WHAT'S IN A PLACE NAME? SPELLING AS A CONFOUNDER IN COORDINATE PROBES OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in language models have enabled emerging applications in geospatial fields, involving tasks ranging from geocoding to spatial question answering. These applications have raised the question of whether, and to what degree, language models encode spatial information about real-world entities. Language model probes are increasingly taken as evidence of learned representations, but the extent to which spatial information is available from the name alone has not been systematically measured. We show through a series of experiments that probes trained to predict coordinates from place names recover a large share of their signal from the name alone. This share rises with entity granularity, from under a tenth at metros to above two fifths at venues, in the same order across model families and does not diminish as model parameters scale, revealing a confounder in probe-based claims of learned spatial representations.

1 INTRODUCTION

Advances in language models have given rise to a wide variety of applications in geospatial domains, and probes of their activations have been read as evidence that the models hold basic ingredients of a more comprehensive world model(!) (Gurnee & Tegmark, 2024). Models are prompted to geocode place names and addresses (Hu et al., 2024; Masis & O'Connor, 2026), to answer spatial questions about distance, direction and containment (Roberts et al., 2023; Truong et al., 2026; Abayomi et al., 2026), to estimate the population and wealth of a location from its coordinates (Manvi et al., 2024b), and to plan routes through a city (Vafa et al., 2024), and their answers are audited for geographic bias (Manvi et al., 2024a; Decoupes et al., 2025).

These applications have raised the question of whether, and to what degree, a language model encodes spatial information about named entities. Three instruments have majorly been explored to answer this question: prompting the model and scoring the answer (Roberts et al., 2023; Manvi et al., 2024a), comparing the geometry of the model's representations of place names with real geography (Liétard et al., 2021), and fitting a linear probe from a model's activations to an attribute of a named entity, reading the fit as evidence of the model's learned representation. In our work, we tackle the question of how much of a probe's fit is available from the name alone.

Such probes have been fitted to the coordinates of places (Gurnee & Tegmark, 2024; Godey et al., 2024), the birth and death years of people (Heinzerling & Inui, 2024), the economic statistics of counties (Buckmann et al., 2025), and the colour a word names and the state of a board (Abdou et al., 2021; Li et al., 2023). In each case the named entity containing suffixes, character patterns, and etymology of settlement names tracks the language of a place, roughly marking the region it lies in. A probe can therefore recover part of the attribute from the name alone. Comparisons reported control for the difficulty of the target, the capacity of the probe, and the architecture it is drawn from (Gurnee & Tegmark, 2024; Hewitt & Liang, 2019; Conneau et al., 2018; Li et al., 2023). However, no probing study of an entity's attribute has reported a baseline that predicts the target from the entity's string.

To address this gap, we fit the same probe to a second copy of each model, built from the model's own configuration file with the same depth, the same width and the same tokenizer, whose weights are drawn at random. We refer to this second copy as the untrained control. The representations of a

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.

trained model can carry the way place names co-occur in text (Liétard et al., 2021; Barenholtz, 2026) and other learned features, which makes the untrained control a more conservative baseline.

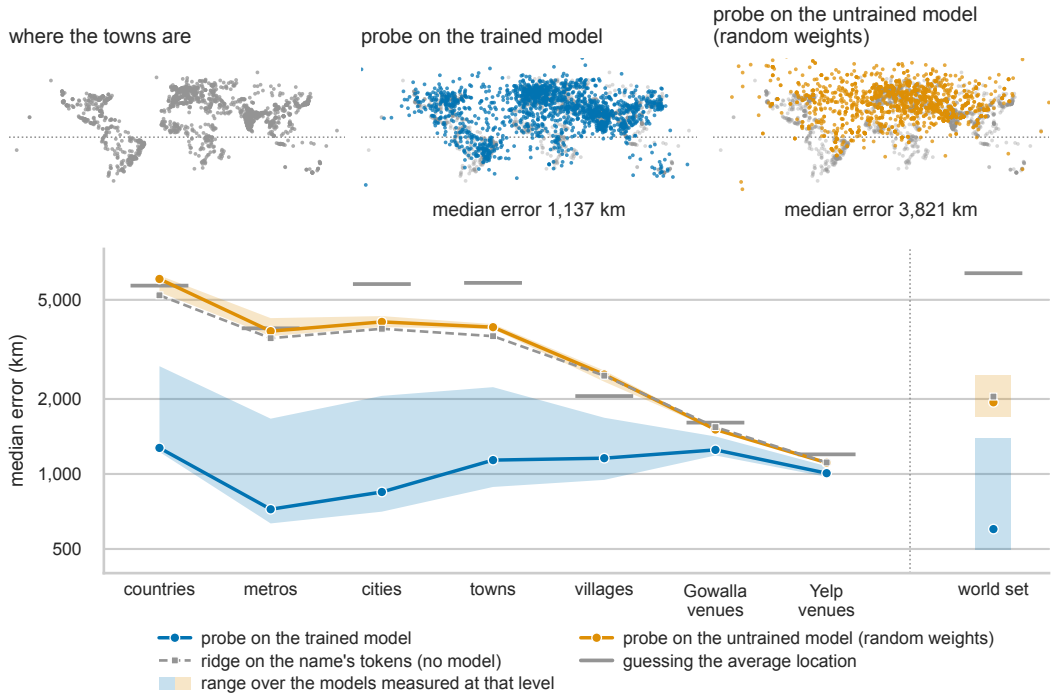

Figure 1: Top: the 1,600 towns of one test split on the Llama-3.2-1B. Left, their true locations on a map. Middle, where a linear probe fitted to the trained model places them, median error 1,137 km. Right, where the same probe fitted to a copy of the model with random weights places them, 3,821 km. Guessing the average location of all towns misses by 5,847 km. Bottom: the median error of the same two probes (median over all models) at each of the seven levels of place, measuring the untrained copy's share of the trained probe's improvement.

We also define the spelling share of a probe: the fraction of its improvement over an average-location baseline that the probe of an untrained copy also reaches (§ 3.3). We measure it with two predictors of the name, the untrained copy and a ridge regression on the name's token counts (Table 2). We first show that on 36,795 places at seven levels of granularity, probed on seven models from 0.41B to 13B parameters, the share is a property of the level: under a tenth at metros, a fifth to two fifths at towns and above two fifths at venues, in that order on every model (Fig. 1, § 4.2). We then show that on the 39,585-place world set of Gurnee & Tegmark (2024), on models to 70B parameters, the share does not fall with model size under the untrained copy and falls by 0.04 under the token ridge (§ 4.3), and that the two predictors of the name give the same share at both venue levels (§ 4.2). We then locate the channel at venues: replacing the place name inside a venue name with a made-up word removes the advantage of the names that contain one, and substituting another place name moves the prediction toward it (§ 4.4). Finally, we show that the same measurement applies to a scalar attribute, the death year of 37,539 historical figures, where the name gives at least a quarter of the probe's gain (§ 5.2), and we recommend the token ridge as the baseline for any probe of an attribute a name predicts, since it requires no forward pass of the model and does not grow with its width (§ 5.1).

2 Related Work

Baselines for probes. The untrained encoder of Conneau et al. (2018) forms the basis for our paper's control. They probed sentence embeddings beside the same encoder with random weights and beside a Naive Bayes classifier on the words of the sentence, and both baselines recovered most of the trained encoder's accuracy on the tasks that depend on the words. Zhang & Bowman (2018)

and Wieting & Kiela (2019) found the same with frozen untrained encoders. Probes of a board game have been reported beside a randomly initialised copy of the model (Li et al., 2023; Nanda et al., 2023; Karvonen, 2024), a coordinate probe on a small network beside a probe on an untrained one (Altmeyer et al., 2024), and automated interpretability scores have failed to separate trained from random transformers (Heap et al., 2025). The control task of Hewitt & Liang (2019) keeps the representation and gives each word type a random label fixed across its occurrences. Static word vectors have served as a baseline for the coordinates of cities (Liétard et al., 2021) and for colour (Abdou et al., 2021), and co-occurrence statistics alone recover the coordinates of cities (Barenholtz, 2026).

Probing geography. Gurnee & Tegmark (2024) report the coordinate probe's R^2, which is zero at the mean predictor, a proximity error whose value under random guessing is 0.5, and a prompt of ten random tokens as a baseline. Godey et al. (2024) fit a probe to models from 14M to 70B parameters and report R^2. De Sabbata et al. (2025) probe places in the United Kingdom, Italy and four states of the United States on one model but lack a baseline, and Decoupes & Guille (2026) probe 100 French municipalities on 29 models from 0.5B to 72B parameters. Heinzerling & Inui (2024) probe the birth and death years, population and coordinates of people and places beside shuffled labels and random representations, the closest of these studies to a control. Some studies measures a centroid baseline on a sample of GeoNames cities (one we also partially leverage), 6,145 km, and shows that a prompt supplying the parent place raises the probe above the bare name (Neira, 2026), a finding we confirm. A few studies hint at place names carrying regional cues or trace the error to name ambiguity (De Sabbata et al., 2025; Decoupes & Guille, 2026; Neira, 2026). None measures how much of the probe's fit those cues account for (Appendix A).

Evaluating geography by prompting. A prompted model answers a question about a named place, and evaluations cover coordinates, distance, direction, containment and geographic bias (Roberts et al., 2023; Abayomi et al., 2026; Decoupes et al., 2025; Masis & O'Connor, 2026; Truong et al., 2026; Manvi et al., 2024b;a; Hu et al., 2024). Prompt ablations on an address show that the state and country strings account for much of the signal (Manvi et al., 2024b). Half of the evaluations tabulated in Appendix A lack any control on the surface form of the name. One reports an untrained model of the same architecture, on route sequences that contain no place name (Vafa et al., 2024).

3 Setup

3.1 Problem Setting

We consider the problem of predicting the coordinates of a named place from a language model's activations, and of measuring how much of the probe's accuracy the name contributes. The error of

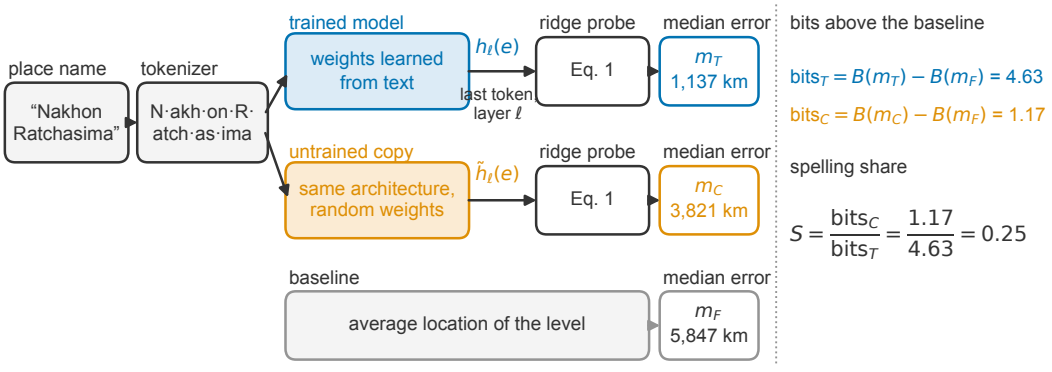

Figure 2: The measurement pipeline at towns on Llama-3.2-1B. A place name goes through the trained model and through an untrained copy as the ridge (Equation 1) is fitted at the last token of the name, and each probe's median held-out error is converted to bits above the average-location baseline. The spelling share is the ratio of the two.

a prediction is its great-circle distance d_g in kilometres to the true coordinate on a sphere of radius $R = 6{,}371$ km. True coordinates come from GeoNames (GeoNames, 2026) and from the Gowalla and Yelp datasets (Cho et al., 2011; Yelp Inc., 2024), and the death year of historical figures comes from DBpedia (Gurnee & Tegmark, 2024). The seven levels of place granularity are listed in Table 1, and Figure 2 shows the measurement of one cell.

Table 1: Selection rules for the seven levels of places. The baseline is the median held-out error of guessing the average location of the set, in kilometres. On the death-year set it is the constant predictor at the training median, in years.

Level	Source	Population band or selection rule	Entities	Average-location baseline (km)
countries	GeoNames	every country, the capital as its coordinate	241	5,700
metros	GeoNames	population of at least one million	554	3,843
cities	GeoNames	100,000 to one million	5,000	5,783
towns	GeoNames	15,000 to 100,000	8,000	5,847
villages	GeoNames	500 to 15,000	8,000	2,054
Gowalla venues	Gowalla and Overture	check-in venues matched to an Overture place within 25 m, sampled evenly over ten bands of check-in count	10,000	1,607
Yelp venues	Yelp Open dataset	5,000 drawn at a fixed seed from the 15,000 most reviewed businesses with a unique name	5,000	1,199
world set	Gurnee & Tegmark (2024)	DBpedia places whose Wikipedia page had at least 5,000 views over three years	39,585	6,394
death-year set	Gurnee & Tegmark (2024)	historical figures, the death year as the attribute	37,539	270 years

3.2 Probe and Baselines

We use a ridge regression to fit a linear probe to the model's activations. For an entity e with input vector $x(e)$ and target $y(e)$, the weights and intercept are

$$\hat{W}, \hat{b} = \arg\min_{W,b} \sum_{e \in \text{train}} \left\| y(e) - Wx(e) - b \right\|_2^2 + \alpha \|W\|_F^2, \qquad \hat{u}(e) = \frac{\hat{W}x(e) + \hat{b}}{\|\hat{W}x(e) + \hat{b}\|_2}, \quad (1)$$

with α chosen on the validation split and the prediction $\hat{u}(e)$ scaled to the sphere. The target is the true coordinate as a unit vector, $y(e) = u(e)$, on every row of Table 2 except the shuffled-label baseline whose coordinates are permuted across the training places.

The models are Pythia-410m, 1B and 1.4B (Biderman et al., 2023), trained on the Pile (Gao et al., 2020), and Llama-3.2-1B and 3B (Meta AI, 2024), Llama-2-7B and 13B, and Llama-3.1-8B and 70B. Seven models are measured across seven place levels and eight on the world set (Appendix C). A cell is one level on one model, and each cell has the trained probe and the control. The probe's input $x(e) = h_\ell(e)$ is the residual stream at the last token of the name at layer ℓ, with layer 0 the embedding output. Each coordinate of $x(e)$ is standardised on the training split and the target is centred on its training mean which is added back to every prediction before scaled to the sphere. Five seeds each split a level into 70/10/20 training, validation, and test sets and $\alpha \in \{0.1, 1, 10, 100, 1000\}$ is chosen per layer and seed on the validation split. The layer ℓ^* of each probe is one whose validation median over the five splits is lowest and every reported median pools the held-out errors per entity, the median over the splits in which the entity was held out, then the median over entities.

Table 2 is an illustration for how Llama-3.2-1B is probed at towns; the remaining runs are detailed in the Appendix B.

Table 2: Every row is the ridge of Equation 1 with the same penalty grid, layer rule, five splits and held-out places. The last column is the median held-out error at towns on Llama-3.2-1B (the cell of Figure 1)..

Predictor	Features $x(e)$	Model weights	Median error at towns (km)
average-location baseline	the mean of u over the training places	none	5,847
shuffled-label baseline	$h_\ell(e)$, trained model, coordinates permuted	trained	6,020
untrained control	$\tilde{h}_\ell(e)$, copy with random weights	random	3,821
ridge on the name's tokens	$c(e)$, token counts of the name	none	3,573
ridge on character n-grams	$g(e)$, hashed 2- to 5-gram counts	none, no tokenizer	4,161
embedding baseline	$h_0(e)$, trained embedding table	trained, layer 0	3,392
trained probe	$h_\ell(e)$, trained model	trained	1,137

Control. The untrained copy has the model's architecture built from its configuration file with every weight drawn independently from a normal distribution with standard deviation 0.02. It shares the tokenizer, the probe, the layer rule and the splits with the trained model, so $\tilde{h}_\ell(e)$ is a fixed function of the name's tokens.

3.3 Spelling share

The spelling share is the fraction of a probe's gain over an average-location baseline that the untrained control reaches, so a share of 0.25 means the name accounts for a quarter of the improvement. We measure this improvements in bits: a prediction with error r has narrowed the answer to a cap of radius r covering $(1 - \cos(r/R))/2$ of the sphere and one bit is one halving of that area.

Definition 1 (Spelling share)**.** Let m_F, m_T and m_C be the median held-out errors of the average-location baseline of the trained probe and of a predictor with access to the entity's string only. Each is read at its own L^* on the same pooled test entities. Let $B(r) = -\log_2\big((1 - \cos(r/R))/2\big)$ be the bits in a spherical cap of radius r as a fraction of the sphere. The bits above the baseline are $\mathrm{bits}_X = B(m_X) - B(m_F)$, and the spelling share and the margin are

$$S = \frac{\mathrm{bits}_C}{\mathrm{bits}_T}, \qquad M = \mathrm{bits}_T - \mathrm{bits}_C. \tag{2}$$

For radii well under the sphere, S is the control's position between the baseline and the trained probe on a log axis of error, $S \approx \log(m_F/m_C)/\log(m_F/m_T)$. A control placed at random on that axis has an expected share of a half, and a simulation with the exact $B(r)$ gives 0.494, which we call the random-placement expectation. A control with no signal has a share of 0.

4 Results

4.1 World set

Finding: The results on the 39,585-place world set show that the name alone gives about half of the trained probe's gain on the Llama models and more than half on the Pythia models. Specifically, the share is 0.61 and 0.68 on the two Pythia models and 0.46 to 0.52 on the six Llama models. On the six models common across the world set and the seven levels, the share is above every settlement level and above the Yelp venue level on four of the six.

4.2 venues to countries

Finding: At towns the name alone gives a fifth to two fifths of the trained probe's gain and at metros under a tenth. Specifically, at towns the control recovers 1.09 to 1.17 bits above the average-location baseline on all seven models, a spread of 0.09 bits over two families and seven sizes. Its share of the trained probe's bits is 0.40 and 0.38 on the two Pythia models and 0.21 to 0.25 on the five Llama

models. At metros the share is under 0.07 (Figure 5, Appendix I). At 14 of the 49 cells it is undefined due to control not beating the average-location baseline (villages on all seven models, countries on five and metros on two). Our belief is that the spelling channel is a coarse regional prior. At villages the level is concentrated in Europe, so guessing the average location already places a name in the right part of the world, and at countries and metros the names are few and famous. Future work should explore a level of small places spread evenly across regions to separate concentration from size (§ 6).

Venues and the toponym split. At both venue levels the trained probe recovers under 1 bit. On the 5,135 held-out Gowalla names without a toponym the trained probe stays within 0.45 bits of the control and at the Yelp level within 0.26 bits, while on the 1,591 Gowalla names that contain a toponym the margin runs from 0.61 to 1.58 bits. Over the entire Gowalla level the ridge on the tokens of the name lies within 0.09 bits of the control so the two predictors of the name agree where the margin is smallest (Appendices E and G).

4.3 Model scaling

Finding: The untrained control on the world set shows that the share does not fall across scaling Llama models (Table 3). Under the ridge on the name's token counts it falls by 0.04. The control's width accounts for part of its movement and Section 6 gives that mechanism. The share is 0.46, 0.48, 0.50, 0.50, 0.49 and 0.52 on the six Llama models from 1.2B to 70.55B scaling by a factor of 57 in parameters. The trend fitted to the six control points is a rise of 0.034 per decade of parameters, a factor of ten, with an interval of [0.025, 0.044]. At towns the share is 0.21 to 0.25 over the five models measured. At the Yelp venue level the 70.55B model has 0.26 [0.21, 0.34] against 0.42 at 8B. The fall of 0.04 holds on all three predictors of the name (Appendix C).

Spelling share predictability. From the five Llama models at and below 13B parameters, we noted spelling shares lying in a band of 0.04 with a fitted rise of 0.040 per decade. This let us predict a share near 0.50 at 70.55B which the run gives 0.522 [0.508, 0.534]. Since the control reads only the last token of the name, an exact lookup on that token caps its gain at 3.90 bits and the share must fall once the trained probe passes about 410 km on the world set. At 70.55B the trained probe reads 522 km and the control sits at 95 percent of that ceiling (Appendix J). The ceiling is a property of the level and the tokenizer set by how many distinct last tokens its names end in and how close together the places ending in the same token lie.

Table 3: The spelling share against model size at every level under the untrained control and on the world set. A dash marks a cell that was not run or has no defined share (Appendix B).

	Pythia			Llama					
Level	410m	1B	1.4B	3.2-1B	3.2-3B	2-7B	3.1-8B	2-13B	3.1-70B
countries	—	—	—	—	—	—	—	—	—
metros	0.06	—	—	0.04	—	0.02	0.03	0.02	—
cities	0.28	—	0.28	0.16	0.17	0.19	0.16	0.20	—
towns	0.40	—	0.38	0.23	0.22	0.25	0.21	0.24	—
villages	—	—	—	—	—	—	—	—	—
Gowalla venues	0.50	—	0.41	0.27	0.21	0.25	0.22	0.22	—
Yelp venues	0.76	—	0.56	0.55	0.44	0.42	0.42	0.40	0.26
world set	0.61	0.68	—	0.46	0.48	0.50	0.50	0.49	0.52
world set, token ridge	0.74	0.70	—	0.49	0.46	0.48	0.44	0.47	0.45

4.4 Toponyms

Finding: The intervention shows that at venues, the name's part of the probe's gain comes from the toponym inside the venue name. Replacing the toponym in a flagged name with a made-up word

of the same character length removes 93 to 112 percent of the advantage of those names over the names without one. Substituting a different real toponym removes it as well and the prediction moves toward the substituted place (Table 4).

For a venue level let $\mathrm{bits}_{\mathrm{top}}$ be the trained probe's bits on the held-out names that contain a toponym, $\mathrm{bits}_{\mathrm{none}}$ its bits on the names that contain none, and $\mathrm{bits}_{\mathrm{edit}}$ its bits on the flagged names after the toponym is replaced without refitting the probe. The fraction of the advantage that an edit removes is

$$\lambda = \frac{\mathrm{bits}_{\mathrm{top}} - \mathrm{bits}_{\mathrm{edit}}}{\mathrm{bits}_{\mathrm{top}} - \mathrm{bits}_{\mathrm{none}}}, \tag{3}$$

so $\lambda = 1$ returns the flagged names to the error of the names without a toponym and $\lambda = 0$ leaves them where they were. λ_{abl} is this fraction for the made-up word and λ_{swap} for the substituted toponym.

Table 4: The toponym ablation on venues across Pythia-410m and Llama-3.2-1B.

		median error (km)					
Cell	n	original	made-up word	swap	λ_{abl}	λ_{swap}	displacement
Pythia-410m, Gowalla	1,591	1,142	1,501	1,476	1.12 [0.99, 1.32]	1.05 [0.91, 1.26]	0.12 [0.10, 0.13]
Pythia-410m, Yelp	300	618	1,092	1,167	0.93 [0.85, 1.15]	1.04 [0.95, 1.25]	0.10 [0.07, 0.17]
Llama-3.2-1B, Gowalla	1,591	850	1,443	1,528	1.10 [1.03, 1.21]	1.22 [1.13, 1.33]	0.34 [0.31, 0.36]
Llama-3.2-1B, Yelp	300	439	1,105	1,275	1.00 [0.91, 1.17]	1.16 [1.10, 1.27]	0.30 [0.22, 0.37]

Both edits act on held-out names and leave the probe as it was fitted, so the string is the only thing that changes. However, both edits can change the token count of the name and the position of the last token that the probe reads. The substituted toponym changes the length as well and the prediction moves toward the substituted place by 0.10 to 0.34 of the distance to it. Table 4 gives every condition of the ablation with its interval, and Appendix E gives the flag, the edits and the hand check.

5 Discussion

5.1 Recommendation

We recommend that a report of probe accuracy on an attribute predictable from the entity's string include the spelling share. An entity's name can encode the answer at a coarse resolution and a probe's accuracy is then read as knowledge of the entity. In the era of large language models, this could be a tricky confounder and lead to over-interpretation of a probe's accuracy.

Which predictor researchers choose can vary, although we offer an argument for the ridge on the tokens of the name. Specifically, five predictors that see only the name are measured at towns and in bits above the average-location baseline they run from 1.77 down to 0.93 (Appendices F and M).

The highest is a ridge on the trained model's averaged embeddings, which averages the trained embedding vectors of every token of the name. The lowest is a ridge on character n-grams of the name. The untrained control is the main predictor of this paper because it was never trained on text. On the death-year set it lies below the ridge on the tokens of the name, so there it is a lower bound on the share. The ridge on the tokens of the name lies above the control at every defined level outside the venues, and needs only the tokenizer and the split with no forward pass of the model.

5.2 Extent of the findings

Claim 1 On language models from 0.41B to 13B parameters across two families, a coordinate probe's accuracy includes a part contributed by the name alone. At metros that part is near zero. At towns it is a fifth to a quarter of the probe's gain on the five Llama models and about two fifths on the two Pythia models (Section 4.2). Under the untrained control that part does not fall with size. We further show that under the ridge on the name's token counts it falls by 0.04 and that the control's width accounts for part of its movement. The world set gives both estimators to 70.55B (Sections 4.3 and 6).

Claim 2 Whenever the attribute a probe predicts is available from the entity's string, a probe fitted to trained activations recovers that part whether or not the model stores the attribute. We further show that the measurement applies to a scalar attribute as well as a spatial attribute (a point on the sphere). On the death year of 37,539 historical figures (Gurnee & Tegmark, 2024) the baseline is the constant predictor which answers the training median year for every person. For a scalar target the bits map is $B(e) = -\log_2(2e/\mathrm{span})$ with span the year range of the training set. Then $\mathrm{bits}_X = B(m_X) - B(m_F) = \log_2(m_F/m_X)$, one bit per halving of the error and free of the span. The control recovers 0.42 and 0.27 of the trained probe's gain over it. A ridge on the tokens of the name recovers 0.83 and 0.56 and Appendix L gives the death-year set.

6 LIMITATIONS

What the spelling share does not say about the model. A share of 0.4 means that a network with random weights reaches two fifths of the probe's gain. The trained model may store the other three fifths as a feature of its own or as a lookup, its computation may or may not use either, and the share does not distinguish such cases. That the model stores the attribute as a feature and that its computation uses it are both unanswered by the spelling share (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022).

The share depends on the predictor. The untrained control is wider on a larger model (wider embedding tables recover more bits from a token) and its share moves with that width. From Llama-3.2-1B to Llama-3.1-8B the control gains 0.65 bits, whic is 86 percent of the trained probe's own gain of 0.75. A control held at its 1.2B width would put the 8B share at 0.41 instead of 0.50, so part of the share's movement with size is a property of the instrument. In contrast, the ridge on the tokens of the name does not grow with the model (one reason we recommend it). The control's dependence on its width is also why it is helpful for future researchers to report the predictor they use (Appendix C).

Frequency of names and their granularity fall together. It is rather challenging to separate the rise of the spelling share from the pretraining frequency of the names as we go down the levels from countries or metros to venues. Smaller places are written about less frequently, so the level and the frequency of the name move together down the seven levels. The trained probe's bits can include a frequency effect through the weights. Separating the two needs a count of each name in a pretraining corpus which the infini-gram index of the Pile provides (Liu et al., 2024; Gao et al., 2020) and which this paper does not use.

Level sampling. Each level is drawn from its own source and population band with its own count of training places and average-location baselines (Section 3.2), so a comparison down the levels compares the sampling of each level as well as the sizes of the places. Whenever a level is concentrated in one region, as villages are in Europe, the average location already lands in the right part of the world, which we believe is why the share is undefined there (Section 4.2). Appendix H gives how far the share moves when the baseline does.

7 CONCLUSION

This paper attempted to ask how much of a coordinate probe's accuracy is available from the name of the place alone. We measured it as the spelling share, which is the part of the probe's gain over the average-location baseline that a predictor of the name also reaches. The results show that the share rises with the granularity of the place. At towns an untrained control removes about a third of the baseline's error on every model. In bits above that baseline, the name's part of the probe's gain is under a tenth at metros, a fifth to two fifths at towns and above two fifths at the Yelp venue level. The levels fall in that order on every model measured. On the world set of the published studies the name removes more of the probe's gain than at any settlement level. The trained probe there is further above the baseline than at any of our own seven levels.

We showed that under the untrained control the share does not fall as the model grows. On the world set it rises slightly to the largest model we measured. Under the ridge on the token counts of the name the same share falls by 0.04 from 1.2B to 70.55B, and part of the control's movement is a property of

its width. The share on the world set is near a half, which is also the random-placement expectation of Section 3.3.

We further showed, albeit warranting further investigation, the same on a second attribute of the death year of a person, where the name gives at least a quarter of a probe's gain. We also discovered that on venue names without a toponym the trained probes stay within a fraction of a bit of the control. Substituting a made-up word for the toponym removes the advantage of the names that contain one.

Lastly, we note that a probe's gain reported without a predictor of the name has not been separated from the name. That part is then a confounder in any claim of a learned spatial representation that rests on the probe alone.

AI USE STATEMENT

Generative AI tools were used in this work to primarily execute and monitor experiments on remote machines that provided access to compute and polish the text of the paper. They were briefly used to validate claims across references and to check the soundness of our arguments. AI was not used to generate research hypotheses, to design experiments, to interpret results or to write the paper. The authors were responsible for the research design, the analysis and the writing, and reviewed every use of generative AI. The authors take responsibility for the final content of this work, including the text, the claims and the artefacts produced with the aid of generative AI.

ETHICS STATEMENT

The measurement runs on published datasets and on models with public weights. The settlement levels are built from GeoNames (GeoNames, 2026). The first venue level averages the Gowalla check-in coordinates per venue and keeps only the venue, so no trajectory of any individual enters the data. No user identifier is read. The second venue level uses the Yelp Open Dataset under its own terms (Yelp Inc., 2024). The world set and the death-year set are the released files of Gurnee & Tegmark (2024).

REPRODUCIBILITY STATEMENT

Section 3 states the seven levels, the probe, the four baselines, the layer rule, the pooling convention and the two string predictors. Definition 1 gives the spelling share in closed form. the supplementary material lists every run behind the paper with the machine it ran on, its seeds, its wall clock and the step its output belongs to. Appendix D gives every seeded control draw, and Appendix B gives every cell of the measurement. The code, the analysis scripts and the JSON artefacts that the tables are read from are submitted as anonymised supplementary material, with the file paths of the appendix as their layout.

REFERENCES

Ayomide Abayomi, Jose G. Moreno, Karim Radouane, and Lynda Tamine. GeoBenchmark: Probing large language models for geo-spatial knowledge. In *Proceedings of LREC*, pp. 5335–5348, 2026. ACL Anthology 2026.lrec-1.417.

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of CoNLL*, 2021. arXiv:2109.06129.

Patrick Altmeyer, Andrew M. Demetriou, Antony Bartlett, and Cynthia C. S. Liem. Position: Stop making unscientific AGI performance claims. In *Proceedings of ICML*, 2024. arXiv:2402.03962.

Elan Barenholtz. World properties without world models: Recovering spatial and temporal structure from co-occurrence statistics in static word embeddings. *arXiv preprint arXiv:2603.04317*, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 2022. arXiv:2102.12452.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of ICML*, 2023. arXiv:2304.01373.

Marcus Buckmann, Quynh Anh Nguyen, and Edward Hill. Revealing economic facts: LLMs know more than they say. Staff Working Paper 1150, Bank of England, 2025. arXiv:2505.08662.

Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of KDD*, 2011.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL*, 2018. arXiv:1805.01070.

Stef De Sabbata, Stefano Mizzaro, and Kevin Roitero. Geospatial mechanistic interpretability of large language models. *arXiv preprint arXiv:2505.03368*, 2025.

Rémy Decoupes and Adrien Guille. Determinants of geographic information quality in large language models: Effects of model family, scale, language, quantization, and fine-tuning. In *Proceedings of the Fourth International Workshop on Geographic Information Extraction from Texts (GeoExT 2026)*, 2026. CEUR-WS Vol-4201, paper 3.

Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. Evaluation of geographical distortions in language models. *Machine Learning*, 114:263, 2025. arXiv:2404.17401.

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.

GeoNames. GeoNames geographical database. `https://www.geonames.org/`, 2026. Files `countryInfo.txt`, `cities15000.txt` and `cities500.txt`, downloaded 2026-08-22.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. On the scaling laws of geographical representation in language models. *arXiv preprint arXiv:2402.19406*, 2024.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.02207.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2025.

Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. *arXiv preprint arXiv:2403.10381*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP*, 2019. arXiv:1909.03368.

Xuke Hu, Jens Kersten, Friederike Klan, and Sheikh Mastura Farzana. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science*, 40(3):670–697, 2024. DOI 10.1080/13658816.2024.2405182.

Adam Karvonen. Emergent world models and latent variable estimation in chess-playing language models. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2403.15498.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.13382.

Bastien Liétard, Mostafa Abdou, and Anders Søgaard. Do language models know the way to Rome? In *Proceedings of BlackboxNLP*, 2021. arXiv:2109.07971.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2401.17377.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased. In *Proceedings of ICML*, 2024a. arXiv:2402.02680.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. GeoLLM: Extracting geospatial knowledge from large language models. In *International Conference on Learning Representations (ICLR)*, 2024b. arXiv:2310.06213.

Tessa Masis and Brendan O'Connor. Coordinates from context: Using LLMs to ground complex location references. In *Proceedings of EACL*, 2026. arXiv:2510.08741.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Blog post, `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, 2024. 25 September 2024.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of BlackboxNLP*, 2023. arXiv:2309.00941.

Mateo Neira. Probing geographic world model. Blog post, `https://mateoneira.com/probing-geographic-world-model/`, 2026. Ridge probes on Llama-3.1-8B; median error in km for cities and Greater London places.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of EACL*, 2021. arXiv:2005.00719.

Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. GPT4GEO: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020*, 2023.

Thinh Hung Truong, Jey Han Lau, and Jianzhong Qi. GPSBench: Do large language models understand GPS coordinates? *arXiv preprint arXiv:2602.16105*, 2026.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.03689.

John Wieting and Douwe Kiela. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1901.10444.

Yelp Inc. Yelp open dataset. `https://www.yelp.com/dataset`, 2024.

Kelly Zhang and Samuel R. Bowman. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of BlackboxNLP*, 2018. arXiv:1809.10040.

A CENSUS OF RELATED WORK

Table 5 lists the probing census including 34 studies from 2021 to 2026 and Table 6 the behavioural census, 42 evaluations from 2023 to 2026. In the control columns *yes* means the control is run and a number is reported, *part* means a weakened form of it, and *no* means the paper reports nothing of the kind. A control is marked only where the method, experiment, or appendix text states it. All three studies with an untrained control in the probing census use a transformer trained on board games by the authors themselves and the one behavioural evaluation with an untrained model gives it a route sequence with no place name in it (Vafa et al., 2024).

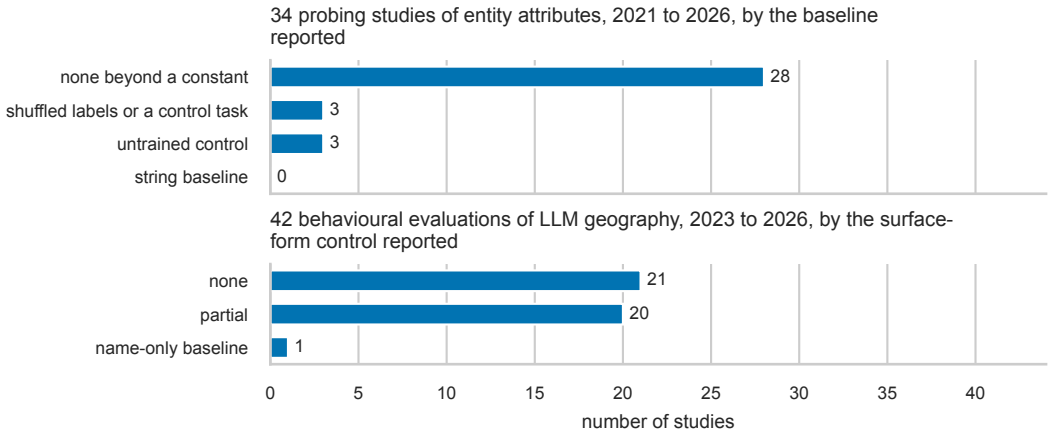

Figure 3: The probing census (upper) and the behavioural census (lower).

Table 5: The probing census, 34 studies that read an attribute of a named entity from activations. C is chance, a majority class or an average location, S shuffled or permuted labels, CT a control task, RI an untrained model of the same architecture, L0 an embedding baseline or a static word vector, SF a baseline from the characters or tokens alone, MDL description length or selectivity, SD several seeds with a spread, and STR a target that the string predicts in principle.

	Study	Identifier	Year	Target	Model	C	S	CT	RI	L0	SF	MDL	SD	STR
A1	Liétard et al.	2109.07971	2021	Latitude and longitude	BERT base	part	yes	yes	no	yes	no	yes	part	yes
A2	Abdou et al.	2109.06129	2021	CIELAB colour coordinates	BERT	no	yes	yes	no	yes	part	yes	part	part
A3	Li et al.	2210.13382	2023	Board tile state	8-layer 8-head	yes	no	no	yes	no	no	no	yes	no
A4	Nanda et al.	2309.00941	2023	Board state relative to	The 8-layer	yes	no	no	yes	part	no	no	no	no
A5	Hazineh et al.	2310.07582	2023	Board tile state	8-layer GPT	no	no	no	no	no	no	no	no	no
A6	Chen et al.	2312.16257	2023	Latitude and longitude	DeBERTa-v2	no	no	no	no	part	no	no	part	yes
A7	Gurnee and Tegmark	2310.02207	2024	Latitude and longitude	Llama-2 7B	part	no	no	no	part	no	no	no	yes
A8	Godey et al.	2402.19406	2024	Latitude and longitude	GPT-2	part	no	no	no	no	no	no	no	yes
A9	Heinzerling and Inui	2403.10381	2024	Birth year	Llama 2 7B	part	yes	no	no	no	no	no	no	yes
A10	Park et al.	2311.03658	2024	Binary concepts read	LLaMA-2 7B	no	no	no	no	no	no	no	no	part
A11	Hernandez et al.	2308.09124	2024	Relation objects	GPT-J	yes	no	no	no	no	no	no	no	part
A12	Karvonen	2403.15498	2024	Board state	8-layer GPT	part	no	no	yes	no	no	no	no	no
A13	Zhu et al.	2401.03735	2024	Numeric value of an operand	LLaMA-2 7B	no	part	part	no	part	no	no	no	yes
A14	Levy and Geva	2410.11781	2024	Digit values in base 10	Llama 3 8B	part	no	no	no	no	no	no	no	yes
A15	Engels et al.	2405.14860	2025	Day of the week	Mistral 7B	no	no	no	no	no	no	no	no	yes
A16	Kantamneni and Tegmark	2502.00873	2025	Numeric value as a helix	GPT-J	no	part	no	no	part	no	no	no	yes
A17	Buckmann et al.	2505.08662	2025	Economic statistics	Llama 3 8B	no	no	no	no	no	no	no	yes	part
A18	De Sabbata et al.	2505.03368	2025	Latitude and longitude	Mistral-7B-Inst	no	no	no	no	no	no	no	no	yes
A19	Tehenan et al.	2506.02996	2025	Spatial relation labels	Llama-3.2-3B-In	no	no	no	no	no	no	no	no	part
A20	AlquBoj et al.	2502.16147	2025	Birth year	Pythia-2.8B	no	no	no	no	part	part	no	part	yes
A21	Sakata et al.	2604.07886	2026	Parent node in a concept	Llama 3.2 3B	yes	no	no	no	no	no	no	yes	yes
A22	Barenholtz	2603.04317	2026	Latitude	GloVe 6B 300d	yes	no	no	no	n/a	no	no	yes	yes
A23	Decoupes and Guille	CEUR-WS Vol-4201	2026	Latitude and longitude	29 models	no	no	no	no	no	no	no	no	yes
A24	Neira	URL only	2026	Latitude and longitude	Llama-3.1-8B	yes	no	no	no	no	no	no	no	yes
A25	Modell	2605.18537	2026	Release year	Llama 2 7B	part	no	no	no	no	no	no	no	yes
A26	Radouane et al.	2608.07353	2026	Geo-spatial concept classes	Qwen3 0.6B	yes	no	no	no	no	no	no	no	part
B1	Burns et al.	2212.03827	2023	Truth of yes-or-no statements	T5	part	n/a	no	no	no	no	no	part	no
B2	Azaria and Mitchell	2304.13734	2023	Truth of a statement	OPT 350m	yes	part	no	no	part	no	no	part	part
B3	Marks and Tegmark	2310.06824	2024	Truth of a simple statement	LLaMA-2 7B	part	no	no	no	no	part	no	no	part
B4	Bürger et al.	2407.12831	2024	Truth of a statement	LLaMA2-13B	part	no	no	no	no	no	no	yes	part
B5	Gottesman and Geva, KEEN	2406.12673	2024	The model's own QA accuracy	GPT-2 XL	no	no	no	no	no	no	no	no	part
B6	Orgad et al.	2410.02707	2024	Correctness of the model's	Mistral-7B	no	no	no	no	no	no	no	yes	no
B7	Ferrando et al.	2411.14257	2025	Whether the model recognises	Gemma 2 2B	no	no	no	no	no	no	no	no	part
B8	Elbadry et al.	2605.09195	2026	Temporal drift class of	six models	yes	no	no	no	no	no	no	no	no

Table 6: The behavioural census, 42 evaluations in which a model is prompted about a named place and scored against a true answer. NB is a baseline that reads the name alone, AB an ablation of the name, FQ a frequency or population covariate, STR a target that the input string predicts in principle, and n/a an input that names no place.

	Study	Identifier	Year	Task	Granularity	NB	AB	FQ	STR
1	Roberts et al., GPT4GEO	2306.00020	2023	coordinates	country	no	no	yes	part
2	Bhandari et al.	2310.13002	2023	coordinates from	city	no	part	yes	part
3	Abbasi et al.	2506.00203	2025	geocoding of one	venue	no	no	no	yes
4	Decoupes and Guille	CEUR Vol-4201	2026	coordinates of French	town	no	no	no	part
5	Masis and O'Connor	2510.08741	2026	geocoding	mixed	no	part	no	part
6	Rodrigues et al., GeoBenchLLM	2608.07411	2026	eight tasks	country to venue	no	no	no	part
7	Mioduski	2508.08266	2025	coordinates	parcel	part	yes	no	part
8	Fernando et al.	2512.14228	2025	coordinates	locality	part	no	no	part
9	Decoupes et al.	2404.17401	2025	country of a capital	city	no	no	yes	yes
10	Salmas et al.	CEUR Vol-3577	2023	fill-mask X is a city	city	no	no	no	part
11	Stillman and Kruspe	CEUR Vol-3969	2025	reverse geocoding	country	n/a	n/a	yes	no
12	Truong et al., GPSBench	2602.16105	2026	17 tasks	country	no	no	yes	no
13	Li et al., STBench	2406.19065	2024	13 tasks	venue	no	n/a	no	no
14	Fulman et al.	2401.04218	2024	intercardinal	city	no	no	no	part
15	Abayomi et al., GeoBenchmark	ACL 2026.lrec-1.	2026	direction	town	no	no	no	part
16	Radouane et al.	2608.07353	2026	binary and MCQ	town	no	no	no	part
17	Holtermann et al.	2506.03984	2025	time and time-zone	city	no	no	yes	part
18	Balsebre et al., LAMP	2403.09059	2024	conversational venue	venue	no	no	no	no
19	Wang et al., GeoHaluBench	2507.19586	2025	21 hallucination	venue	no	no	no	part
20	Feng et al.	2406.13948	2025	CityEval	venue	no	no	no	part
21	Feng et al., CityBench	2406.13945	2025	eight tasks over 13	road	no	no	no	part
22	Gao et al., LocalBench	2511.10459	2025	14	county	no	no	part	part
23	Bockling et al., MultiGlobeQA	2608.03882	2026	geospatial reasoning	mixed OSM	no	no	yes	part
24	Saeedan et al., GS-QA	2605.22811	2026	2	venue to city	no	no	no	part
25	Dihan et al., MapEval	2501.00316	2025	map-based reasoning	venue	no	no	no	no
26	Manvi et al., GeoLLM	2310.06213	2024	population density	coordinate	part	yes	no	yes
27	Manvi et al.	2402.02680	2024	zero-shot ratings for	coordinate	no	yes	yes	yes
28	Schwobel et al.	2310.14777	2023	the country	country	n/a	n/a	yes	no
29	Moayeri et al., WorldBench	DOI 10.1145/3630106.365896	2024	recall of 11 World	country	no	no	part	no
30	Bhagat et al.	2411.07320	2025	story generation	city and town	no	no	yes	part
31	Lalai et al.	2508.05525	2025	20 Questions	entity	n/a	n/a	yes	no
32	Mai et al.	2304.06798	2023	toponym recognition	toponym	no	no	no	yes
33	Hu et al.	2310.09340	2023	extraction of nine	address	no	no	no	yes
34	Hu et al.	DOI 10.1080/13658816.2024.	2024	toponym resolution	any toponym	part	no	part	part
35	Hu et al.	CEUR Vol-3969	2025	parallel toponym	any toponym	no	no	no	part
36	Cafferata et al.	2602.08872	2026	toponym tagging	any toponym	no	no	part	yes
37	Yin et al.	2310.14360	2023	address parsing on 239	address	part	no	no	yes
38	Ji et al.	2505.17136	2025	topological relation	parcel	no	no	no	no
39	Zhang et al., GeoAnalystBench	2509.05881	2025	50 geoprocessing	not applicable	no	no	no	no
40	Krechetova and Kochedykov, GeoBenchX	2503.18129	2025	202 multi-step	country	no	no	no	no
41	Buckmann et al.	2505.08662	2025	county-level	county	no	no	no	no
42	Vafa et al.	2406.03689	2024	next-turn prediction	intersection	yes	n/a	no	no

B Detailed Experiments

Table 7: Every run of the measurement each with its control. L^* is the validation-selected layer, T/C for the trained probe and its control. Medians are in kilometres and bits are above the average-location baseline. Brackets are the bootstrap interval and round brackets the median and interquartile range of S over the seeded control draws. A double dagger marks a cell whose lowest shuffled-label median over layers is under 0.8 of the baseline with S in square brackets and not used.

Level	n	L^* T/C	m_F	m_T	m_C	bits$_T$	bits$_C$	M	S (seeds)	
Pythia-410m										
countries	241	20/0	5,700	2,708	6,206	2.07	−0.23	2.30 [1.54, 2.92]	undefined	
metros	554	22/0	3,843	1,667	3,648	2.37	0.15	2.23 [1.83, 2.65]	0.06 [−0.07, 0.18]	
cities	5,000	20/1	5,783	2,058	4,303	2.89	0.81	2.09 [1.97, 2.21]	0.28 [0.24, 0.31]	
towns	8,000	24/1	5,847	2,231	3,937	2.69	1.09	1.61 [1.51, 1.72]	0.40 [0.38, 0.43]	(0.40 [0.39, 0.40])
villages	8,000	21/1	2,054	1,681	2,563	0.57	−0.63	1.21 [1.09, 1.31]	undefined	
Gowalla venues	10,000	24/15	1,607	1,415	1,508	0.37	0.18	0.18 [0.13, 0.24]	0.50 [0.39, 0.63]	(0.50 [0.46, 0.52])
Yelp venues	5,000	19/2	1,199	1,075	1,091	0.31	0.27	0.04 [−0.01, 0.11]	0.86 [0.67, 1.04]	(0.76 [0.74, 0.82])
Pythia-1.4B										
countries	241	19/15	5,700	1,708	5,911	3.39	−0.10	3.49 [2.94, 3.82]	undefined	
metros	554	24/1	3,843	1,194	3,986	3.33	−0.10	3.44 [3.04, 3.86]	undefined	
cities	5,000	20/1	5,783	1,727	4,074	3.40	0.96	2.44 [2.31, 2.57]	0.28 [0.26, 0.31]	
towns	8,000	21/0	5,847	1,942	3,815	3.09	1.17	1.92 [1.80, 2.05]	0.38 [0.34, 0.41]	
villages	8,000	23/1	2,054	1,646	2,602	0.64	−0.67	1.31 [1.20, 1.43]	undefined	
Gowalla venues	10,000	23/11	1,607	1,350	1,498	0.50	0.20	0.30 [0.25, 0.36]	0.41 [0.31, 0.47]	
Yelp venues	5,000	18/0	1,199	1,051	1,114	0.38	0.21	0.17 [0.09, 0.24]	0.56 [0.43, 0.72]	
Llama-3.2-1B										
countries	241	12/12	5,700	1,287	6,132	4.20	−0.20	4.40 [3.94, 4.66]	undefined	
metros	554	16/0	3,843	761	3,576	4.63	0.20	4.43 [4.08, 4.85]	0.04 [0.00, 0.10]	
cities	5,000	13/0	5,783	847	4,168	5.45	0.90	4.55 [4.39, 4.70]	0.16 [0.14, 0.19]	
towns	8,000	14/0	5,847	1,137	3,821	4.63	1.17	3.46 [3.34, 3.58]	0.25 [0.23, 0.27]	(0.23 [0.22, 0.23])
villages	8,000	13/0	2,054	1,132	2,551	1.71	−0.62	2.33 [2.23, 2.44]	undefined	
Gowalla venues	10,000	16/1	1,607	1,259	1,536	0.70	0.13	0.57 [0.51, 0.63]	0.18 [0.13, 0.24]	(0.27 [0.27, 0.28])
Yelp venues	5,000	16/0	1,199	1,019	1,111	0.47	0.22	0.25 [0.18, 0.32]	0.47 [0.36, 0.60]	(0.55 [0.54, 0.58])
Llama-3.2-3B										
countries	241	21/0	5,700	1,258	6,263	4.27	−0.25	4.52 [4.01, 5.01]	undefined‡	
metros	554	22/21	3,843	633	4,222	5.16	−0.26	5.42 [4.97, 5.71]	undefined	
cities	5,000	20/0	5,783	770	4,090	5.72	0.95	4.77 [4.63, 4.94]	0.17 [0.15, 0.18]	
towns	8,000	20/0	5,847	986	3,883	5.04	1.12	3.91 [3.79, 4.03]	0.22 [0.21, 0.24]	
villages	8,000	19/0	2,054	1,072	2,507	1.87	−0.57	2.44 [2.28, 2.59]	undefined	
Gowalla venues	10,000	25/0	1,607	1,238	1,520	0.75	0.16	0.59 [0.53, 0.65]	0.21 [0.16, 0.26]	
Yelp venues	5,000	18/0	1,199	1,007	1,110	0.50	0.22	0.28 [0.20, 0.34]	0.44 [0.36, 0.56]	
Llama-2-7B										
countries	241	23/29	5,700	1,261	5,586	4.26	0.05	4.21 [3.66, 4.70]	[0.01]‡ [−0.09, 0.12]	
metros	554	22/4	3,843	699	3,703	4.87	0.10	4.77 [4.39, 5.07]	0.02 [−0.03, 0.05]	
cities	5,000	21/0	5,783	853	3,937	5.42	1.06	4.37 [4.22, 4.52]	0.19 [0.17, 0.22]	
towns	8,000	23/0	5,847	1,142	3,830	4.61	1.16	3.45 [3.32, 3.55]	0.25 [0.24, 0.27]	
villages	8,000	32/0	2,054	1,157	2,368	1.65	−0.41	2.05 [1.92, 2.19]	undefined	
Gowalla venues	10,000	32/0	1,607	1,250	1,511	0.72	0.18	0.55 [0.48, 0.61]	0.25 [0.19, 0.32]	
Yelp venues	5,000	23/0	1,199	1,006	1,113	0.51	0.21	0.29 [0.21, 0.37]	0.42 [0.33, 0.54]	
Llama-3.1-8B										
countries	241	22/2	5,700	1,226	6,181	4.34	−0.22	4.56 [4.15, 5.00]	undefined	
metros	554	21/0	3,843	652	3,649	5.08	0.14	4.93 [4.64, 5.31]	0.03 [−0.02, 0.06]	
cities	5,000	22/0	5,783	706	4,068	5.97	0.96	5.01 [4.85, 5.15]	0.16 [0.14, 0.18]	(0.16 [0.16, 0.16])
towns	8,000	20/0	5,847	887	3,919	5.34	1.10	4.24 [4.09, 4.36]	0.21 [0.19, 0.23]	(0.21 [0.21, 0.21])
villages	8,000	22/0	2,054	947	2,428	2.23	−0.48	2.70 [2.54, 2.85]	undefined	
Gowalla venues	10,000	32/0	1,607	1,187	1,498	0.87	0.20	0.67 [0.60, 0.75]	0.23 [0.18, 0.29]	(0.22 [0.22, 0.23])
Yelp venues	5,000	19/0	1,199	974	1,097	0.60	0.26	0.34 [0.25, 0.43]	0.43 [0.36, 0.53]	(0.42 [0.41, 0.42])
Llama-2-13B										
countries	241	28/1	5,700	1,272	5,362	4.24	0.17	4.07 [3.57, 4.58]	[0.04]‡ [−0.08, 0.15]	
metros	554	25/3	3,843	722	3,743	4.78	0.07	4.71 [4.35, 5.07]	0.02 [−0.04, 0.08]	
cities	5,000	26/0	5,783	826	3,885	5.52	1.09	4.42 [4.29, 4.59]	0.20 [0.18, 0.22]	
towns	8,000	25/0	5,847	1,102	3,845	4.72	1.15	3.57 [3.44, 3.66]	0.24 [0.23, 0.27]	
villages	8,000	23/0	2,054	1,168	2,346	1.62	−0.38	2.00 [1.86, 2.14]	undefined	
Gowalla venues	10,000	23/0	1,607	1,246	1,518	0.73	0.16	0.57 [0.50, 0.63]	0.22 [0.18, 0.28]	
Yelp venues	5,000	32/0	1,199	999	1,114	0.52	0.21	0.31 [0.23, 0.39]	0.40 [0.28, 0.52]	

C SHARE AGAINST MODEL SIZE

Table 8: The world set by model: S with its bootstrap interval, the trained and control medians in kilometres, the bits above the average-location baseline on 26,632 held-out places, and S under the published split of the set in (Gurnee & Tegmark, 2024).

Model	params (B)	S	m_T	m_C	bits$_T$	bits$_C$	S, paper split
Pythia-410m	0.41	0.607 [0.59, 0.62]	1,388	2,501	4.29	2.60	0.616
Pythia-1B	1.0	0.685 [0.67, 0.70]	1,265	2,088	4.56	3.12	0.701
Llama-3.2-1B	1.2	0.459 [0.45, 0.47]	642	2,187	6.51	2.99	0.432
Llama-3.2-3B	3.2	0.482 [0.47, 0.49]	551	1,925	6.95	3.35	0.445
Llama-2-7B	6.7	0.499 [0.49, 0.51]	612	1,946	6.65	3.32	0.484
Llama-3.1-8B	8.0	0.501 [0.49, 0.52]	495	1,742	7.26	3.64	0.465
Llama-2-13B	13.0	0.495 [0.48, 0.51]	590	1,932	6.75	3.34	0.483
Llama-3.1-70B	70.55	0.522 [0.51, 0.53]	522	1,700	7.11	3.71	0.503

Table 9: The world-set share under four predictors of the name each divided by the trained bits of Table 8..

		world-set share from			
Model	params (B)	control	n-grams	tokens, Pythia	tokens, Llama
Pythia-410m	0.41	0.607	0.646	0.739	0.741
Pythia-1B	1.0	0.685	0.608	0.696	0.698
Llama-3.2-1B	1.2	0.459	0.426	0.487	0.488
Llama-3.2-3B	3.2	0.482	0.399	0.456	0.457
Llama-2-7B	6.7	0.499	0.417	0.477	0.478
Llama-3.1-8B	8.0	0.501	0.381	0.437	0.438
Llama-2-13B	13.0	0.495	0.410	0.470	0.471
Llama-3.1-70B	70.55	0.522	0.390	0.446	0.447

D CONTROL DRAWS

Table 10: Every seeded control draw, pooled per entity: k the seed, L_C the draw's layer, S its share against the same trained probe, a dagger an undefined S, a double dagger a failed shuffled-label check, median rows the median of m_C over the draws.

Model	Level	k	L_C	m_C	bits$_C$	S	Model	Level	k	L_C	m_C	bits$_C$	S
Pythia-410m	towns	1	0	4,010	1.034	0.384	Llama-3.2-1B	Yelp venues	1	0	1,096	0.257	0.549
Pythia-410m	towns	2	1	3,919	1.098	0.408	Llama-3.2-1B	Yelp venues	2	0	1,099	0.251	0.537
Pythia-410m	towns	3	0	3,953	1.074	0.399	Llama-3.2-1B	Yelp venues	3	0	1,087	0.283	0.605
Pythia-410m	towns	4	1	3,993	1.046	0.388	*median*	Yelp venues			1,096	0.257	0.549
Pythia-410m	towns	5	1	3,958	1.071	0.398	Llama-3.1-8B	countries	1	2	6,181	−0.216	†
median	towns			3,958	1.071	0.398	Llama-3.1-8B	countries	2	28	5,346	0.173	0.040
Pythia-410m	Gowalla venues	1	21	1,531	0.139	0.379	Llama-3.1-8B	countries	3	32	6,063	−0.165	†‡
Pythia-410m	Gowalla venues	2	4	1,490	0.217	0.592	*median*	countries			6,063	−0.165	†
Pythia-410m	Gowalla venues	3	7	1,515	0.170	0.464	Llama-3.1-8B	metros	1	0	3,649	0.145	0.029
Pythia-410m	Gowalla venues	4	1	1,508	0.183	0.499	Llama-3.1-8B	metros	2	3	4,272	−0.295	†
Pythia-410m	Gowalla venues	5	9	1,503	0.192	0.523	Llama-3.1-8B	metros	3	26	4,101	−0.182	†
median	Gowalla venues			1,508	0.183	0.499	*median*	metros			4,101	−0.182	†
Pythia-410m	Yelp venues	1	2	1,089	0.276	0.880	Llama-3.1-8B	cities	1	0	4,068	0.964	0.162
Pythia-410m	Yelp venues	2	6	1,107	0.228	0.727	Llama-3.1-8B	cities	2	0	4,077	0.958	0.161
Pythia-410m	Yelp venues	3	1	1,103	0.240	0.763	Llama-3.1-8B	cities	3	0	4,119	0.930	0.156
median	Yelp venues			1,103	0.240	0.763	*median*	cities			4,077	0.958	0.161
Llama-3.2-1B	towns	1	0	3,995	1.044	0.226	Llama-3.1-8B	towns	1	0	3,919	1.098	0.206
Llama-3.2-1B	towns	2	0	4,006	1.037	0.224	Llama-3.1-8B	towns	2	0	3,891	1.118	0.209
Llama-3.2-1B	towns	3	0	3,971	1.062	0.229	Llama-3.1-8B	towns	3	0	3,870	1.133	0.212
Llama-3.2-1B	towns	4	0	3,984	1.052	0.227	*median*	towns			3,891	1.118	0.209
Llama-3.2-1B	towns	5	0	4,057	1.002	0.216	Llama-3.1-8B	villages	1	0	2,428	−0.477	†
median	towns			3,995	1.044	0.226	Llama-3.1-8B	villages	2	0	2,474	−0.531	†
Llama-3.2-1B	Gowalla venues	1	0	1,517	0.165	0.235	Llama-3.1-8B	villages	3	0	2,399	−0.442	†
Llama-3.2-1B	Gowalla venues	2	0	1,499	0.201	0.286	*median*	villages			2,428	−0.477	†
Llama-3.2-1B	Gowalla venues	3	0	1,504	0.190	0.270	Llama-3.1-8B	Gowalla venues	1	0	1,498	0.202	0.231
Llama-3.2-1B	Gowalla venues	4	0	1,501	0.197	0.280	Llama-3.1-8B	Gowalla venues	2	0	1,503	0.192	0.220
Llama-3.2-1B	Gowalla venues	5	0	1,506	0.187	0.266	Llama-3.1-8B	Gowalla venues	3	0	1,504	0.191	0.219
median	Gowalla venues			1,504	0.190	0.270	*median*	Gowalla venues			1,503	0.192	0.220
							Llama-3.1-8B	Yelp venues	1	0	1,097	0.256	0.427
							Llama-3.1-8B	Yelp venues	2	0	1,098	0.252	0.421
							Llama-3.1-8B	Yelp venues	3	0	1,105	0.234	0.391
							median	Yelp venues			1,098	0.252	0.421

E TOPONYM SPLIT

Table 11: The toponym split of both venue levels with the average-location baseline recomputed inside each subset. Medians are in kilometres, M is in bits, and S is against the subset's own baseline.

	Without a toponym					With a toponym				
Cell	n	m_T	m_C	M	S	n	m_T	m_C	M	S
Pythia-410m, Gowalla	5,135	1,502	1,537	0.07 [0.01, 0.13]	0.69	1,591	1,142	1,411	0.61 [0.48, 0.72]	0.34
Pythia-410m, Yelp	3,062	1,105	1,114	0.02 [−0.04, 0.08]	0.91	300	618	805	0.76 [0.49, 1.10]	0.62
Pythia-1.4B, Gowalla	5,135	1,447	1,550	0.20 [0.13, 0.26]	0.39	1,591	1,009	1,370	0.88 [0.76, 1.04]	0.31
Pythia-1.4B, Yelp	3,062	1,097	1,137	0.11 [0.04, 0.17]	0.59	300	534	592	0.30 [−0.37, 1.15]	0.88
Llama-3.2-1B, Gowalla	5,135	1,414	1,551	0.27 [0.21, 0.33]	0.32	1,591	850	1,474	1.58 [1.42, 1.74]	0.10
Llama-3.2-1B, Yelp	3,062	1,071	1,140	0.18 [0.12, 0.25]	0.44	300	439	446	0.04 [−1.00, 0.79]	0.99
Llama-3.2-3B, Gowalla	5,135	1,360	1,565	0.40 [0.34, 0.47]	0.19	1,591	851	1,350	1.33 [1.16, 1.47]	0.25
Llama-3.2-3B, Yelp	3,062	1,061	1,143	0.22 [0.14, 0.28]	0.39	300	416	408	−0.06 [−1.77, 0.82]	1.02
Llama-2-7B, Gowalla	5,135	1,384	1,560	0.34 [0.28, 0.40]	0.24	1,591	874	1,333	1.22 [1.07, 1.40]	0.28
Llama-2-7B, Yelp	3,062	1,059	1,137	0.21 [0.13, 0.28]	0.42	300	448	610	0.89 [0.22, 1.63]	0.70
Llama-3.1-8B, Gowalla	5,135	1,320	1,543	0.45 [0.38, 0.51]	0.23	1,591	818	1,336	1.41 [1.24, 1.58]	0.25
Llama-3.1-8B, Yelp	3,062	1,027	1,125	0.26 [0.17, 0.33]	0.41	300	470	353	−0.82 [−2.41, 0.55]	1.30
Llama-2-13B, Gowalla	5,135	1,358	1,556	0.39 [0.33, 0.46]	0.23	1,591	872	1,350	1.26 [1.09, 1.40]	0.26
Llama-2-13B, Yelp	3,062	1,048	1,142	0.25 [0.17, 0.32]	0.36	300	542	580	0.20 [−0.32, 1.08]	0.92

Toponym flagging logic The toponym flag fires when a name contains (whole words) a United States city or state name from GeoNames `cities15000` longer than three characters, one of six short forms, or a postal abbreviation. It marks 2,364 Gowalla and 467 Yelp names whose advantage over the names without a toponym is 0.70 and 1.38 bits on the Gowalla level and 1.76 and 2.65 bits on the Yelp level. A stricter flag built from names outside the United States fires on 4.3 and 4.7 percent of the two subsets without a toponym. The made-up word alternates consonants and vowels and is never a listed toponym, the substitution draws a toponym of the same type from the same list, and neither refits the probe. Both edits change the token count of the name with the median going from 6 tokens to 8 under the made-up word and 7 under substitution on the Gowalla level, and a change of length alone gives the prediction no direction, whereas the substitution moves it a median 0.10 to 0.34 of the way toward the substituted place (Table 4). The edit list is `results/toponym_ablation_edits.csv` and the 50-name hand check `results/HANDCHECK-toponyms-labelled.csv` both in the supplementary material.

F STRING BASELINES

Table 12: The string predictors per level: the median in kilometres and the bits above the average-location baseline.

	character n-grams		tokens of the name, Pythia		tokens of the name, Llama		one-feature baselines	
Level	km	bits	km	bits	km	bits	tokens	ids
countries	5,355	0.17	5,481	0.11	5,214	0.24	0.04	0.09
metros	3,353	0.38	3,431	0.32	3,507	0.26	−0.01	−0.02
cities	4,221	0.86	3,787	1.16	3,826	1.14	0.00	0.04
towns	4,161	0.93	3,523	1.40	3,573	1.36	−0.01	0.10
villages	3,142	−1.21	2,531	−0.60	2,481	−0.54	−0.12	−0.31
Gowalla venues	1,483	0.23	1,544	0.12	1,542	0.12	0.00	0.00
Yelp venues	1,170	0.07	1,089	0.28	1,112	0.22	0.00	−0.02

G CONTROL

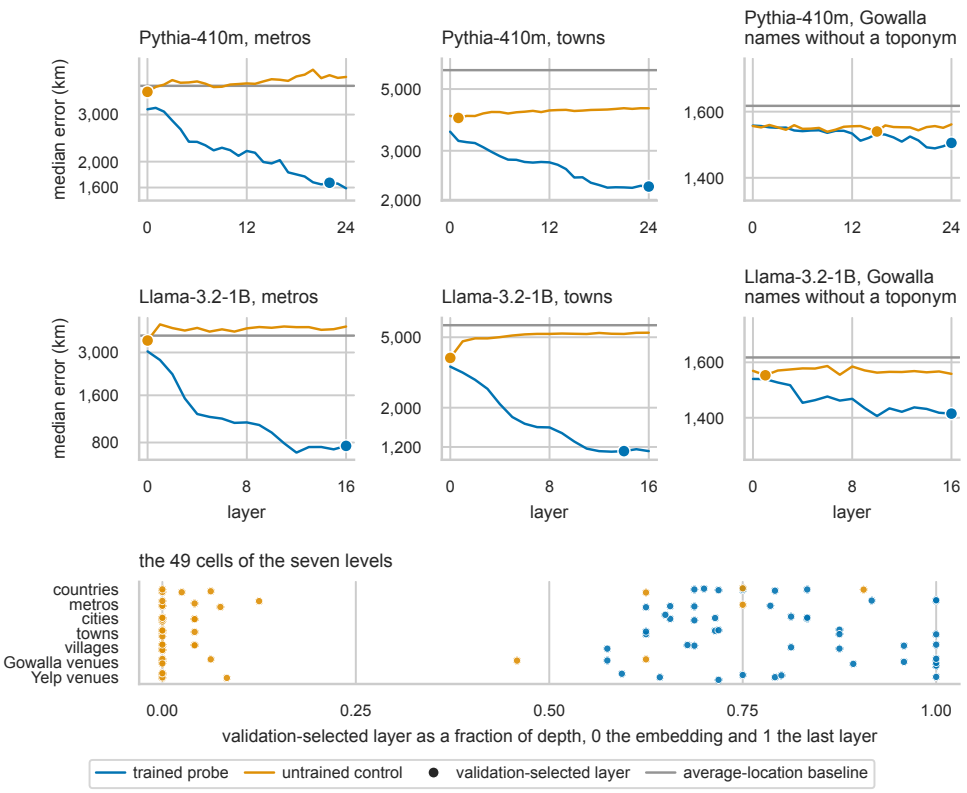

Figure 4: The median error by layer of the trained probe (blue) and the untrained control (orange) on Pythia-410m and Llama-3.2-1B at metros, towns and the Gowalla level, with a dot at each validation-selected layer and the grey line the average-location baseline.

H WHAT THE SPELLING SHARE DEPENDS ON

The spelling share compares two improvements over the same baseline of the error of guessing the average location by level. With a baseline that already knows each place's continent (1,357 km on the world set) or its country (395 km), the untrained control does not beat the baseline at all so the share is undefined on every model. Table 13 holds each model's trained and control errors fixed and recomputes the share as the baseline is moved from the paper's 6,394 km down to 2,100 km. Every model's share falls and the models drift apart from a spread of 0.063 to 0.164. The paper therefore uses the loosest baseline of the level's own average location where the models agree most.

Table 13: The world-set share of the six Llama models under a swept average-location baseline, with the trained and control medians of Table 8 held fixed. The first row is the baseline of the paper.

m_F (km)	1.2B	3.2B	6.7B	8B	13B	70.55B	spread
6,394	0.459	0.482	0.499	0.501	0.495	0.522	0.063
5,000	0.398	0.428	0.444	0.451	0.440	0.473	0.075
4,000	0.327	0.365	0.380	0.394	0.377	0.417	0.090
3,000	0.203	0.260	0.270	0.300	0.269	0.323	0.120
2,300	0.039	0.124	0.125	0.180	0.127	0.203	0.164
2,100	undefined	0.065	0.061	0.129	0.065	0.151	undefined

I SPELLING SHARE BY LEVEL

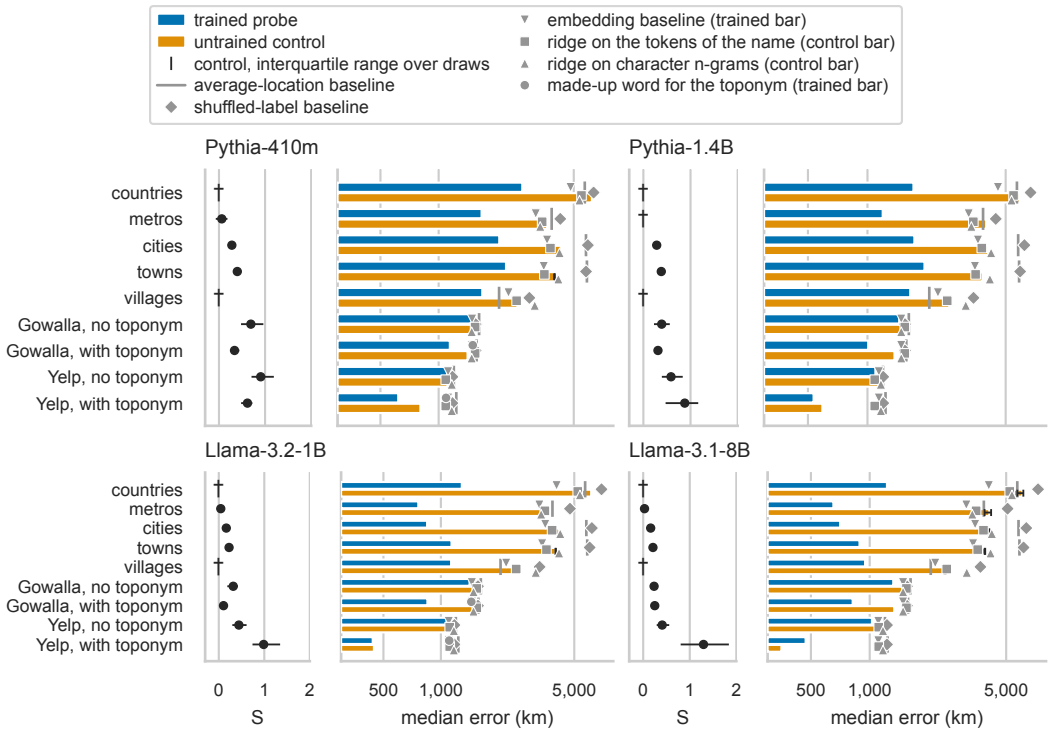

Figure 5: The spelling share and the bits at every level on four models with the venue levels split by toponym. The left panel of each pair is S with its interval..

J LLAMA 3.1-70B ON THE WORLD SET AND YELP VENUES

Table 14: Llama-3.1-70B on the world set under the five seeded splits, under the published split with the lat/lng target, and on the Yelp venues with one control draw: the held-out count, the trained probe's selected layer of 81, the medians in kilometres, the bits above the average-location baseline, S with its bootstrap interval, and the control's ceiling, the exact lookup on the name's last token under the Llama-3 tokenizer. The other six levels were not run.

Set	n	L_T^*	m_F	m_T	m_C	bits$_T$	bits$_C$	S	ceiling (bits)
world set, seeded splits	26,632	46	6,394	522	1,700	7.11	3.71	0.522 [0.508, 0.534]	3.90
world set, published split	7,917	—	—	854	2,281	—	—	0.503 [0.48, 0.53]	3.90
Yelp venues	3,362	80	1,199	874	1,103	0.91	0.24	0.264 [0.212, 0.340]	—

A share of 0.50 against the control's ceiling of 3.90 bits needs a trained probe at or below 7.80 bits above the 6,394 km baseline, which on this set is a median error of about 410 km, so the share must fall once the trained probe passes that point.

K STABILITY CHECKS

Table 15: The learning curve at towns per-entity pooled on the 5,333 paired entities: m_T and m_C in kilometres at their own validation-selected layers, the last row is the full training set of 5,600.

	Pythia-410m			Llama-3.2-1B		
Training entities	m_T	m_C	S	m_T	m_C	S
250	2,661	5,072	0.18	1,602	5,177	0.09
388	2,555	4,985	0.20	1,502	5,024	0.12
500	2,521	4,859	0.22	1,439	4,999	0.11
1,000	2,403	4,504	0.28	1,289	4,549	0.16
2,000	2,320	4,353	0.31	1,242	4,347	0.19
4,000	2,264	4,086	0.37	1,173	4,000	0.23
5,600	2,231	4,010	0.38	1,137	3,995	0.23

Table 16: Stability checks at towns: the selectivity gap in bits between the real task and a control task of coordinates sampled with replacement and fixed per name, the ratio of the error of a two-layer network (256 hidden units, a ReLU, Adam at 10^{-3}, early stopping on the validation median) to the ridge's at the same layer on the first 4,000 training entities of the seed-0 split.

	Pythia-410m	Llama-3.2-1B
selectivity gap, trained probe (bits)	2.74	4.73
selectivity gap, control (bits)	1.14	1.25
network over ridge, trained probe	1.08	1.50
network over ridge, control	1.00	1.07
penalty sweep, movement 10^3 to 10^4 (km)	41	61
effective degrees of freedom at 10^3	592	1,016
effective degrees of freedom at 10^5	39	70

L DEATH-YEAR SET

The second attribute is the death year of the 37,539 historical figures of Gurnee & Tegmark (2024). The baseline is the constant predictor at the train median which errs by 270 years. The bits of an error e in years are $B(e) = -\log_2(2e/\text{span})$ with span the range of the training-set years so $\text{bits}_X = B(m_X) - B(m_F) = \log_2(m_F/m_X)$, one bit per halving of the error and independent of the span (`scripts/spelling_share_year.py`). On 25,228 paired held-out entities the trained probes reach 158.1 and 122.1 years against the control's 216.2 and 218.3 years. The share is 0.42 [0.39, 0.44] and 0.27 [0.25, 0.28]. Birth year on the same saved activations gives 0.38 and 0.27.

The baseline from shuffled labels is above the constant predictor on both models and the embedding baseline of the last token is 212.7 and 198.7 years. Three seeded control draws on Pythia-410m give an interquartile range of 0.01 in the share and a full spread of 0.02. On the published test split the probe reaches an R^2 of 0.618 and 0.731 against the 0.785 that Gurnee & Tegmark (2024) report for a 7B model.

The prior in the string differs between the two attributes with (1) a region prior in place names and (2) an era prior in given names. On this attribute the string predictors do not bracket the control as they do at towns where they lie either side of it at 0.93 and 1.40 bits against 1.09. Here the character n-gram ridge recovers 0.520 bits above the constant predictor and the ridge on the tokens of the name 0.64 bits against the control's 0.32 and 0.31 bits so the control is a lower bound. A network that was never trained recovers at least 0.42 and 0.27 of the trained probe's gain on this set and the ridge on the tokens of the name recovers 0.83 and 0.56.

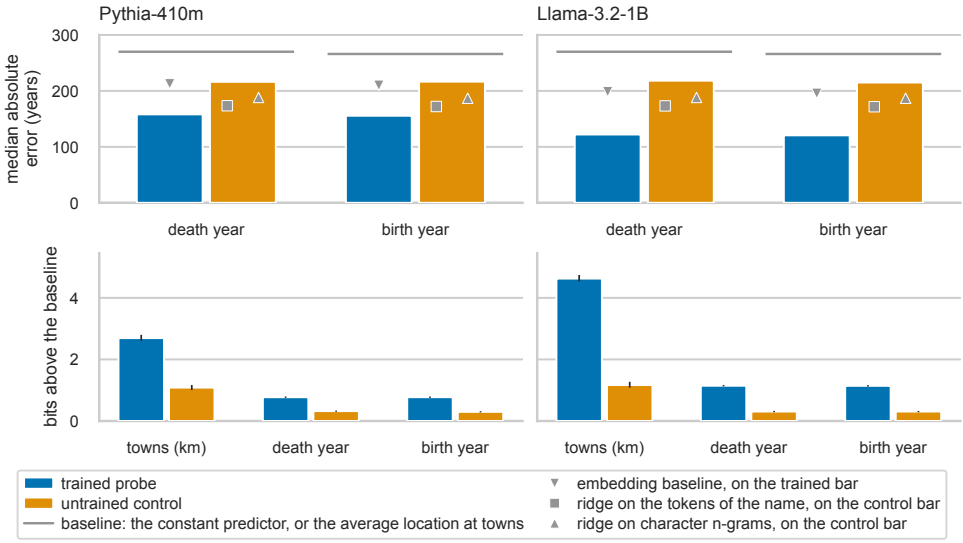

Figure 6: The death-year set on Pythia-410m and Llama-3.2-1B: the trained probe (blue) and the untrained control (orange) at their validation-selected layers, in years, the grey line the constant predictor at the training median, the triangle on the trained bar its embedding baseline, and the square and the triangle on the control bar the ridge on the tokens of the name and the character n-gram ridge. The lower panels give the same in bits beside the towns level. Every bar carries its bootstrap interval.

M PREDICTORS OF THE NAME

Table 17: Predictors that see only the name at towns, in bits above the 5,847 km average-location baseline. Row 2 is fitted under each tokenizer, row 3 is the first seeded control draw where one exists, row 4 the single draw of Table 7, and row 5 sees no model.

		bits at towns		
Predictor	Usage	Pythia-410m	Pythia-1.4B	Llama-3.2-1B
1. Ridge on the trained model's averaged embeddings	the trained embedding of every token	1.63	1.61	1.77
2. Ridge on the tokens of the name	the token ids of the name	1.40	1.40	1.36
3. Ridge on the control's averaged embeddings	a random embedding of every token	1.13	—	1.29
4. Control at its own layer	a random embedding of the last token	1.09	1.17	1.17
5. Ridge on character n-grams of the name	the characters of the name	0.93	0.93	0.93

Row 1 is the upper estimate and row 5 the lower one. The gap between row 1 and row 2 is 0.21 to 0.42 bits. That gap is the contribution of the trained embedding table over token identity alone.