rising to 1.00. Under a few-shot prompt with one declining example, checkpoints abstain on up to 0.59 of questions (189B, birth city), so the falsifier as written was triggered. It is saved only by Corollary 1b's context clause, which the pre-registration did not state, and the ledger records it as under-specified. A control prompt without the "Unknown" example was not run, and the 9B checkpoint copies the example for 0.99 of birth-city prompts, so part of the few-shot abstention is template copying.

F.3 Definition 3: a protocol for identifying a prior carrier

Let $X \to \Phi \to \{C, G\} \to Y$ be the structural model: input, a representation, the content written at the answer position, the gate at the last prompt position, the output. A representation Φ is a *prior carrier* for $\phi \to y$ if (i) *sufficiency*: setting Φ to a donor's value at the same norm moves C to the donor prior while random directions do not; (ii) *necessity*: removing Φ's readers removes both fabrication and true recall while matched random components do not; (iii) *dose*: the effect increases with strength and saturates; (iv) *shrinkage*: on seen entities the same intervention scales with $(1 - w)$ of Proposition 2. Because Φ and G are set by intervention, the controlled direct effects of the interventions actually made (a shift along one direction of Φ, or one latent of G held at its clean value) are identified without an ignorability assumption; reading them as the whole of X's effect on C additionally needs the graph's omission of any path from X to C that bypasses Φ (Observation 3b finds one in a single arm). Interventions rather than probes are the evidence. Both Gemma models meet (i)–(iii) (Table 1, §6, Figure 4b), and Gemma-2-2B meets (iv) with exposure measured by corpus counts, on a test of little power (the one-parameter test above), not by sitelinks. Condition (i) fails its bar in Qwen (0.58), (ii) holds only in part in Qwen and Llama (zero-ablation removes a quarter to a half of each ability and mean ablation little, Appendix I), and its norm-matched floor fails in Gemma-2-2B.

Observation 3a (one representation, two readers). We detect no effect of the gate on the tradition of what is answered. Moving refusal along the knowledge-refusal direction leaves the origin of the answers still given at 1.00, and holding the entity latent at its clean value leaves the patch's effect unchanged (a weak check, since that latent does not predict refusal). Setting Φ, in turn, moves G (§8, rule UA1 of Appendix I). $C \perp\!\!\!\perp G \mid \Phi$ is then a property of the assumed graph rather than a tested one. It fails if X reaches both C and G by paths that bypass Φ, and we find evidence for each

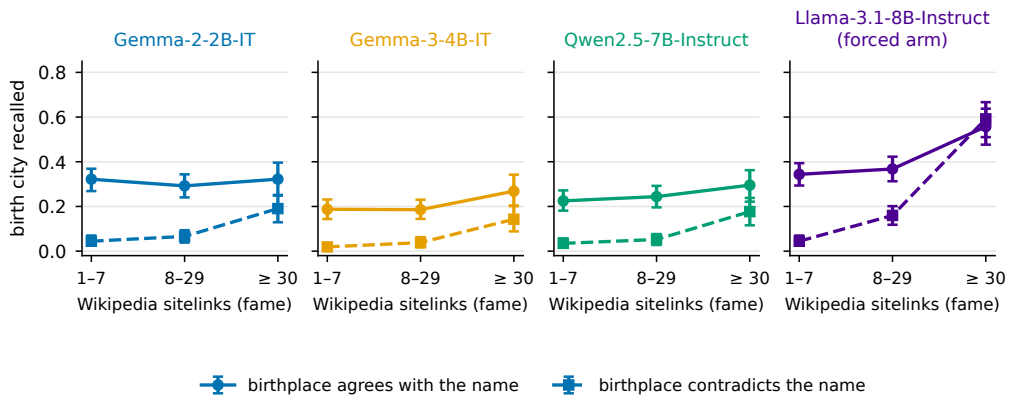


Figure 7: **The prior replaces the record for real people whose names mislead it.** Birth-city recall for real people whose birthplace contradicts their name (dashed) and fame-matched people whose birthplace agrees with it (solid), by Wikipedia-sitelink band, with 95 percent bootstrap intervals and cells under 15 omitted. The gap closes only where the model knows the person, as in Llama at 30 or more sitelinks, where the lines meet.

such path, in the Qwen arm of Observation 3b for C and in the gate's reading of tradition beyond the origin subspace (Appendix I) for G. "Two independent mechanisms" is therefore the wrong description. Both are children of Φ, which is why deleting Φ costs recall without buying abstention (§9) and why a mitigation must act on the $\Phi \to G$ edge.

Observation 3b (selection on the gate), and its test. Conditioning on Y = answered under a patch that raises refusal conditions on a collider of Φ's two effects, so the answers that survive could be the names the patch moved least. We tested this on the Qwen arm where rule UA3 failed (Appendix I), and it is not what happened. Every patched Baltic name, survivor or refused, is read as Tamil Nadu by the eleven-way decoder at block 10 (0.986 in both groups, Fisher $p = 0.71$). The representation moved completely. Of the 71 surviving answers, 39 name a university of no tradition, 27 still name a Baltic one, 2 a Tamil Nadu one and 3 one of another tradition. For the university question in Qwen, the content does not follow the representation the patch moved while the gate does. This is a limit of Observation 3a, and a later read of the name's own surface tokens is a candidate path the name-span patch does not touch.

G THE PRIOR AGAINST THE FACT: REAL PEOPLE WHOSE NAMES MISLEAD IT

This appendix gives the data, criteria and results behind §10 and Figure 7. The criteria were fixed before the data were fetched.

The people. From Wikidata, for each of the eight European traditions and three fame bands (Wikipedia sitelinks 1–7, 8–29, ≥ 30): *counter-stereotypical* people, whose one attested birthplace is a human settlement none of whose countries, historical ones included, is the tradition's home or a neighbour, and fame-matched *stereotypical* people born in the home country, 1,548 in all. Both groups pass the same battery: a two-token Latin-script name, one birthplace, a Latin-script place label, no duplicate across groups, and a name gate. The paper's invented-name classifier could not serve as that gate, because it places real names in their tradition 0.41 of the time, having memorised the curated pools. A classifier was therefore trained on 19,307 real people born in the home countries (character 2–4-grams, logistic regression, out-of-fold accuracy 0.817 over ten classes) and its threshold, 0.6, fixed by a rule stated before any person was gated (the smallest of four with out-of-fold precision at least 0.90; 0.928). The birthplace filter checks a place's *historical* countries as well as its current ones. A filter on current countries alone admits 90 people whose birthplace had a home country among its historical ones (Austria-Hungary towns now in Slovakia or Romania). As fixed in advance, four high-fame cells with fewer than 15 people are excluded from the rules, and 20 of 24 remain.

Reading. The paper's own prompts and labeller are used, and correct means the attested city appears in the answer, the rule of §9. Each wrong placed answer is scored twice more: in the *name's* tradition, and in the *record's* country (city resolved to country through Wikidata), the second being the rival hypothesis, that the model half-knows the person and lands near the truth.

	R1: birth city recalled			R2: where the wrong answers land		
model	counter	stereo	gap	n wrong	name's tradition	record's country
Gemma-2-2B	0.081	0.311	+0.229	233	0.95	0.02
Gemma-3-4B	0.051	0.203	+0.152	222	0.89	0.07
Qwen2.5-7B	0.069	0.246	+0.177	212	0.93	0.04
Llama-3.1-8B (forced)	0.196	0.395	+0.199	216	0.88	0.07

R1 passes in all four models (the rule asked for a gap of at least 0.10 in three). A person whose name points the wrong way is recalled at a quarter to a half of the rate of a matched person whose name points the right way ($n = 750$ and 760 in the 20 cells the rules use). **R2 passes**, with 88 to 95 percent of the wrong answers for the counter-stereotypical group in the name's tradition and 2 to 7 percent in the record's country. The prior does not merely blur the record. It replaces it, and almost as completely as it fills a blank, since 98 to 99 percent of the stereotypical group's placed answers sit in the name's tradition. It does so among people the model knows as well (Appendix J). **R3 fails, 2 of 4** (the rule asked for a gap at least 0.10 larger in the low band than the high, in three). The direction is right in every model and the margin is met in two (Gemma-2-2B $+0.15$ and Llama $+0.33$, against Gemma-3-4B $+0.04$ and Qwen $+0.07$). The cause is the rule's margin. Llama is the clean case. Below 8 sitelinks the prior wins by 0.30, and at 30 or more the gap is zero and the record wins. Abstention is reported beside the content, as everywhere in this paper, and it does not move. The two groups abstain at the same rate (birth city 0.009 against 0.018 in Gemma-2-2B, 0.041 against 0.030 in Qwen, under 0.005 in Gemma-3-4B). The gate does not know that the record contradicts the name.

R4 passes. The clamp of §9 (the origin projection removed at the name span, one direction at block 6, the configuration fixed before the run) was applied to the people each Gemma recalled correctly when clean. Recall retained falls 0.208 for the stereotypical group and 0.048 for the counter-stereotypical group in Gemma-2-2B ($n = 240$ and 62), and 0.291 against 0.105 in Gemma-3-4B ($n = 158$ and 38), differences of $+0.160$ and $+0.186$ against the rule's 0.05, positive in all eight configurations run. When the record agrees with the name, a fifth to a third of correct recall leans on the prior. When it contradicts the name, almost none does, because that recall was achieved despite the prior. The same prior supports the people who fit it and works against those who do not. Two caveats travel with the number. The quantity is retention among people recalled correctly when clean, so R4 shows a differential cost rather than a rescue, and the counter-stereotypical cells are small.

Is the gap a fact-frequency effect? Counter-stereotypical facts could simply be rarer in the corpus. Counting the Dolma documents that contain both the person's name and their attested city, and adding its logarithm to a logistic regression of correct recall on group, the counter-stereotypical coefficient keeps 0.94 to 0.97 of its size in all four models ($p < 10^{-13}$). Co-occurrence helps recall (+0.12 to +0.24 per log unit) and does not explain the gap. At equal exposure to the fact, the name's prior still wins.

H WHERE THE PRIOR COMES FROM: OLMO 2 ACROSS PRETRAINING AND POST-TRAINING

This appendix follows the mechanism through the training of a fully open model. Four criteria were fixed before any weights were loaded. The main study's code and configuration were run unchanged on ten checkpoints of OLMo-2-7B: six stage-1 pretraining checkpoints at log-spaced token counts (1B, 9B, 42B, 189B, 856B, 3,896B), the released base model (end of stage 2), and the SFT, DPO and Instruct releases. OLMo-2's stage-1 corpus is one of the three in which every one of our 1,206 names has a zero count, so the stage-1 checkpoints never saw any of them. The stage-2 anneal mix is not indexed, and the statement is a stage-1 one for the later models. Base checkpoints are asked with a completion prompt (`Question: ... Answer:`), fixed in advance. Their answers

are continuations, read by the forced arm's rule where the chat matcher finds nothing, and their abstention is never pooled with a chat model's. A second prompt, three-shot with one "Unknown" answer, tests whether zero abstention under the completion prompt reflects the prompt (§11).

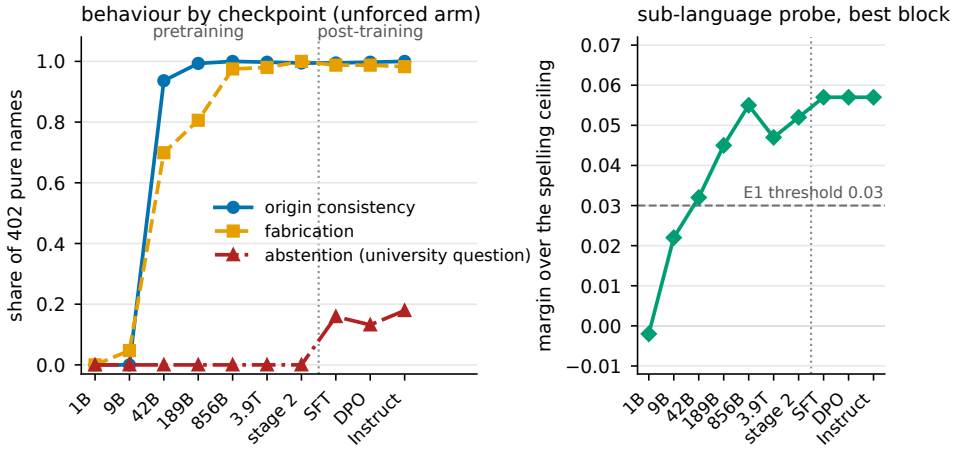


Figure 8: **The prior is learned in pretraining. Under a bare question, declining appears only after post-training.** OLMo-2-7B at six stage-1 checkpoints, the base model, and its SFT, DPO and Instruct stages, on the 402 certified names. Left: origin consistency, fabrication and abstention. Right: the sub-language probe's margin over its spelling ceiling, with the pre-registered 0.03 line. Representation and behaviour arrive together at 42B tokens.

The four rules. **E1, order of emergence: pass.** The first checkpoint whose probe clears its ceiling by $\geq$ 0.03 is 42B (+0.032). The first with fabrication $\geq$ 0.60 and consistency $\geq$ 0.85 is the same checkpoint (0.70, 0.94). The first with abstention above 0.10 is SFT (chat prompt, since no base checkpoint abstains under the completion prompt). **E2, content is a pretraining property: pass.** At each model's best qualifying point under the main study's guard, the origin patch makes the base model follow the donor tradition for 0.924 [0.879, 0.962] of held-out names and Instruct for 0.841 [0.773, 0.902], a difference of 0.083 against the allowed 0.10, with random directions at 0.000 in both. **E3, the gate is a post-training property: fail, on one clause by 0.021.** Under the completion prompt every pretraining checkpoint abstains on nothing (0.000), as does Instruct under that prompt. Under a few-shot prompt that allows "Unknown", first-stage checkpoints decline by tradition (§11). Under the chat prompt, Instruct's per-tradition χ^2 is $5.2\text{e}{-4}$, and its abstention is 0.179 [0.142, 0.216] against a threshold of 0.20 set from the main study's models before this model was run. **E4, the circuit is shared across stages: pass.** The top-ten heads by attribution of the fabricated city are shared 8 of 10 between the base model and Instruct on the same items. Within each model the real-against-invented overlap is 7 and 9 of 10, and ablating the shared heads cuts real recall (0.33→0.20; 0.32→0.16) and invented consistency (0.93→0.65; 0.96→0.66) together.

Results outside the criteria. The knowledge-refusal direction moves abstention both ways at SFT (0.37→0.53 pushed, 0.23 pulled, origin consistency 0.98) and not at DPO or Instruct, where a norm-matched random direction at the same site drops Instruct's abstention to 0.09 while the fitted direction leaves it near 0.32, and why is not understood. At 9B tokens the forced prefill already yields real cities (Vienna, Rome, Kandahar) of the wrong tradition, so the slot-filler exists before the prior does. A direction fitted from refusals on the base model, which has none, is undefined and degenerates every steered answer, so that stage is reported as not applicable.

Reading. The table below gives the few-shot abstention by checkpoint. Which group is declined more changes from checkpoint to checkpoint; at 856B tokens European names are declined far more than South Asian ones, the reverse of the chat models' pattern. Post-training leaves the content computation within a tenth of where it was (E2). Whether the annealing stage or post-training sets which traditions a chat model declines is not separated here.

checkpoint, three-shot prompt	birth city South Asian / European	χ^2 p	university South Asian / European	χ^2 p
9B	1.00 / 0.98	3e−5	0.00 / 0.00	-
42B	0.14 / 0.03	2e−4	0.68 / 0.00	8e−63
189B	0.52 / 0.61	4e−18	0.08 / 0.05	1e−7
856B	0.01 / 0.70	3e−47	0.03 / 0.15	2e−46
3,896B	0.48 / 0.33	4e−34	0.16 / 0.27	2e−28
released base	0.00 / 0.00	-	0.00 / 0.01	5e−3
Instruct, bare chat prompt (no few-shot)	0.00 / 0.00	-	0.00 / 0.00	-

At 9B the model copies the example. Abstention is by group, and χ^2 is over the eleven traditions. The 1B checkpoint's output is degenerate. The answers given stay in the name's tradition at every checkpoint from 42B (0.91 to 1.00).

What the gate could have been learned from. OLMo 2's SFT mix is public. A scan of its 939,344 examples for biographical questions about a named person (an explicit question form with the name in the subject slot, with a detector hand-checked on 100 names and 100 labels) finds 620 such questions, 258 with a name the paper's classifier places in one of its traditions, ten of them South Asian. The declining rate is 0.16 over the 161 Anglo-classified names and 0.00–0.29 over cells of 2 to 26. The rule fixed for this audit needed 30 examples per tradition and is met by one, so the reading is *inconclusive* as pre-registered. The scan does show that per-tradition refusal examples are almost absent, so the chat model's per-tradition pattern is unlikely to have been copied from them.

When the map forms in the weights. The OV readout of Appendix I was repeated at each stage-1 checkpoint on the heads that carry the map in the final model, with each checkpoint's own tradition directions. It passes for 0 of 11 traditions at 1B tokens, 6 at 9B, **9 at 42B**, 9 at 189B, 10 at 856B and at the end of stage 1 (own cities above every unrelated tradition's and above both random floors). The map is in the recall heads' weights from the checkpoint at which the tradition becomes decodable and the behaviour appears (the rule, by 42B or 189B, is met). The heads persist. The top ten by attribution overlap 8 of 10 between 856B and the end of stage 1, and 8 of 10 between the end of stage 1 and the final Instruct model (a rule fixed in advance, met), and the final model's three carriers are among the top ten by 9B (L19H5) and 189B (L22H3, L25H2). Pretraining builds the map, and post-training keeps it.

Is the prior shared across model families? The criteria were fixed in advance. Per tradition, we compared the distributions over placed fabricated cities of Gemma-2-2B, Gemma-3-4B, Qwen2.5-7B, Llama-3.1-8B (forced arm) and OLMo-2-7B (base) by Jensen–Shannon divergence, against a within-model reference (100 seeded half-splits) and the Dolma co-occurrence prior of Appendix D. Rule U1 asked the median cross-model divergence to be at most twice the within-model median in at least 8 of 11 traditions, and it holds in 5. Rule U2 asked every model to sit closer to the Dolma prior than to uniform in at least 8 of 11, and it holds in 5. Both fail. The prior is shared at the level of the tradition, which every model agrees on at 0.99, and model-specific in its shape within it. Gemma-3-4B in particular invents different Dutch, Portuguese and Welsh–Irish cities from every other model (Jensen–Shannon 0.5 to 1.0 from each of them, against 0.1 to 0.6 between the others). This is the same fact as the models' agreement on the tradition but not on the city (0.18) in §4 and the failure of the per-name corpus prediction in Appendix D, under a rule fixed in advance.

Is the prior sharper than the population it was learned from? The criterion was fixed in advance. The reference is the birth-city distribution of real Wikidata people selected by *name* (family-name items of the tradition's language for the European traditions, and the real-people classifier of Appendix G for the South Asian ones), born in the home countries, at two or more sitelinks (6,006 city rows). Rule M1 asked the share of fabrications on "the modal city" to exceed the share of real births there in at least 8 of 11 traditions in every model. The rule did not say whose modal city, and the two readings disagree. On the reference population's own modal city, the reading taken as the criterion of record, the pooled Nordic reference peaks at Copenhagen where every model says Oslo, so the fabricated share reads 0.000 and M1 fails (Gemma-2-2B 7, Gemma-3-4B 6, Qwen 9, Llama 9, OLMo 10 of 11). On the fabrications' modal city and the share of real births there, the reading clos-

est to the criterion's wording "the tradition's capital", M1 holds in 11 of 11 traditions in every model, with ratios such as 9.2 for Dutch names on Amsterdam and 4.9 for Finnish names on Helsinki. The criterion of record is not met. The second reading is reported beside it, and the criterion is entered in Table 2 as under-specified. Under either reading the prior is not population-matched, since it puts far more of its mass on one city than the people with such names do.

I ADDITIONAL CONTROLS, AND THE TRADITION-TO-CITY MAP IN THE RECALL HEADS

This appendix reports controls on the head-ablation, patching and gate results and locates the tradition-to-city map in the recall heads' weights.

Random-head floors matched on magnitude. Section 6 draws its floor uniformly over heads, which mostly picks heads that do little. We add two floors matched to the shared set over ten quantile bins of a per-head statistic, five seeds each: the absolute mean attribution to the produced city over the invented set, and the mean ℓ_2 norm of the head's output at the answer position. Drops are read relative to the clean rate, and the rule was a mean relative drop below 0.20 on both abilities in both Gemma models. Because the shared heads fill the top attribution decile in Gemma-2-2B, its attribution-matched floor is necessarily one set, the ten next-strongest heads (ranks 11–20). In Gemma-3-4B the five sets differ.

		Gemma-2-2B		Gemma-3-4B		
floor	heads	real recall	invented consistency	real recall	invented consistency	rule
shared set (§6)	10 / 9	0.99	0.93	0.94	0.69	≥ 0.50: met
shared set, mean ablation	10 / 9	0.63	0.77	0.88	0.63	
uniform random, 5 sets	10	0.01	0.05	0.02	0.03	
attribution-matched	10 / 9	0.07	-0.00	-0.08	-0.01	< 0.20: met
norm-matched, 5 sets	10 / 9	0.16	**0.26**	-0.14	0.01	< 0.20: not met

Relative drops in the two abilities under zero-ablation, and, for the shared set, under mean ablation (each head replaced by its mean output over the invented people, at the positions that predict the answer), which keeps the model in its natural activation range (Zhang & Nanda, 2024). Under mean ablation validity stays 0.91 / 0.87 and 1.00 / 1.00 in the two Gemmas, against 0.55 / 0.59 under zero ablation in Gemma-2-2B. The norm-matched floor fails its clause in Gemma-2-2B for the reason §6 already notes, that some random sets silence the model. Two of its five sets zero large-norm early heads, and the model stops answering (abstention 0.44 and 0.49). Among the answers that remain, 0.97 are on-tradition and recall equals clean. Under the shared set the remaining answers are on-tradition 0.12 of the time and recall among them is 0.008. Matched heads silence, and the shared heads redirect. The criterion should have conditioned on validity and did not, so it is recorded as not met.

The circuit in Qwen2.5-7B and Llama-3.1-8B. The same decomposition and arms were run on the two remaining main-study models, with the reconstruction check passing on the Llama layout at under one part in a thousand. Llama, which declines the free question, is read on the forced prefill for both sets. The top-ten overlap between real and invented people holds in both, **8 of 10** in Qwen and **9 of 10** in Llama, so the same heads write the city for real and invented people in all four models. Neither matched floor moves either ability in either model (relative drops -0.04 to 0.06). What differs is how much of the answer the shared set carries: zeroing it removes 0.53 of real recall and 0.37 of invented consistency in Qwen and 0.32 and 0.26 in Llama, with validity unchanged and the displaced answers landing in neighbouring geography (Berlin for a Dutch name, Vienna for a Hungarian one). The criterion of at least 0.50 on both, set from the Gemma results, fails in the two larger models, the damage falling with the head count (208, 272, 784, 1,024 heads). Under mean ablation the shared set removes 0.11 and 0.34 in Qwen and 0.11 and 0.00 in Llama, with validity 0.95 / 0.97 and 0.98 / 1.00, so in these two models the case rests on the overlap and the weights, not on ablation.

Where the prior is stored: the tradition-to-city map in the recall heads' weights. Appendix L shows that what encodes a name's tradition at the name tokens is not what writes the city at the answer. The connection is in the shared heads' own weights. For a head (ℓ, k) of §6 and a tradition

t, push the whitened origin direction as it stands in the residual entering block ℓ through the head's value and output projections (its OV circuit) and read the result through the unembedding, $\mathbf{z} = \mathbf{W}_U(\boldsymbol{\gamma}_{\text{final}} \odot \boldsymbol{\gamma}^{(\ell)}_{\text{post}} \odot \mathbf{W}_O^{(\ell,k)}\mathbf{W}_V^{(\ell,g)}(\boldsymbol{\gamma}^{(\ell)}_{\text{in}} \odot \hat{\mathbf{u}}_t^{(\ell-1)}))$, with every RMSNorm entering as its per-dimension weight, every scalar dropped, and the head's key-value group g resolved. No forward pass is involved. The joint readout sums over the shared heads. Two rules were fixed. The positive control requires that, for real people the model recalls correctly, the same readout of their own name-span representation ranks the attested city in the vocabulary's top ten for at least half of them, or the instrument has no standing. The claim requires that, for at least 8 of the 11 traditions, the tradition's own cities score above every unrelated tradition's cities and above the 95th percentile of two floors, ten random head sets and ten random directions.

model	positive control	random heads	traditions passing	carriers (traditions)
Gemma-2-2B-IT	43 / 71 = 0.61	0.00	11 of 11	L22H0 (9), L25H4 (9), L18H5 (8), L15H5 (6), L20H3 (4)
Gemma-3-4B-IT	26 / 50 = 0.52	0.00	11 of 11	L24H1 (11), L31H7 (5)
Qwen2.5-7B-Instruct	32 / 64 = 0.50	0.00	10 of 11	L23H4 (9), L23H6 (7), L22H19 (5)
Llama-3.1-8B-Instruct	50 / 102 = 0.49	0.00	10 of 11, *no standing*	L24H3 (10), L31H7 (6), L17H23 (4)
OLMo-2-7B-Instruct	30 / 70 = 0.43	0.00–0.14	10 of 11, *no standing*	L19H5 (9), L22H3 (9), L25H2 (9)

Figure 3 shows the readout for three traditions. The map is concentrated. Five of Gemma-2-2B's ten shared heads carry it, and the other five, found by attribution, do not read out geography on their own. In Gemma-3-4B one head carries it for every tradition. Two of the Gemma-2-2B carriers, L18H5 and L20H3, are the attribute-extraction heads Ferrando et al. (2025) name for that model. In OLMo-2-Instruct the same pattern appears for 10 of 11 traditions but the positive control recovers a known city for fewer than half of the people, so under the rule fixed in advance that reading is reported and not claimed. This is a weights-only statement with scalars dropped. It says what the heads would write from a pure tradition direction, and not how much they write in a forward pass, which the attribution and ablation of §6 measure. The causal test says the same. Recording the shared heads' attention patterns on the clean pass and holding them fixed while the origin patch of §7 is applied leaves the patch's logit shift intact (-0.11 [-0.29, $+0.08$] and -0.08 [-0.53, $+0.37$] of shifts of $+32.7$ and $+49.9$ logits, Gemma-2-2B and Gemma-3-4B, 132 items, both pre-registered clauses met), and giving the clean model the patched heads' patterns moves at most 0.55 logits. The heads attend to the name as they always did and carry a different value from it, so the prior rides their OV circuit.

The real-people patch by fame band. Table 1's 0.69 was measured on 108 people all in the highest fame band (30 or more Wikipedia sitelinks). Under Proposition 2 (Appendix F), people with weaker entity evidence should move more easily. The same configuration applied to the mid and low bands (108 people each, and repeating the high band reproduces 0.69 at every strength) gives follows-donor at strength 2 of **0.69** [0.60, 0.78], **0.78** [0.70, 0.85] and **0.81** [0.73, 0.88] for the high, mid and low bands, low minus high $+0.11$ [$+0.02$, $+0.21$], with validity 0.98, 0.92 and 0.97 and the clean model recalling the attested city for 0.35, 0.22 and 0.38 of each band, so the low band is not one the model never knew. Both pre-registered clauses (low at least high, and monotone within 0.05) are met. The high-band effect among people the clean model recalls, and the cost to occupation and birth year, are given in §7; in the mid and low bands too few such facts are known to read.

Removing the feed-forward writers. Section 5's decomposition says feed-forward blocks write most of the origin into the name tokens. Zeroing the feed-forward output of the early blocks that hold 80 percent of that contribution up to the patch block (blocks 4–6 in Gemma-2-2B), at the name span only, takes a sub-language probe at the patch block from 1.00 to 0.80, below the 0.92 string ceiling, while five random triples of blocks leave it at 0.98. The name is still echoed in every answer, the fabricated city stays on-tradition (0.93 to 0.89), and the late blocks' removal costs 0.06. The early writers carry the representation a probe reads at that depth. The answer does not depend on it, because later blocks rewrite it. In Gemma-3-4B the same rule names nine of the first thirteen blocks, and zeroing them destroys the name itself (echoed in 0.2 percent of answers, validity 0.43), so its collapse of consistency to 0.02 is corruption rather than removal and is reported as such.

What the unequal gate reads. Section 9 shows the refusal signal tracks the name's tradition without saying what produces it. There are two candidates, the tradition itself and how rare the name's parts are in pretraining. We regressed the refusal projection of 1,310 and 1,351 invented

names (Gemma-2-2B and Qwen, names outside the direction's fit half) on tradition, on the Dolma log-counts of the given name and surname, and on both. Tradition explains 0.12 and 0.10 of the variance, frequency 0.04 and 0.03, and tradition keeps three quarters of its share once frequency is in the model. Frequency explains none of the eleven tradition means. The causal read agrees. The origin shift of Table 1, applied to a Baltic name on the university prompt, carries its refusal projection 0.63 [0.52, 0.75] and 0.87 [0.66, 1.07] of the way to the Tamil Nadu level, against 0.12 and -0.02 for a random shift of the same norm. The reverse direction moves it less (0.48 and 0.09, random 0.38 and 0.24). Under the two readings fixed in advance the record is *demographic* in both models. The gate reads the name's tradition, from the same representation the content route uses; the rarity of the name's parts adds little.

The gate's behaviour follows the shift. A moved refusal projection is not yet a refusal. On the confirmation set's 220 Baltic and 78 Tamil Nadu names (none used to fit a direction), with Table 1's configuration on the university question, the forward shift raises Baltic abstention from 0.39 to 0.91 in Gemma-2-2B and from 0.17 to 0.63 in Qwen (1.95 and 0.95 of the clean gap A between Tamil Nadu and Baltic abstention, 0.26 and 0.49; McNemar $p = 8.3\text{e}{-32}$ and $5.4\text{e}{-26}$), against $-0.41A$ and $0.03A$ for a random shift of the same norm, so **UA1 is met in both**. The reverse shift lowers Tamil Nadu abstention by $0.73A$ and $0.32A$, but in Gemma-2-2B a random shift lowers it by $0.39A$, above the $0.25A$ allowed (**UA2 not met**). Surviving answers follow the donor in Gemma-2-2B (7/9, 28/28) and in Qwen's reverse arm (20/22). In Qwen's forward arm 2 of 32 do and 27 still name a Baltic university (**UA3 not met**), the one place where the gate moves and the content of what is still answered does not. Under the reverse shift Qwen also respells 16 of 78 Tamil names in Baltic form ("Tamīlsēvs Ramāsis attended Riga Technical University").

Fluency under the origin patch. At the reported configuration of Table 1 the degeneration rate is 0.030 against a clean 0.008 in Gemma-2-2B, 0.061 against 0.015 in Gemma-3-4B, 0.015 against 0.008 in Qwen and 0.000 in both arms of Llama. Validity is 0.87 to 1.00 against 0.94 to 1.00, and the name is echoed in 0.985 or more of answers. The pre-registered clause, degeneration within 0.02 of clean, is met in Qwen and Llama and not in the two Gemmas, where the patch costs two to five percent more degenerate generations. The second clause used an external model, Phi-3.5-mini. The per-token negative log-likelihood of each answer conditioned on its question rises to 1.46–2.48 times the clean arm's at the reported configuration, and to 1.18–1.25 times with the name removed from the context. The rule allowed 1.10, so it fails in every model. The answers stay fluent English ("Ruairí Pennant was born in Chennai, India"). What the external model penalises is a city it does not expect after that name, which is the prior under study, and, in 4 to 8 percent of patched answers in the three models where it occurs, against 1 to 2 percent clean, a name whose spelling has drifted under the shift ("Rauruí" for "Ruairí").

The map as a matrix. Pushing all eleven whitened tradition directions, each estimated on the residual stream entering the head's block, through the shared heads' value and output weights gives a $d \times 11$ matrix with singular values $s_1 \geq \cdots \geq s_{11}$. They decline gently over the first nine components and then collapse (Gemma-2-2B: 4.61 down to 1.90, then 0.74 and 0.07). The participation-ratio rank $(\sum_i s_i)^2 / \sum_i s_i^2$ is 8.85, 9.06, 8.91 and 9.05 of 11 in Gemma-2-2B, Gemma-3-4B, Qwen and Llama (6.83 to 7.44 in the eigenvalue form $(\sum_i s_i^2)^2 / \sum_i s_i^4$), and eight components hold 90 percent of the energy in every model. Read through the unembedding as an 11×11 matrix of tradition-direction against tradition-cities, the diagonal is the argmax in 10, 10, 9 and 9 of 11 rows and the own cities score above every unrelated tradition's in 11, 11, 10 and 10 of 11. The map is a low-rank bilinear lookup with about one column per tradition, the two related pairs sharing columns. Both rules fixed for it (rank ≥ 6, and diagonal dominant in ≥ 8 rows) are met in all four models, under either form of the rank.

An evidence estimator that failed its tests. Proposition 2 of Appendix F names the variable a gate should read, which is how much the model knows about the person rather than the name's tradition. The detector experiment (Appendix E) had a candidate, σ, the share of the name representation's squared norm (grand mean removed) that lies outside the tradition subspace, which separated real people from invented ones at 0.954. Eight tests were fixed in advance for it as an evidence estimator, on 935 real people across three fame bands, the 1,128 confirmation names and the 1,548 people of Appendix G. It fails the four that matter, in all four models: real against invented 0.84–0.88 (rule

0.90); tradition explains 0.26–0.44 of its variance among invented names (rule 0.05), with within-tradition separation as low as 0.74; its correlation with fame is 0.06–0.14; and among real people the model answers, higher σ goes with *wrong* answers (0.40–0.51, rule 0.70), at every block and with raw or whitened directions. On 61 real people whose full names occur in none of the three corpora, σ separates them from invented people (0.80–0.85) but hardly from famous ones (0.55–0.61), so it reads whether a name looks like a real person's name rather than whether the model has seen the person. A five-fold truthfulness probe separates real from invented at 0.95–0.98 but carries tradition too (R^2 0.14–0.43). Under the criterion fixed in advance, σ is not used as the evidence term of Appendix J.

The gate's own direction, pushed. Removing the tradition subspace at the gate's input leaves knowledge intact (0.99–1.00 of correct answers about famous people kept) but trims the tradition spread of refusal only by a third in Gemma-2-2B and a sixth in Qwen, short of the halving fixed for it. Nor does it change how often the answers about real people are right. Adding the knowledge-refusal direction of §8 at one strength for every prompt, the contrast arm of that experiment, was therefore tested across strengths under criteria fixed before that run. Refusal follows exposure in both models (low minus high 0.45 in Gemma-2-2B and 0.25 in Qwen against the 0.30 fixed, so that clause fails in Qwen; the per-band numbers are the uniform-push columns of the evidence-conditioned gate table in Appendix J), and the invented names still answered stay in their tradition. It is not fair across traditions. Gemma's university question becomes nearly flat (spread 0.38 to 0.16) while its birth-city question opens a new gap with Welsh–Irish names declined least, and Qwen's university spread narrows by a third with Baltic names declined least. The model's gate is exposure-graded, and its read of the name is not only in what the direction adds. Appendix J scales the same push by the name's corpus evidence.

J USES OF THE MECHANISM: METHODS AND RESULTS

This appendix gives the methods and full results behind §12 and the known-record test of §10. Every criterion was fixed before its data (Table 2). Models, snapshots, prompts, decoding and the labeller are those of the main study. Llama answers on its forced arm, as in the counterfactual test, and is marked.

Does the prior replace a known record? For the 1,548 people of Appendix G we asked two further questions, "What is the occupation of *Name*?" and "In what year was *Name* born?", scored against Wikidata (an occupation label, alias or listed synonym as a whole word, and a birth year within one year). A person is *known* (K) if either answer is correct. D_K is stereotypical (s) minus counter-stereotypical (c) birth-city recall inside K, with a bootstrap that resamples within group and fame band (10,000 draws). The criteria asked for $D_K \geq 0.05$ with the interval above zero in at least three models, and for at least 0.70 of the counter-stereotypical group's wrong placed answers inside K to name a city of the name's tradition.

model	$\lvert K\rvert$ s/c	recall in K s/c	D_K [95% CI]	in tradition (n)
Gemma-2-2B	236 / 277	0.347 / 0.134	$+0.214$ [$+0.143$, $+0.285$]	0.95 (59)
Gemma-3-4B	200 / 227	0.235 / 0.115	$+0.120$ [$+0.051$, $+0.191$]	0.86 (44)
Qwen2.5-7B	240 / 256	0.329 / 0.141	$+0.189$ [$+0.116$, $+0.261$]	1.00 (49)
Llama-3.1-8B (forced)	380 / 425	0.461 / 0.315	$+0.145$ [$+0.082$, $+0.209$]	0.87 (71)

Both criteria are met in all four models. Restricting to known people removes little of the gap, and the two groups' K are of similar size, so the comparison is between equally known people. K means that the model knows one other fact about the person, not its birthplace.

Forecasting fabrication from a name list. The tool uses only the name: the study's name classifier (ten classes, abstaining below probability 0.6) and, per model and class, tables built from the answers of §4 over the 1,206 names, giving the probability of an answer, the modal country of the invented birthplace and the modal city. An answered city is resolved to a country through Wikidata (P17). Forecasts were scored on the 1,128 confirmation names, on real people with one to seven sitelinks, and across models.

model	country hit, fresh names	name-blind	city hit	real people, wrong answers
Gemma-2-2B	0.704 [0.674, 0.734]	0.068	0.599	0.339
Gemma-3-4B	0.688 [0.660, 0.720]	0.069	0.398	0.351
Qwen2.5-7B	0.658 [0.627, 0.688]	0.072	0.462	0.366
Llama-3.1-8B (forced)	0.655 [0.624, 0.685]	0.069	0.440	0.356

Tables built from one model forecast another model's country at 0.59 to 0.69. The classifier covers 0.82 of the fresh names. Of the distinct answered cities, 530 resolved to a country through Wikidata, and 92 did not and are excluded. Traditions with one country are forecast at 0.96 to 0.98 (Dutch, Finnish, Hungarian, Tamil Nadu), traditions spanning several countries at 0.44 to 0.78, which is why the 0.85 criterion fails. At the level of the tradition the forecast hits 0.95 to 0.97. The probability of an answer is calibrated to within 0.05 per tradition on the birth-city question but not on the university question (0.11, 0.18). The tool does not forecast real people's wrong answers, many of which leave the name's tradition.

A name-twin detector, and selective answering. From the 1,548 people we drew 400, balanced over group and fame band. For each, four *twins* kept the real given name with a pool surname or the real surname with a pool given name, of the same tradition. Each passed the full certification battery (Appendix B), with up to three redraws. 1,487 twins were certified for 397 people. The twin score is the share of a person's twins whose answer names the same city as the person's. The baselines are the probability of the first token of the answered city and self-consistency over ten samples at temperature 1. The target is a wrong answer among asserted ones. A person and its twins were answered under identical settings in one pass.

model	asserted (wrong)	twin score	token probability	self-consistency	self-consistency, counter group
Gemma-2-2B	391 (312)	0.40 [0.34, 0.47]	0.67	0.66	0.72
Gemma-3-4B	395 (353)	0.38 [0.30, 0.46]	0.66	0.65	0.71
Qwen2.5-7B	381 (310)	0.42 [0.35, 0.49]	0.71	0.73	0.73
Llama-3.1-8B (forced)	395 (258)	0.49 [0.43, 0.54]	0.78	0.76	0.81

AUROC for detecting a wrong answer. All four criteria fail: self-consistency does not fail on the counter-stereotypical group, the twin score is below chance, it adds nothing to the token probability, and a selective-answering rule that uses it has the same risk at 80 percent coverage as one without it (ratio 1.00). Split by group, which was not fixed in advance, the twin score points the right way for counter-stereotypical people (0.52 to 0.60) and the wrong way for stereotypical ones (0.33 to 0.39): where the name fits the record, the correct city is the one the prior names, so agreeing with the prior marks a correct answer as often as an invented one.

An evidence-conditioned gate. In Gemma-2-2B and Qwen we added the knowledge-refusal direction of §8 at the gate block, scaled per prompt by $1-e(x)$ with $e(x) = \min\{1, \log_{10}(1+c(x))/3\}$ and $c(x)$ the Dolma count of the full name (zero for certified names), at the strength fixed by the dose study of Appendix I (2.0, 3.0). A placebo arm permuted e within tradition among real people. The sets and scoring are those of the gate experiments of Appendix I.

	Gemma-2-2B		Qwen2.5-7B	
	evidence	uniform push	evidence	uniform push
known famous answers kept	0.992	0.929	1.000	0.954
invented names declined (university)	0.918	0.918	0.764	0.764
tradition spread (clean 0.380; 0.498)	0.157	0.157	0.335	0.335
real people declined, high / mid / low	0.02 / 0.06 / 0.23	0.09 / 0.37 / 0.54	0.02 / 0.07 / 0.17	0.13 / 0.29 / 0.38
placebo, high / mid / low	0.02 / 0.09 / 0.12		0.03 / 0.07 / 0.09	

On invented names $e = 0$, so the evidence arm gives the uniform arm's responses exactly, and the fairness criterion re-tests the uniform push, as stated in advance, and it fails in both models. On real people the evidence term keeps almost every answer the model knows and declines the least-known people about twice as often as the placebo. The criterion comparing it with the placebo was read in the high band, where both arms are near ceiling, and fails there.

Real people born in Tamil Nadu or Sri Lanka. We asked the same questions about them, fame-matched by Wikipedia sitelinks to people born in Europe (§10). This analysis has no pass rule. In the least famous band, Gemma-2-2B declines the university question for 0.47 of them against 0.31

($p = 4\text{e}{-5}$) and Qwen for 0.52 against 0.38 ($p = 9\text{e}{-4}$), and both decline the birth-city question more often. Their birth-city recall is equal or higher in Gemma-3-4B and Qwen. In Gemma-2-2B it is lower on the pre-registered set (0.23 against 0.32) and equal only on a post hoc subset whose names the classifier places in the tradition (0.31 against 0.32), and Gemma-3-4B declines no one. Refusal and knowledge are measured on different attributes here, and university recall was not scored, so calibrated ignorance of their universities is not excluded. Sitelink matching may not equate exposure across regions.

What was not tested. Training OLMo 2 with declining answers balanced across traditions would test whether the unequal gate is a property of the post-training data; it requires fine-tuning a model under study and is left for future work, as are retrieval with conflicting documents, prompting mitigations, removal of the origin subspace at several layers, and whether attributes other than birthplace move together.

K Extended related work

Section 2 cites the work our design depends on directly; this appendix places the rest.

Name-derived demographic directions. Nguyen & Tan (2025) recover a race subspace from applicant names with 74–85% interchange accuracy, Karvonen & Marks (2025) show a whitened difference of means, fit on synthetic prompts, that transfers to realistic resume-screening prompts, Bouchaud & Ramaciotti (2025) steer gender and race directions activated by names, and Ahsan & Wallace (2026) publish labelled national-origin latents in the exact Gemma Scope release and layer we search. Every one of these papers moves a decision, a rating or a classification of the person's *own* group; none relocates a factual claim about a third party.

The nearest result, in another domain. In Mo et al. (2026), drug affixes are an engineered nomenclature that encodes class by design, the patch restores an affix rather than substituting a donor, and the direction moves the affix-consistent answer and the "non-existent drug" choice together. Person names have no designed nomenclature, but they carry origin in their spelling, since a character classifier reads the tradition at 0.993 (Appendix N). The mechanism is of the same kind, a form-level cue read as a class; what differs is that the class selects a fact about a third party.

Methodological instruments. The character-n-gram baseline refit on the probe's own folds, the paired bootstrap with an exact McNemar test, the token-embedding internal baseline and the logit-preference patching criterion are from Chelombitko et al. (2026). Patching conventions follow Zhang & Nanda (2024), and donor-swap designs judged against criteria fixed before the run have been used to separate what a residual band carries from what it routes (Thavarasa & Sukumar, 2026). For the removal of the origin projection in §9 we use the redistribution metric of Shan & Mueller (2026). In Gemma-2-2B the answers that change move to vaguer places (0.114 against a same-tradition null of 0.024), not to refusals.

Names as cues, without a fabricated fact. Salinas et al. (2024) show that a name alone shifts model advice and valuations along race and gender lines. Dasanaike (2026) recovers ethnicity from names above Bayesian surname geocoding, and Inoshita (2026) predicts nationality from names through what he describes as associative memory rather than reasoning, the alternative our toponym control (§4) tests. Hu et al. (2026) locate race-associated neurons, one of which captures Chinese surnames, while arguing that the representation is distributed. In all of these the dependent variable is a judgement about the named person's own group. Ahsan et al. (2025) come closest to our intervention, since patching one layer flips an invented patient's gender, but the cue is an explicit demographic sentence and the invented attribute is the demographic itself.

Knowledge storage, exposure and name-origin classifiers. Allen-Zhu & Li (2024; 2025) train models on synthetic biographies and show that a person's attributes are extractable only when the training text presents that person's knowledge in varied forms, and that stored attributes are easy to retrieve but hard to manipulate. Their individuals are seen in training, and ours are not. Kandpal et al. (2023) and Mallen et al. (2023) show that factual accuracy follows how often an entity or fact

appears in pretraining. The exposure weight of Proposition 2 is the same variable, and our fame bands and corpus counts reproduce that dependence for birth cities. Name-based nationality and ethnicity classifiers, such as NamePrism (Ye et al., 2017) and ethnicolr (Sood & Laohaprapanon, 2018), show that origin is readable from the name string alone, which is why every decoding result here is compared with a string classifier.

L What "the origin direction" does and does not name

The feed-forward share. *Where* the origin is written is a division of labour. Projecting each block's sublayer outputs at a position i onto that block's direction for the name's own tradition and accumulating over blocks, the feed-forward share f_i is

$$f_i \;=\; \frac{\sum_\ell \langle \mathbf{m}_i^{(\ell)}, \hat{\mathbf{u}}_t^{(\ell)} \rangle}{\sum_\ell \langle \mathbf{a}_i^{(\ell)}, \hat{\mathbf{u}}_t^{(\ell)} \rangle + \sum_\ell \langle \mathbf{m}_i^{(\ell)}, \hat{\mathbf{u}}_t^{(\ell)} \rangle}, \tag{2}$$

each sum averaged over names (they share a sign at every position reported). **At the name's own last token** it is 0.84 (Llama), 0.82 (Gemma-2-2B) and 0.81 (Gemma-3-4B), and 0.13, 0.34 and 0.55 at the final prompt token, where attention dominates. Feed-forward layers write the origin into the name's own representation, the stored-prior shape.

Two constructions of the direction. Two natural constructions of "the origin direction" do not coincide. A residual difference-of-means direction at the name tokens and a weights-only construction from the candidate cities' unembedding rows agree at cosine -0.005 to 0.153 across models and traditions, against chance floors of 0.016 to 0.021. What encodes "this name is Latvian" at the name tokens is therefore not what writes "Riga" at the answer position, and the phrase needs a qualifier every time it is used. What connects the two is the shared heads' own output map, which reads the name-token direction out as the tradition's demonym, country and cities (Appendix I, Figure 3).

M The second labeller, and which way its errors run

A second, model-based judge, Phi-3.5-mini-instruct, relabelled 600 stratified answers. Agreement is $\kappa = 0.95$ on assert-versus-not, $\kappa = 0.98$ on the origin tradition (405 answers both call assertions) and $\kappa = 0.97$ on origin-consistency, but $\kappa = 0.73$ [0.68, 0.77] (95% CI) on the three-way assert/decline/other label, against the 0.80 fixed in advance. The shortfall is structural, because a pattern matcher with an unmatched-residual bucket cannot reach 0.80 against a judge with no such bucket. Of the 69 answers our matcher calls OTHER, the judge calls 64 declines. Where the matcher commits to a label, it agrees with the judge 524 times in 531.

We hand-read all 84 disagreements and recorded a verdict on each. The judge is right in 73 (64 declines our phrasing list does not cover, 5 assertions in unusual syntax, 3 answers about a name the model altered, 1 asserting a value while calling it fictional), the matcher in 3, and in 8 both labels agree and only the origin mapping differs. **The errors therefore run in the conservative direction.** Our labeller under-detects refusal and does not mislabel assertions, so every abstention rate in the paper is a floor and every fabrication rate is under-counted, by at most 5 in 600. Every abstention result is additionally reported under both codings of the ambiguous bucket, and no verdict changes between them.

N The lexical ceiling

This appendix shows why every decoding accuracy in the paper is compared with a string-only baseline, which is that the spelling of a name alone predicts its tradition almost perfectly. The sub-language task splits the three pooled traditions (Baltic into Estonian, Latvian and Lithuanian; Welsh–Irish and Slovene–Croatian into two each). A name's label is the sub-language its given name and surname share. Where the two components differ, or the tradition is not split, the label is the tradition itself. Seventeen labels occur among the 402 names.

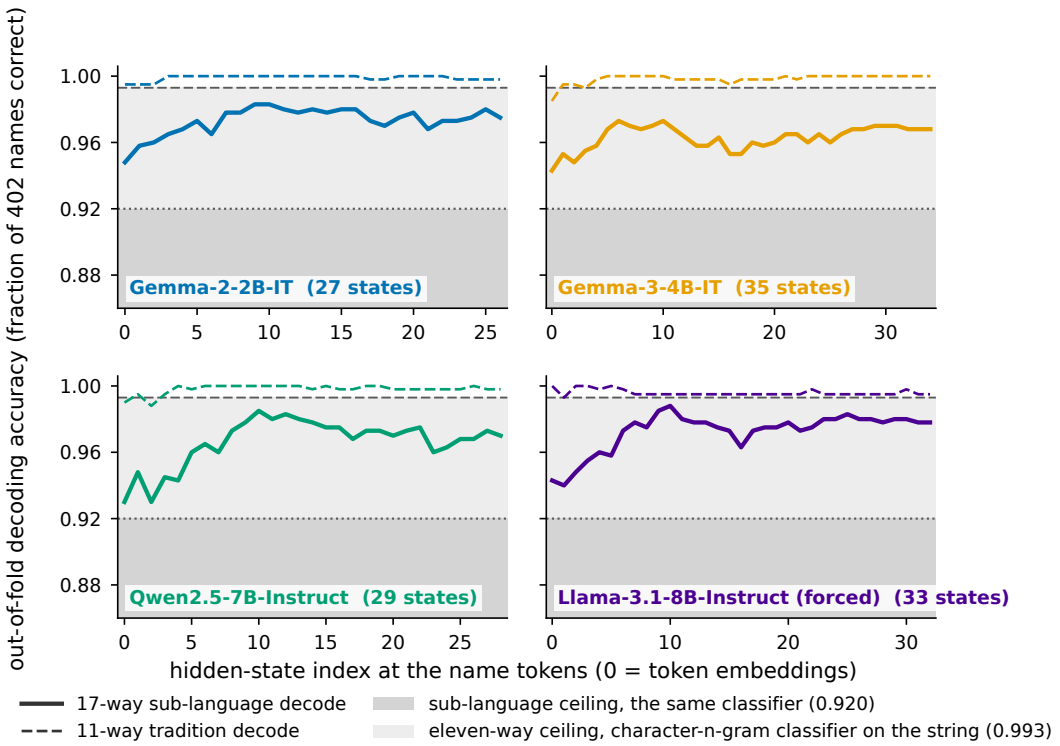


Figure 9: **A decode is only evidence above its lexical ceiling.** Out-of-fold accuracy of a linear probe on the residual stream at the name tokens, at each hidden state (0 is the token embeddings), for the eleven-way tradition (dashed) and the seventeen-way sub-language (solid). Grey bands: the accuracy of a character-n-gram classifier on the name string alone, refit on the probe's folds. The eleven-way curve rides its ceiling from the token embeddings on and is not reported. The sub-language curve clears its ceiling by $+0.05$ to $+0.07$ in every model, which is the decoding claim.