TESTING THE PLATONIC REPRESENTATION HYPOTHESIS WITH LIFE-EVENT MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Do models trained on life histories and language learn similar relationships between events? We test the Platonic Representation Hypothesis by comparing (a) models trained on nationwide registry histories with (b) independently trained language models given descriptions of the same events. Across life models of increasing size, we measure local alignment, whether the two representation spaces agree on which events are neighbours, and global alignment, whether they agree on the similarities between all pairs of events. Following the Aristotelian critique, we calibrate these comparisons for chance agreement and layer selection. Local alignment increases at every life-model size step for every language model in the original comparison. Better language performance is also associated with greater local alignment, both in this cohort and in a broader comparison of language models. These local-alignment findings persist after controls for shared wording and coding categories. Global alignment also exceeds chance for every pairing, but it stops increasing with life-model size beyond a few million parameters. The results support the local form of the Platonic Representation Hypothesis in a modality that shares no training data with language.

1 INTRODUCTION

As neural networks become more capable, do they learn increasingly similar relationships among the things they represent? The Platonic Representation Hypothesis (PRH) proposes that models converge towards a shared statistical model of reality (Huh et al., 2024). Its vision–language experiments show that better language models tend to agree more closely with vision models about which examples are neighbours. That observation leaves open how much of the representation is shared: agreement about neighbours does not require agreement about distances or the geometry of the whole space.

The Aristotelian Representation Hypothesis makes this distinction central (Gröger et al., 2026). Its authors show that model width and layer selection can inflate similarity scores, and use permutation calibration to account for these effects. Local alignment survives this calibration. However, the apparent scaling of global alignment largely disappears, and agreement about fine-scale distances is not retained. Their results support the local component of the PRH. We ask whether this pattern also holds for life-event models, and whether their agreement with language extends from local to global alignment.

We bring a new modality to this comparison: life histories assembled from linked Danish administrative registries. The registries are extraordinarily comprehensive. They cover the entire Danish population, about 10 million people, from the 1970s through 2023, and link hospital care, doctor visits and prescriptions with education and grades, employment and income, criminal justice and social care, and demographic events such as marriage, migration and changes of address (Statistics Denmark, 2026b). For people born during this period, the records follow them from birth. Because the registries can be linked across years, one history can contain an occupation, a medical diagnosis and a family event, so a model trained on these histories can learn relationships that cross domains.

Our life models are trained to predict the next token in these chronological histories, following the sequence-modelling approach of Savcisens et al. (2024), and they are not trained on the codes' descriptions. Language models read descriptions of the same events but have never seen a registry history, because the histories are confidential. The two sides therefore share no training data, and

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.

their only link is the human-written description of each code. Consider the occupation "weaving- and knitting-machine operators" and the industry "production of pullovers, cardigans etc." The two descriptions share no word. A life model can place these codes close together because the same people hold that occupation and work in that industry; a language model can do so because text links knitting machines to knitwear. When both models treat the two codes as neighbours, they agree about a relationship that neither learned from the other. Agreement can also arise for less interesting reasons, such as similar wording or membership of the same coding category, and we control for both.

We compare local and global alignment under the same calibration, first for the language models of the original PRH experiment (the PRH cohort) and then for a broader set ranked by HellaSwag accuracy (the HellaSwag cohort). Our contributions are:

- A test of the PRH in a new modality, in which life-event and language models are trained separately and linked only at evaluation, through the descriptions of shared event codes.
- Evidence that local alignment, corrected for chance and layer selection, grows with life-model size across the whole range we train and with language performance in both cohorts. Beyond the smallest life model, it follows the life model's validation loss almost linearly.
- Evidence that global alignment is above chance but scales differently: it stops growing once life models reach a few million parameters, and its relation to language performance holds for every life model in the HellaSwag cohort but only for the larger life models in the PRH cohort.

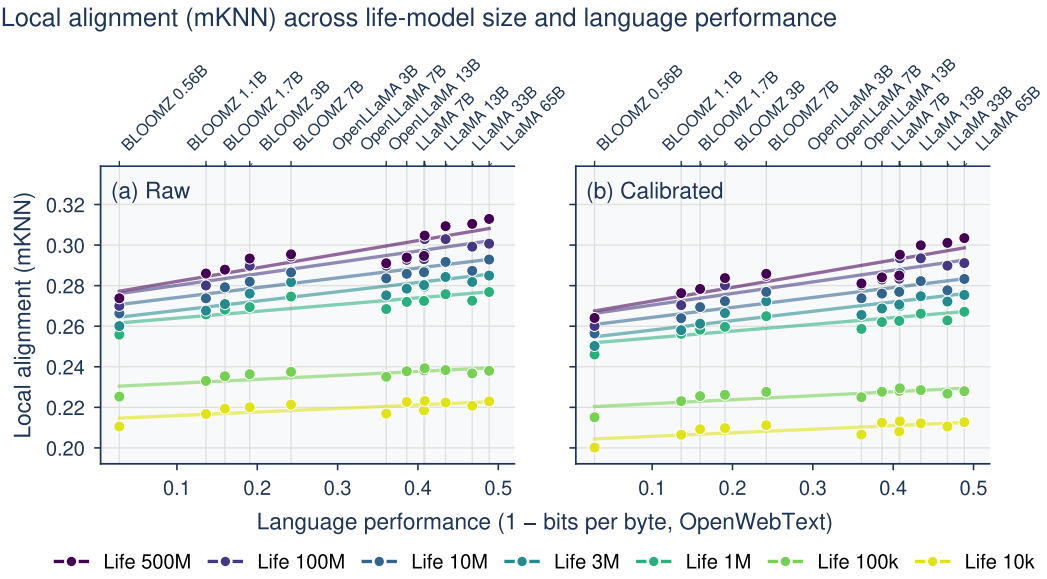


Figure 1: Local alignment across life-model sizes and language performance. (a) Raw and (b) calibrated local alignment (mKNN) of the PRH cohort on 1,024 matched code–description pairs, using the best layer pair (Section 3); calibration repeats this layer search in each of 200 permutations. Colours identify life-model sizes; App. Fig. 6 shows the range across training runs.

2 Related work

Representational convergence. Huh et al. (2024) report increasing vision–language alignment with capability and propose a co-occurrence kernel as its mechanism. Subsequent studies examine convergence in chemistry, proteins and interatomic potentials (Edamadaka et al., 2025; Li & Walsh, 2026), world models (Li et al., 2026), astronomy (Duraphe et al., 2025), and across layers and training (Kapoor et al., 2025). Network stitching provides a functional test of whether representations are interchangeable (Bansal et al., 2021); our comparisons measure geometry instead. Token-level comparisons with non-linguistic spaces follow work on colour and vision (Abdou et al., 2021; Merullo et al., 2023; Sharma et al., 2024).

Measurement. Gröger et al. (2026) identify dimension-dependent chance baselines for global alignment measures and inflation from layer selection, and show that calibration can remove apparent global convergence. We apply their calibration across sample sizes and separately examine how global alignment measures (CKA) depend on width (Kornblith et al., 2019; Davari et al., 2023). Koepke et al. (2026) question both gallery-size robustness and continued capability scaling. We treat the representation and its comparison metric as separate evaluation choices (Sucholutsky et al., 2025).

Codes and language. Representations of medical codes have long been learned from patient records. Med2Vec learns them from codes that occur within and across visits (Choi et al., 2016), BEHRT trains a transformer on longitudinal diagnosis histories (Li et al., 2020), and cui2vec combines concept embeddings from insurance claims, clinical notes and biomedical articles and evaluates them against known clinical relations (Beam et al., 2020). These studies assess code representations against curated relations; they do not ask whether separately trained record and language models converge as they scale. The closest comparison is Noori & Manda (2026), who find weak agreement between clinical co-occurrence and text similarity in MIMIC-IV notes. We add a trained sequence model, scaling axes and measured baselines. Other EHR encoders represent codes through their text descriptions (Su et al., 2025; Hegselmann et al., 2026); Liu et al. (2026) instruction-tune an LLM on life courses written as text; and Tchuitcheu et al. (2026) apply PRH terminology to tables. Our comparison instead uses separately trained models linked at evaluation by event descriptions.

3 SETUP

Life models and event representations. We train autoregressive transformers to predict the next token in chronological registry histories (App. B). A history is a sequence of records, such as a hospital contact or a year's income, and each record becomes one or more tokens. To represent an event code, we average a block's hidden states over the occurrences of that code in a sample of 20,000 lives; these contextual vectors carry information from the preceding events. Static input embeddings, which represent each code independently of its surroundings, serve as a supplementary representation. The seven life models share a corpus and vocabulary and range from 10k to 500M non-embedding parameters; width grows from 16 to 1,600 while depth varies between 2 and 16 blocks (App. B). Validation loss falls with size, and the gains diminish at the largest sizes (App. F).

Matched registry sample. The primary comparisons use the same 1,024 codes. Each code belongs to one of 43 coding systems, such as ICD-10 diagnoses, prescribed drugs, occupations or industries. Codes are paired with their English descriptions. This sample size matches the original PRH and Aristotelian image–language experiments (Huh et al., 2024; Gröger et al., 2026), and neighbours are found within the sample.

Language models and performance. We compare the life models with two cohorts of large language models (LLMs). The *PRH cohort* consists of the 12 LLMs of the original PRH experiment: five BLOOMZ checkpoints (0.56B–7.1B), three OpenLLaMA checkpoints (3B, 7B and 13B) and four LLaMA checkpoints (7B–65B). Their language performance is the authors' published score, $1-$bits per byte on about 4M OpenWebText tokens, so higher means better text prediction. The *HellaSwag cohort* consists of 25 dense LLMs from six families with comparable ten-shot HellaSwag normalised-accuracy scores (Zellers et al., 2019). Both cohorts follow the PRH protocol on the same 1,024 codes. We mean-pool each description's hidden states at every LLM layer and preprocess them as in the PRH code. The candidate representations on each side are the individual layers (for life models, blocks) and their concatenation. We compare every LLM candidate with every life-model candidate and report the best-scoring pair, the *best layer pair*. Some supplementary controls instead use text embedders, models trained to embed text rather than predict it.

Local and global alignment. Local alignment (mKNN) is the average fraction of nearest neighbours shared by the two representations,

$$\mathrm{mKNN}(X, Y) = \frac{1}{Nk} \sum_{i=1}^{N} |\mathcal{N}_X(i) \cap \mathcal{N}_Y(i)|. \tag{1}$$

We use $k = 10$ cosine neighbours and exclude the query itself, so a score of 0.20 means that two of a code's ten neighbours are shared on average. Global alignment asks whether the two spaces agree on the similarities between all pairs of events. We measure it with centred kernel alignment (CKA) using a Gaussian (RBF) kernel (Kornblith et al., 2019) with bandwidth $\sigma = 1$ on row-normalised features, the default in the Aristotelian experiment code; App. D gives the formula and the checks with other kernels.

Permutation calibration. We adopt the calibration of Gröger et al. (2026). For each pair of models, 200 permutations shuffle which description belongs to which code while keeping each representation's geometry, and every permutation repeats the entire layer search. The calibrated score is the amount by which the observed score exceeds the 95th percentile of these shuffled scores, rescaled so that perfect agreement stays at one; scores at or below that threshold become zero.

4 Results

4.1 Calibrated local alignment increases with language performance

Figure 1 reproduces the local scaling pattern of the original Platonic Representation Hypothesis experiment: local alignment increases with life-model size and language performance. For every one of the 12 LLMs of the PRH cohort, calibrated local alignment (mKNN) increases at each of the six steps from the 10k to the 500M life model. Averaged over the LLMs, it rises from 0.209 to 0.287. Across these LLMs, better language performance is associated with greater alignment at every life-model size, with Pearson correlations of 0.72–0.89. Calibration lowers the scores only slightly and leaves this pattern intact (panel b).

Figure 2 shows that both spaces group event codes by domain, the coarsest structure they share; Section 5 tests how much of the alignment goes beyond such categories.

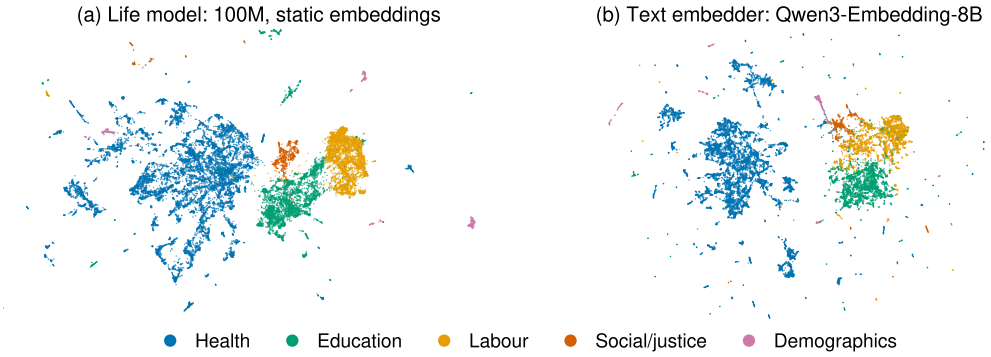


Figure 2: Event-code representations in a life model and a text embedder. Both panels show the same 23,657 event codes coloured by domain. (a) Static embeddings from the three-epoch checkpoint of the 100M life model. (b) Qwen3-Embedding-8B embeddings of the corresponding descriptions. Two-dimensional layouts (UMAP) are fitted directly on the full-width embeddings, independently in each space (cosine metric, 15 neighbours).

4.2 The HellaSwag cohort strengthens the association

Following the broader capability comparison in Huh et al. (2024), we repeat the protocol with the HellaSwag cohort, keeping the codes, preprocessing, layer search and calibration unchanged (App. Fig. 7). Calibrated local alignment correlates strongly with HellaSwag accuracy at every life-model size ($r = 0.88$–0.93). These associations are stronger than in the PRH cohort and change little when any single LLM is omitted. Every LLM aligns more with the largest life model than with the smallest, although for some, alignment dips slightly at intermediate steps.

4.3 LOCAL ALIGNMENT FOLLOWS LIFE-MODEL VALIDATION LOSS

Local alignment tracks the life model's validation loss more closely than its parameter count. Figure 3 compares each run's mean local alignment with its parameter count and validation loss, including seed replicates. From 100k upwards, local alignment follows an almost linear relationship with validation loss in both cohorts ($r = -0.99$). A decrease of 0.1 in loss corresponds to about 0.01 more calibrated local alignment. The 10k model aligns more than this relationship predicts, so the fit applies only from 100k upwards.

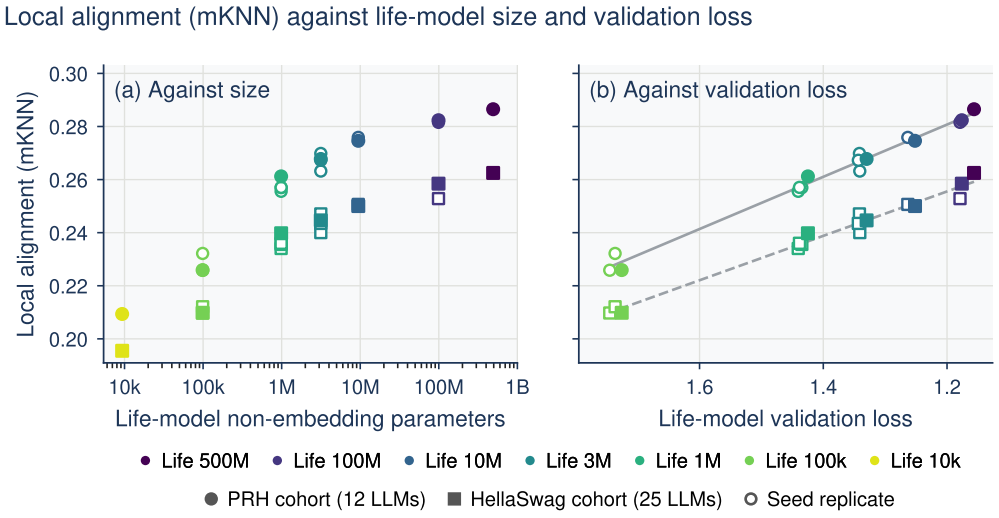


Figure 3: Local alignment against life-model size and validation loss. Each point averages one life-model run's calibrated local alignment (mKNN) over an LLM cohort. Circles represent the PRH cohort and squares the HellaSwag cohort. Colours indicate life-model size, and open markers indicate seed replicates. (a) Against non-embedding parameters; (b) against best validation loss, with better models to the right. Panel (b) omits the 10k runs (validation loss 2.59); its lines are least-squares fits from 100k to 500M.

At a fixed parameter budget (App. B), a deeper model matches the reference in both loss and alignment, while a single-block model predicts less well and aligns less. The small gain in local alignment from 100M to 500M is consistent with the small improvement in validation loss. Both gains diminish on this fixed corpus, suggesting data limitation under the tested training recipe.

4.4 GLOBAL ALIGNMENT IS ABOVE CHANCE BUT STOPS GROWING EARLY

Global alignment (RBF CKA; same models, codes and layer search) is above chance for all 84 model pairs, but it stops growing early. Before calibration, larger life models score higher, particularly compared with the smallest (Fig. 4a). After calibration, the ranges for the 1M to 500M models overlap substantially (Fig. 4b).

Averaged over the LLMs, calibrated global alignment rises up to 1M–3M parameters and then levels off near 0.5, even though validation loss keeps falling (Fig. 5); local alignment, by contrast, keeps following the loss (Section 4.3).

The relationship with language performance depends on the cohort. In the PRH cohort, better LLMs align less with the two smallest life models and more with the two largest (App. Table 2). In the HellaSwag cohort, better LLMs align more at every life-model size ($r \geq 0.84$), including within the Pythia family (App. Fig. 8). Larger LLMs in this cohort also align more: the correlation with log-parameters is almost as strong ($r = 0.81$–0.84). Gröger et al. (2026) also find above-null global alignment for every image–language pair, but only a weak capability trend after calibration. Life models of different shape can differ in how their global alignment relates to language performance even at nearly equal validation loss (App. B).

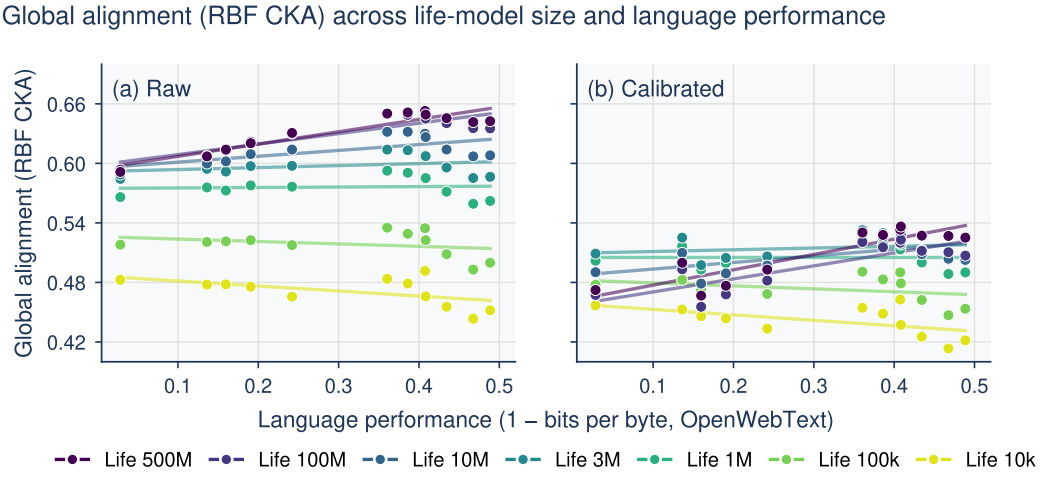


Figure 4: Global alignment (RBF CKA) shows shared structure, but less consistent scaling with language performance than local alignment. (a) Raw and (b) calibrated scores use the PRH cohort and the same 1,024 codes and layer candidates as Figure 1, for all seven life-model sizes. The Gaussian kernel bandwidth is fixed at $\sigma = 1$ on row-normalised features. Each of 200 permutations repeats layer selection. Panels share axis limits, and lines are pooled linear fits. After calibration the larger life models lie closer together, and all remain above chance.

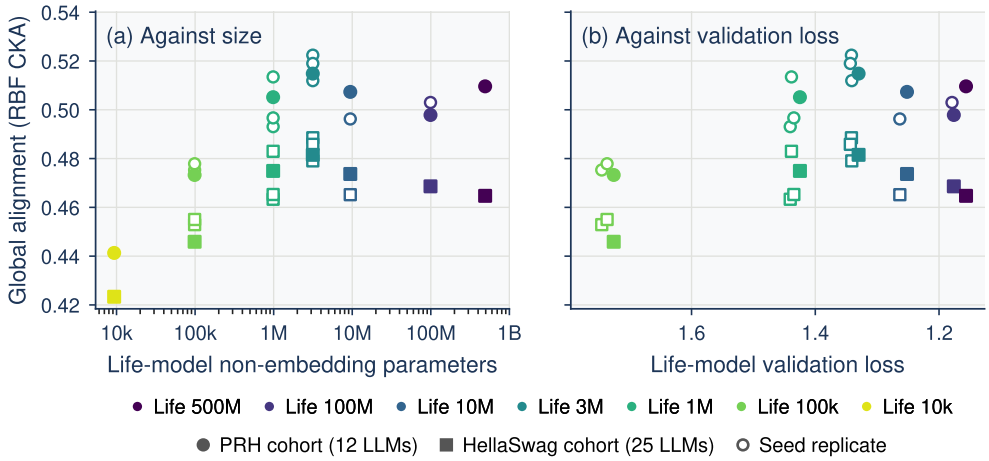


Figure 5: Global alignment against life-model size and validation loss, as in Figure 3. Each point averages one life-model run's calibrated global alignment (RBF CKA) over the LLMs of a cohort: circles for the PRH cohort, squares for the HellaSwag cohort. Open markers indicate seed replicates. (a) Against non-embedding parameters; (b) against best validation loss, with better models to the right, omitting the 10k runs.

5 WHAT DO THE REGISTRY CONTROLS ESTABLISH?

Could coding categories, similar wording or simple co-occurrence explain the agreement? Registry codes belong to categories, and their descriptions often contain category names. The permutation calibration of Section 3 leaves this shared structure in place. We test these explanations below, then examine how the results depend on the representation and code sample.

Table 1 summarises the category, wording and dimension controls for the PRH cohort. Each entry measures whether better language performance is still associated with greater local alignment after a control. The association remains positive for life models of 1M parameters and above under every control in this table. It is less stable for the smallest models, whose local alignment outside coding categories is close to the null.

Agreement beyond coding categories. Removing coding categories weakens the association with language performance only for the smallest life models. We repeat the primary comparison with permutations restricted to codes of the same coding system, with neighbours drawn only from other systems, and, for diagnoses, with finer ICD categories (App. Table 4). For life models of 1M parameters and above, better LLMs still align more under every restriction in both cohorts, and in the HellaSwag cohort the association stays strong from 100k upwards. The smallest life models align little outside their own categories, so there is little left for language performance to explain. Their correlations become unstable: in the PRH cohort they are weak or negative under some restrictions, and in the HellaSwag cohort the 10k model's association nearly disappears when neighbours come from other coding systems.

Growth beyond shared wording. Life models never read descriptions, yet their neighbourhoods partly agree with a character-level TF-IDF representation of those descriptions, while untrained models show little such agreement (App. C). We use this TF-IDF representation as the wording baseline.

We separate each LLM's neighbours into those also found by TF-IDF and those it misses. For each group, we measure the fraction that the life model also finds among its neighbours. Recovery of both groups improves with life-model size, with a larger proportional increase for neighbours that wording does not predict. In the PRH cohort, recovery from 10k to 100M grows by 31% for TF-IDF neighbours, compared with 48% for neighbours that TF-IDF misses and that also share no content word with the query, like the knitting-machine pair of the introduction. Removing character and word TF-IDF similarity also leaves a strong association with language performance (Table 1). On a larger set of 15,933 codes, text embedders also gain more over TF-IDF as the life model grows, levelling off at the largest sizes.

Agreement beyond co-occurrence counts. To test what simple registry statistics capture, we build count baselines from every life of the training cohort (App. E). The strongest baseline records, for each code, how much more often each other token appears in the next record than chance predicts: a first-order Markov chain over records, stored as positive pointwise mutual information (PPMI; Church & Hanks, 1990; Bullinaria & Levy, 2007). These counts align with LLMs about as well as the 100k life model and better than the 10k model. In the HellaSwag cohort, better LLMs also agree more with the counts, in both local and global alignment. Every life model from 1M parameters upwards aligns better than the counts, for every LLM in both cohorts. Between 1M and 100M, the margin over the counts nearly doubles, to 0.05 in the PRH cohort, or about half an extra shared neighbour in ten. The margin is larger for better LLMs from 3M upwards in the HellaSwag cohort, but clearly so only at 100M in the PRH cohort.

Dependence on the representation and gallery. The increase in alignment with size holds for contextual vectors. Static input embeddings behave differently. On matched codes, the 3M model's static embeddings align slightly more than its contextual vectors, but across all codes with descriptions their alignment peaks at 3M instead of rising with size (App. Table 10). The sampled codes also matter: resampling keeps the language-performance associations positive, yet changes their strength substantially in the PRH cohort (App. Table 5). Capping representation width preserves the positive associations as well.

Table 1: Does the association with language performance survive each control? Entries give Pearson r between language performance and calibrated local alignment (mKNN) across the 12 LLMs of the PRH cohort (Figure 1). Columns identify life models; rows identify controls. Scores use the same 1,024 codes and 200 permutations, with layer selection repeated in every permutation. The first row repeats the comparison without registry controls. *Null within coding system*: permutations shuffle codes only within their coding system, discounting agreement explained by membership. *Neighbours outside the coding system*: neighbours are drawn only from other coding systems, in every space. *Similarity after removing TF-IDF*: each LLM's similarities are regressed, query by query, on character and word TF-IDF similarity before neighbours are found (App. C). *Candidates capped at 64 PCs*: representations wider than 64 dimensions are projected onto 64 principal components before the layer search. Full results: App. Tables 4 and 5.

Control	10k	100k	1M	3M	10M	100M	500M
No restriction (Fig. 1)	0.72	0.77	0.81	0.87	0.89	0.88	0.88
Null within coding system	0.38	0.30	0.65	0.79	0.83	0.83	0.84
Neighbours outside the coding system	-0.88	0.89	0.72	0.88	0.84	0.90	0.89
Similarity after removing TF-IDF	0.97	0.96	0.96	0.98	0.98	0.98	0.98
Candidates capped at 64 PCs	0.78	0.90	0.80	0.86	0.89	0.92	0.92

6 Discussion

Our strongest finding concerns local alignment: larger life models agree more with language models about neighbouring event codes, and better language performance is associated with greater local alignment. Both findings survive calibration that repeats layer selection. The HellaSwag cohort extends this result to 25 LLMs (Section 4.2): better performance is associated with greater local alignment at every life-model size tested, so the pattern holds beyond the PRH cohort and its benchmark.

The models also share global structure, but it scales less consistently. Along the life-model axis, calibrated global alignment barely changes from 1M parameters upwards, while validation loss keeps falling. Along the language axis, it grows with language performance in the HellaSwag cohort almost as strongly as local alignment does. In the PRH cohort the association turns from negative for the smallest life model to strongly positive for the largest, and at a fixed budget it changes with life-model shape. In the HellaSwag cohort, the association also appears for LLM size and for the co-occurrence counts (Sections 4.4 and 5), so better LLMs agree more with the global geometry of any registry representation, trained or not. Agreement about fine-scale distances is weak and depends on which codes are sampled (App. D). This matches the Aristotelian account, in which models agree more on neighbours than on detailed distances.

Unlike in vision–language studies, the two sides here share no training data, so their agreement cannot come from shared inputs. From 1M parameters upwards, the life models agree with language models on neighbours beyond what wording, coding categories and co-occurrence counts reproduce, and this extra agreement grows with life-model size.

7 Limitations

The main limitation is data. All life models are trained on the registry histories of one country, about 10 million people, and the 500M model improves validation loss by only 0.020 over the 100M model, consistent with data limitation under the tested training recipe (App. F). Training on more data would test whether further improvements in prediction bring greater alignment. Registries from other countries or larger populations could support this test.

Our controls address specific explanations. The permutation calibration only tests whether codes and descriptions are matched better than chance; shared wording and coding categories are handled by the separate controls of Section 5, and institutional structure that we did not measure may remain. On the language side, performance rises with model size in both cohorts, so our correlations do not separate capability from size: in the HellaSwag cohort, calibrated local alignment correlates as strongly with log-parameters ($r = 0.89$–0.92).

The results are also specific to Danish registries and their descriptions; extending them to other populations or descriptions requires further tests.

Reproducibility statement

The anonymous supplementary material contains the alignment code, null-calibration procedures, result tables and the source data of the figures. Registry microdata cannot be shared, so life-model retraining requires access through Statistics Denmark.

Ethics statement

Danish registry microdata were accessed through Statistics Denmark's research scheme, in compliance with GDPR and the Danish Data Protection Act. Individuals are pseudonymised. All model training and forward passes over person-level sequences took place inside the approved environment. Model weights are not released.

AI use statement

We used an LLM coding assistant for research ideation and execution, including summarising result tables, creating figures, summarising literature, and drafting the manuscript. The authors reviewed the AI-assisted work and take responsibility for the final text and claims.

References

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 the 25th Conference on Computational Natural Language Learning (CoNLL)*, pp. 109–132, 2021. arXiv:2109.06129.

Maria Antoniak and David Mimno. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119, 2018.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 225–236, 2021. arXiv:2106.07682.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing*, volume 25, pp. 295–306, 2020. arXiv:1804.01486.

John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, 2007.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1495–1504, 2016. arXiv:1602.05568.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.16156.

Kshitij Duraphe, Michael J. Smith, Shashwat Sourav, and John F. Wu. The Platonic universe: Do foundation models see the same sky? In *NeurIPS 2025 Workshop on Machine Learning and the Physical Sciences*, 2025. arXiv:2509.19453.

Sathya Edamadaka, Soojung Yang, Ju Li, and Rafael Gómez-Bombarelli. Universally converging representations of matter across scientific foundation models. *arXiv preprint arXiv:2512.03750*, 2025. Oral at the NeurIPS 2025 Workshop on Unifying Representations in Neural Models (UniReps).

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.14486.

Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *npj Digital Medicine*, 9:530, 2026. arXiv:2502.17403.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 30016–30030, 2022. arXiv:2203.15556.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024. arXiv:2405.07987.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Chaitanya Kapoor, Sudhanshu Srivastava, and Meenakshi Khosla. Bridging critical gaps in convergent learning: How representational alignment evolves across layers, training, and distribution shifts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 155718–155732, 2025. arXiv:2502.18710.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into Plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. arXiv:1905.00414.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pp. 2177–2185, 2014.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Wenhow Li, Chengwei Ma, Hui Xiong, Ying-Cong Chen, and Lei Zhang. Platonic Representation Hypothesis on world models. *arXiv preprint arXiv:2608.23720*, 2026.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10:7155, 2020. arXiv:1907.09538.

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. *Nature Machine Intelligence*, 8(5):830–840, 2026. doi:10.1038/s42256-026-01235-7; arXiv:2512.05349.

Samuel Liu, Muchen Xi, William Yeoh, and Joshua J. Jackson. LifeSentence: Language models can encode human life course trajectories from longitudinal panel data. *arXiv preprint arXiv:2606.11220*, 2026.

Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11:5725, 2020.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.15162.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 50358–50376, 2023. arXiv:2305.16264.

Ali Noori and Prashanti Manda. Semantic analysis of SNOMED CT concept co-occurrences in clinical documentation using MIMIC-IV. *Digital Health*, 12:20552076261436869, 2026. arXiv:2509.03662.

Open LLM Leaderboard. Open LLM Leaderboard: Archived evaluation results, 2024. URL `https://huggingface.co/datasets/open-llm-leaderboard-old/results/tree/23474373f8874f9057d23b97e5a41e911d2721c5`. Accessed September 8, 2026.

Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. Using sequences of life-events to predict human lives. *Nature Computational Science*, 4:43–56, 2024. arXiv:2306.03009.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14410–14419, 2024. arXiv:2401.01862.

Statistics Denmark. Documentation of statistics and variables: the metadata bank (formerly TIMES). `https://www.dst.dk/da/Statistik/dokumentation/Times`, 2026a. Variable and classification documentation for Danish register data; documentation of research data at `https://www.dst.dk/da/TilSalg/data-til-forskning/generelt-om-data/dokumentation-af-data`.

Statistics Denmark. Databank of basic data. `https://www.dst.dk/en/TilSalg/data-til-forskning/generelt-om-data/grunddatabanken`, 2026b. Accessed September 13, 2026.

Statistics Denmark. Registers and reference types. `https://www.dst.dk/en/TilSalg/data-til-forskning/generelt-om-data/registre-og-referencetyper`, 2026c. Accessed September 13, 2026.

Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57030–57045, 2025. arXiv:2502.04397.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. arXiv:2310.13018.

Willy Carlos Tchuitcheu, Tan Lu, and Ann Dooms. Towards Platonic representation for table reasoning: A foundation for permutation-invariant retrieval. *arXiv preprint arXiv:2604.12133*, 2026.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2092–2102, 2018.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4791–4800, 2019. URL https://aclanthology.org/P19-1472/.

A PRH-STYLE COMPARISON: PROTOCOL, CONTROLS AND ROBUSTNESS

This appendix separates three checks: whether the primary result depends on the layer search, whether it survives registry-specific restrictions, and how much it varies across code samples and training runs. The primary sample contains 1,024 codes drawn without replacement from the contextual subset with NumPy RandomState(0). Token identifiers and descriptions are fixed before scoring and shared across all models and metrics. This preserves the sample count and $k = 10$ of the original cross-modal PRH comparison, with registry code–description pairs in place of WIT image–caption pairs; absolute scores remain specific to each dataset.

The PRH cohort consists of BLOOMZ-560M/1.1B/1.7B/3B/7.1B, OpenLLaMA-3B/7B/13B and original LLaMA-7B/13B/33B/65B. The checkpoint named `huggyllama/llama-30b` is labelled 33B in the original figure. All 12 checkpoints use unquantised half-precision inference with no description truncation; revisions and precision are recorded per model. Mean pooling masks padding tokens, and the candidate set includes the embedding state, every transformer layer and their concatenation. Life candidates are contextual block outputs and their concatenation, using the same final layer normalisation at each block. The final-block caches agree with independently extracted contextual vectors up to half-precision rounding.

Clipping uses the exact global 95th-percentile absolute feature magnitude across sampled codes and layers, followed by row L2 normalisation of each candidate. The PRH paper includes concatenated features among its comparisons; the released scoring loop only explicitly iterates over saved layers. The same candidate set is used for mKNN and RBF CKA, but each metric selects its own maximum.

For each metric, the same 200 permutations are applied consistently across all layer pairs, and the maximum is recomputed within each permutation. The calibrated score is $s_{\mathrm{cal}} = \max\{0, (s - \tau_{.95})/(1 - \tau_{.95})\}$, where the threshold $\tau_{.95}$ is the $\lceil .95(K+1) \rceil$th order statistic of the observed and null scores, with $K = 200$. For all seven life-model sizes, all 84 observed pairings exceed every sampled permutation maximum for both metrics, giving add-one $p = 1/201$ per pairing; these also pass Benjamini–Hochberg correction across the 84 pairs within each metric. The calibrated score measures alignment above the permutation threshold on a scale where perfect agreement is one. The unrestricted correspondence null does not condition on lexical similarity or coding categories.

Table 2: Language-performance correlations in the PRH cohort. Columns report Pearson r with published OpenWebText performance. Best-layer scores include concatenation. 100M and 500M are the best checkpoints of five- and three-epoch runs (App. A).

	Best-layer mKNN		Best-layer RBF CKA	
Life model	Raw	Calibrated	Raw	Calibrated
10k	.730	.721	-.525	-.542
100k	.780	.774	-.288	-.330
1M	.812	.809	.063	.000
3M	.875	.873	.288	.210
10M	.888	.887	.641	.592
100M	.884	.882	.874	.828
500M	.884	.881	.924	.890

The best-layer mKNN slope remains positive after omitting any one language checkpoint. Fixing the middle language-model layer and final life-model layer gives raw performance correlations of

-0.08, 0.11, 0.43, 0.43, 0.52 and 0.52 across the six sizes through 100M. Repeating the layer search in each permutation corrects its chance advantage; the selected and fixed-layer comparisons still evaluate different representations. The RBF implementation was checked against independent float64 calculations of centring, concatenation, permutation indexing and self-alignment, including an actual 1,024-code comparison, which differed by less than 5×10^{-8}.

Mean calibrated global alignment (RBF CKA) stays between 0.49 and 0.52 from 1M to 500M (Fig. 5), peaking at 3M (0.515, compared with 0.510 at 500M). For comparison, Gröger et al. (2026) report a pooled CKA capability correlation that falls from 0.83 to 0.35 after calibration. Their pooled capability ranking differs from our performance measure and per-life-model correlations.

Agreement within each modality. Table 3 scores pairs of language models and pairs of life models on the same 1,024 codes, with the same clipping, layer search and permutation calibration as the cross-modal comparison. The 12 LLMs share 0.688 of their neighbours on average (0.575–0.827), more within a family (0.780) than across families (0.651). Life models of 1M parameters and above agree at 0.563–0.750, around the agreement of independently trained runs of the same size (0.625–0.695), and the 100M and 500M models agree most (0.750); pairs involving the 10k or 100k model agree less (0.376–0.575). Cross-modal alignment thus reaches about 0.3 of the within-modality level at 10k and 0.4 at 100M and 500M (0.282 and 0.287, against means of 0.64–0.69). Both references benefit from shared inputs, because the LLMs read the same description strings and the life models train on the same histories. Life-model pairs take every size from one 100,000-life extraction sample. Pairing the 20,000-life vectors of Figure 1 with the 100M vectors instead lowers agreement by up to 0.09, whereas swapping the extraction sample changes cross-modal scores by at most 0.004.

Table 3: Local alignment within and across modalities on the primary 1,024 codes (calibrated best-layer mKNN, $k = 10$, 200 permutations repeating the layer search). n counts pairs; means and ranges are over pairs. Life-model pairs use one 100,000-life extraction sample (3M: the seed-73 run); independent runs are the same-size pairs of Figure 6 at 100k, 1M, 3M and 10M. Cross-modal rows repeat Figure 1b.

Pairs of representations	n	Mean	Range
Between language models (12 LLMs)			
all pairs	66	0.688	0.575–0.827
same family	19	0.780	0.721–0.827
different families	47	0.651	0.575–0.735
Between life models			
different sizes, both 1M or larger	10	0.639	0.563–0.750
different sizes, one of them 10k or 100k	11	0.478	0.376–0.575
independent runs of the same size	13	0.658	0.625–0.695
Between life and language models (Figure 1)			
10k life model	12	0.209	0.200–0.213
1M life model	12	0.261	0.246–0.270
100M life model	12	0.282	0.260–0.293
500M life model	12	0.287	0.264–0.303

Training replicates. Independently trained replicates were extracted in the same way at five sizes. At 100k and 1M, we compare the primary five-epoch run with two and three seed replicates, respectively, each trained for three epochs. At 3M, we compare three seed replicates, each trained for three epochs. At 10M and 100M, we compare each primary run with one seed replicate; the replicate validation losses are 1.263 and 1.179, respectively. Figure 6 shows their range; runs that appear in Figure 1 use its values. The widest per-LLM range is 0.011 (100k, LLaMA 13B), and median ranges are 0.001–0.007. The pooled correlation of a single run differs from the main-text value by up to 0.14 (100k).

The validation-loss fits in Figure 3 use 16 runs from 100k to 500M in each LLM cohort. No primary or replicate run lies more than 0.005 in calibrated mKNN from its cohort's fitted line. In the PRH cohort, the 10k model lies above the extrapolated line (0.209 observed versus 0.144 predicted). For global alignment (Fig. 5), the correlation with validation loss over the same range is $r = -0.72$.

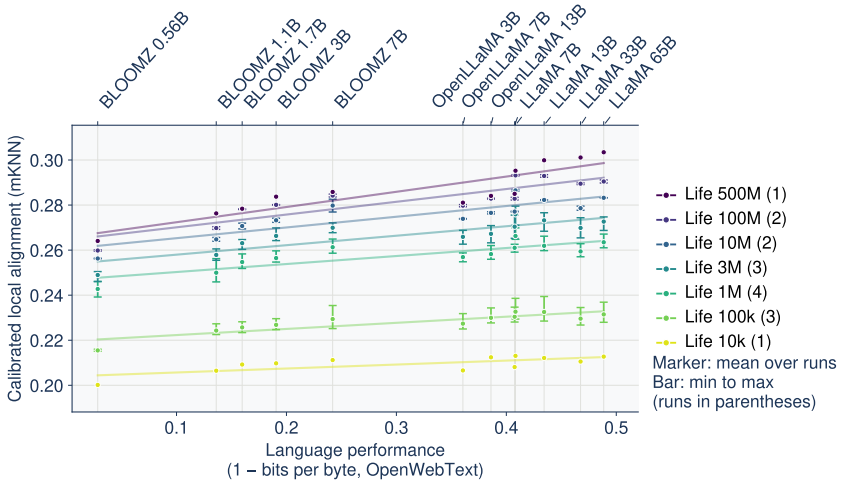


Figure 6: Run-to-run spread of calibrated local alignment in the PRH cohort. Same axes and LLMs as Figure 1b; runs shown in Figure 1 use its values. Markers are means over the independently trained runs of a size (number in the legend) and bars span their range: the primary five-epoch run and three-epoch seed replicates at 100k, 1M and 10M, the primary run and one seed replicate at 100M, three-epoch seed replicates at 3M, and single runs at 10k and 500M. The widest range is 0.011 (100k, LLaMA 13B).

HellaSwag cohort. Figure 7 shows the comparison of Section 4.2; it uses 25 LLMs at the checkpoint revisions of their HellaSwag evaluations. The cohort spans 14M–72B parameters across Pythia, Qwen2.5, Qwen3, OPT, Falcon and SmolLM2.

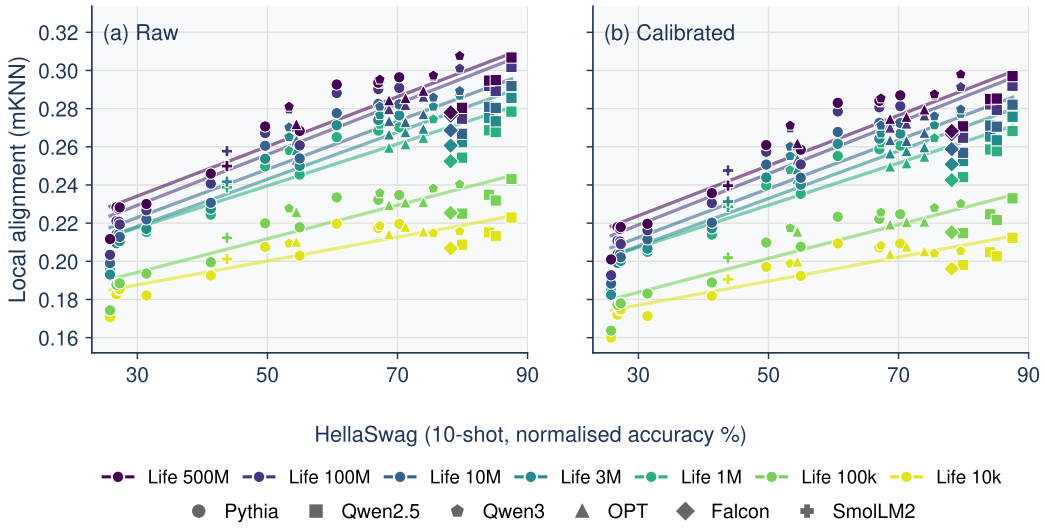


Figure 7: HellaSwag cohort: language performance versus local alignment under the protocol of Figure 1. (a) Raw and (b) permutation-calibrated best-layer local alignment (mKNN) for 25 dense LLMs with comparable ten-shot HellaSwag normalised-accuracy scores and all seven life-model sizes, on the same 1,024 codes as Figures 1–4. Calibration repeats the layer search, including concatenations, in each of 200 permutations. Colours identify life-model size and marker shapes identify LLM family. Panels share axis limits, and lines are pooled linear fits. Benchmark sources are described under *HellaSwag scores* in App. A.

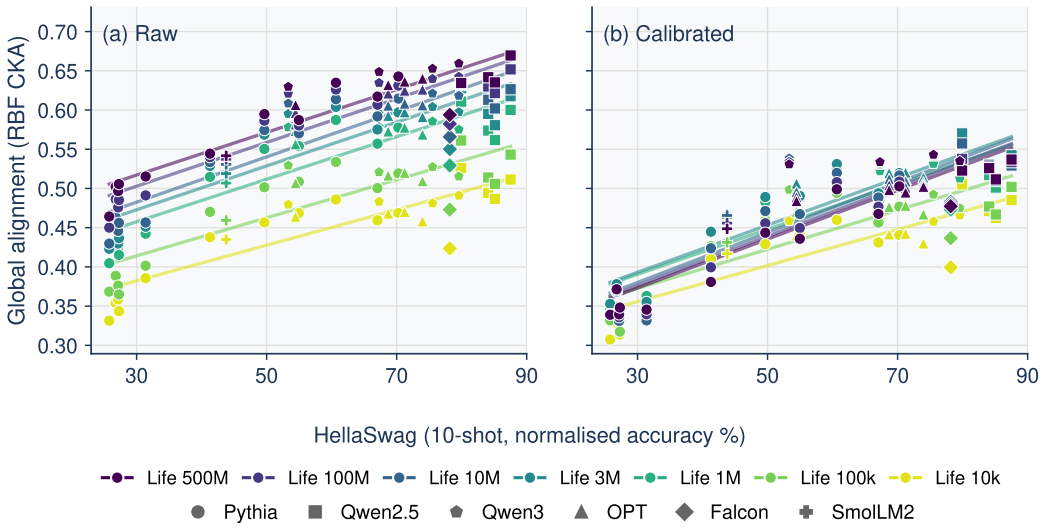


Figure 8: HellaSwag cohort: global alignment (RBF CKA) against language performance, the counterpart of Figure 7. (a) Raw and (b) permutation-calibrated best-layer RBF CKA ($\sigma = 1$ on row-normalised features) for the 25 LLMs and the 1,024 codes of Figure 7; each of 200 permutations repeats the layer search, including concatenations. Colours identify life-model size and marker shapes identify LLM family; lines are pooled linear fits per life-model size.

For 22 of the 25 LLMs, alignment increases at every life-model size step from 10k to 100M (for 21 through 500M). All 25 align more with the 500M model than with the 10k model, and 24 align more with 500M than with 100M. Every hidden state, including the embedding state, is mean-pooled over each description's tokens on the same 1,024 codes, in BF16 on one GPU with CPU and disk offloading for the largest models; no quantisation or truncation is applied. The life-side candidates, clipping, best-pair search and 200-permutation calibration are identical to the original-cohort analysis. Calibrated correlations with HellaSwag are 0.88–0.93 across life-model sizes, and 0.89–0.92 with log-parameters. For global alignment (RBF CKA), they are 0.84–0.91 with HellaSwag, 0.92–0.96 within the 11 Pythia models and 0.81–0.84 with log-parameters. Within-family calibrated correlations are 0.96–0.98 for Pythia (eleven models, including the standard-Pile 70M checkpoint, so this is not a same-training-data estimate), 0.91–0.99 for Qwen2.5, 0.89–1.00 for OPT and 0.65–1.00 for the four Qwen3 models. Leaving out any single LLM keeps every pooled correlation within 0.86–0.94. The change from the full-cohort correlation is at most 0.02. The smallest Pythia models select the input embedding state. Among 119 pairings involving LLMs of at least 1B parameters, 113 select a single layer at 62–97% of model depth and the rest select concatenation. SmolLM2 selects a layer about a third of the way through the model. The optimal representation therefore changes with the language model; it is not consistently the final layer. All 175 calibrated scores exceed their permutation thresholds.

HellaSwag scores. Figure 7 combines archived ten-shot HellaSwag scores (Zellers et al., 2019; Open LLM Leaderboard, 2024) with new evaluations on the same 10,042 archived prompts. Archived scores cover Pythia, Qwen2.5-7B/14B and OPT-1.3B/6.7B/13B. We evaluate Qwen2.5-32B/72B, OPT-30B, Falcon-7B, SmolLM2-135M and the Qwen3 base models (0.6B, 1.7B, 4B and 8B) in BF16, without quantisation, truncation or a sample limit. Tokenisation, answer labels and length normalisation were checked against archived predictions; a Pythia-1B rerun differs by one question. The archived GPT-2 run truncates prompts and is excluded, as are the two Mixtral MoE models.

The HellaSwag cohort spans six families, and eleven of its 25 checkpoints are Pythia models. Family-specific fits address this imbalance, but do not separate size from capability. Checkpoint revisions, prompt hashes, source scores and evaluator details are retained in the supplementary evaluation records.

Registry-specific controls on the same codes. Table 4 repeats both cohort comparisons on the same 1,024 codes under the same best-layer protocol, with the layer search repeated inside every permutation, but with category-conditioned nulls and outside-category neighbours. The permutation null shuffles the code correspondence within coding systems (43 on the full gallery) or within ICD chapters (on the 322 ICD-10 codes of the sample); the neighbour restrictions exclude candidates from the query's coding system, 3-character category or chapter in every space. In the PRH cohort the pooled correlation with language performance remains positive for every life model of 1M parameters and above under every restriction (0.49–0.90); for the 10k and 100k models, whose outside-category alignment is within 0.02 of the null, it is unstable. In the HellaSwag cohort it remains 0.75–0.93 for every life model from 100k upwards and 0.62–0.82 for the 10k model, except outside the coding system (0.16); within families it is 0.77–1.00 under the within-system null. A description-only character TF-IDF baseline scored under the same protocol reaches raw mKNN of 0.047–0.106 outside the chapter on the ICD gallery against 0.062–0.205 for the LLMs; because the baseline is one number per life model and restriction, the increment over it has the same pooled correlation as the LLM score.

Table 4: Pooled Pearson correlations of calibrated best-layer mKNN with language performance in the primary cohorts, on the same 1,024 codes, under category-conditioned nulls and outside-category neighbour restrictions (200 permutations, layer search repeated in every permutation). The 100M and 500M columns use the checkpoints of App. A.

Neighbours	Null	10k	100k	1M	3M	10M	100M	500M
PRH cohort (12 LLMs)								
all 1,024 codes	unrestricted	0.72	0.77	0.81	0.87	0.89	0.88	0.88
all 1,024 codes	within coding system	0.38	0.30	0.65	0.79	0.83	0.83	0.84
outside the coding system	within coding system	-0.88	0.89	0.72	0.88	0.84	0.90	0.89
ICD codes (322)	within chapter	-0.28	0.50	0.49	0.64	0.68	0.70	0.70
ICD, outside 3-char category	within chapter	-0.35	0.60	0.55	0.69	0.70	0.74	0.74
ICD, outside chapter	within chapter	0.74	0.53	0.56	0.70	0.73	0.61	0.69
HellaSwag cohort (25 LLMs)								
all 1,024 codes	unrestricted	0.88	0.92	0.92	0.92	0.93	0.92	0.93
all 1,024 codes	within coding system	0.82	0.91	0.91	0.92	0.93	0.93	0.93
outside the coding system	within coding system	0.16	0.75	0.87	0.89	0.91	0.93	0.93
ICD codes (322)	within chapter	0.72	0.86	0.90	0.89	0.91	0.90	0.91
ICD, outside 3-char category	within chapter	0.66	0.85	0.90	0.88	0.89	0.90	0.90
ICD, outside chapter	within chapter	0.62	0.84	0.86	0.88	0.89	0.90	0.90

Gallery sampling and dimension caps. Table 5 draws 20 random sub-galleries of 512 codes from the sample and recomputes clipping, neighbours and the layer selection inside each. All raw language-performance correlations remain positive, but the PRH cohort is more sensitive to which codes are sampled than the HellaSwag cohort. The table reports the means and ranges at each life-model size. Calibration on the first five sub-galleries preserves this pattern: mean correlations range from 0.85 to 0.93 for the PRH cohort and from 0.89 to 0.92 for the HellaSwag cohort.

We also cap each candidate representation at 256 or 64 principal components before selecting layers, leaving narrower candidates unchanged. Both caps preserve positive calibrated correlations in both cohorts (Table 5). These sampling checks reuse the original codes; independent samples of new codes would require re-extracting every language model.

Sensitivity to the number of neighbours. We repeat the PRH-cohort comparison with $k \in \{5, 10, 20\}$ on the same 1,024 codes and all seven life-model sizes. Each k uses the same primary checkpoints, clipping and candidate layers, including concatenations; the layer search is repeated in each of 200 shared correspondence permutations. Table 6 tests the two scaling patterns in Figure 1: whether better language performance is associated with greater alignment, and whether alignment increases at every life-model size step.

The language-performance association remains positive at every k and life-model size, with r between 0.57 and 0.93; for every size it is highest at $k = 20$, and calibration changes each correlation by at most 0.011. For all 12 LLMs, alignment increases at every life-model size step from 10k to 100M at $k = 10$ and $k = 20$; at $k = 5$ it does so for eight, and the largest stepwise decrease in calibrated mKNN is 0.0016. Thus the language-performance scaling pattern persists across the three neighbourhood sizes, while strict stepwise life-model scaling through 100M holds at $k = 10$ and

Table 5: Gallery-sampling and dimension-cap checks of the primary cohort correlations. Sub-galleries recompute clipping, neighbours and layer selection; PCA is fitted on the 1,024 codes for candidates wider than the cap; narrower candidates are unchanged. Dimension controls use 200 permutations; five sub-galleries use 100, and the remaining sub-galleries report raw scores.

Check	Statistic	10k	100k	1M	3M	10M	100M	500M
PRH cohort (12 LLMs)								
fixed sample	raw r	0.73	0.78	0.81	0.88	0.89	0.88	0.88
20 sub-galleries of 512	mean raw r	0.86	0.86	0.82	0.84	0.88	0.90	0.85
	min–max	0.40–0.97	0.59–0.98	0.47–0.94	0.62–0.97	0.72–0.96	0.75–0.96	0.68–0.94
PCA to 256 dimensions	calibrated r	0.68	0.84	0.76	0.88	0.92	0.90	0.93
PCA to 64 dimensions	calibrated r	0.78	0.90	0.80	0.86	0.89	0.92	0.92
HellaSwag cohort (25 LLMs)								
fixed sample	raw r	0.88	0.92	0.92	0.92	0.92	0.92	0.93
20 sub-galleries of 512	mean raw r	0.89	0.91	0.90	0.91	0.92	0.91	0.91
	min–max	0.83–0.91	0.86–0.93	0.87–0.93	0.88–0.93	0.90–0.95	0.89–0.94	0.90–0.94
PCA to 256 dimensions	calibrated r	0.87	0.91	0.91	0.92	0.92	0.93	0.93
PCA to 64 dimensions	calibrated r	0.87	0.91	0.91	0.92	0.92	0.92	0.92

$k = 20$ but not at $k = 5$. All 252 calibrated scores across the three neighbourhood sizes exceed their permutation thresholds ($p = 1/201$). The $k = 5$ and $k = 20$ values are recomputed on CPU using prefixes of the same cosine ordering.

Table 6: Language-performance correlations at different neighbourhood sizes. For each life-model size, entries give Pearson r between OpenWebText performance and raw or calibrated mKNN across the same 12 LLMs. All comparisons use the same 1,024 codes and one primary run per life-model size, with layer selection repeated in each calibration permutation at each k. The $k = 10$ columns repeat the Figure 1 correlations of Table 2.

	$k = 5$		$k = 10$		$k = 20$	
Life model	Raw	Calibrated	Raw	Calibrated	Raw	Calibrated
10k	0.727	0.716	0.730	0.721	0.929	0.928
100k	0.576	0.571	0.780	0.774	0.929	0.926
1M	0.819	0.813	0.812	0.809	0.870	0.869
3M	0.698	0.691	0.875	0.873	0.917	0.914
10M	0.828	0.825	0.888	0.887	0.899	0.898
100M	0.846	0.841	0.884	0.882	0.929	0.926
500M	0.845	0.842	0.884	0.881	0.928	0.927

B REGISTRY DATA, LIFE MODELS AND REPRESENTATION CHOICES

Registry life histories. The source corpus links administrative records for 10 million people in Denmark, spanning the 1970s through 2023. Coverage begins at different dates across registries and varies with each person's observed history. The records describe five broad domains:

- *Health*: hospital contacts and diagnoses, including ICD-10 codes; visits to general practitioners and other healthcare providers; and prescription medications classified by ATC codes.
- *Education*: enrolment, qualifications and graduation, alongside school and higher-education grades, national tests, attendance and student wellbeing.
- *Labour*: employment status, occupations and industries, earnings, income and labour-market support.
- *Demographics*: births and deaths, family relationships, marital status, residential moves, immigration and emigration.
- *Social/justice*: charges, legal decisions, victim records, imprisonment and parole, out-of-home placements, preventive social interventions and substance-use treatment.

Records from these sources are combined chronologically for each person. A life history includes both dated events and periodic observations, such as annual income; these are different forms of

registry information (Statistics Denmark, 2026c). Categorical values are represented by discrete tokens, and numerical quantities such as income are discretised into bins. Age and calendar-year tokens place the observations in time. An individual record can contribute several tokens, so predicting the next token includes learning relationships within records as well as across successive events. For contextual-representation extraction, we use an analysis sample of 20,000 lives from the validation cohort (App. B).

Life-model architecture and objective. The life models are decoder-only transformers trained on these token sequences. A learned input table maps each token to a vector, which passes through a stack of transformer blocks. Each block combines causal multi-head attention with a feed-forward network, using residual connections and layer normalisation; rotary position embeddings encode sequence order. Causal attention lets a token use preceding observations in the sequence. The output layer shares its weights with the input embedding table and predicts the next token, with cross-entropy as the training objective. No event descriptions or text-model representations enter this training.

This architecture provides two ways to represent an event code. Its *static embedding* is its learned input vector, which is the same at every occurrence. Its *contextual representation* is a hidden state after a transformer block and depends on the preceding history. We average these states over occurrences to obtain one contextual vector per code; the pooling and layer-selection procedures are detailed below. Increasing model size changes the width of the vectors and the number of blocks. The 10k, 100k, 1M, 3M and 10M models use (width, blocks) of $(16, 3)$, $(64, 2)$, $(128, 5)$, $(256, 4)$ and $(256, 12)$, respectively. The 100M and 500M models use $(768, 14)$ and $(1{,}600, 16)$ (App. A).

Analysis vocabulary and training series. The training vocabulary contains 23,657 tokens; the full comparison uses the 23,657 event codes with descriptions. Filtering and description provenance are recorded in the supplementary data. The 1,024-code primary sample is drawn from the contextual subset below. Two larger analysis sets support the additional controls; because mKNN depends on the number of candidate neighbours, their absolute scores are not directly comparable with those on the 1,024-code sample:

- *Full set*: all 23,657 event codes used for static alignment.
- *Contextual subset*: 15,933 full-set codes occurring at least five times in the sample of 20,000 lives used to compute contextual vectors.

Description sources are Statistics Denmark's variable and classification documentation (Statistics Denmark, 2026a; Savcisens et al., 2024) and official WHO, ISCO and NACE titles. The alternative-description control compares registry translations with distinct official titles on matched codes (App. C). The supplementary scaling comparison uses six sizes from 10k through 100M, with the three-epoch checkpoint of the 100M run (validation loss 1.181); a second 100k checkpoint supports the training comparisons below. All seven sizes, including 500M, enter the primary neighbourhood and CKA comparisons and the lexical-baseline comparison. Other supplementary analyses on larger code sets stop at 100M because the 500M static table and final-block states were not exported for those sets. Comparisons with description embedders select a life-model block while holding the text representation fixed; controls using a fixed life-model layer are labelled separately. All life models use an 8,192-token training context. We use each run's best-validation checkpoint. Table 7 lists the primary checkpoints. Seed replicates, the half-population runs and the width and depth controls were trained for three epochs with 120k tokens per update.

Training, validation and optimisation. Training and validation use a person-level split: the training cohort holds 7.4 million individuals (12.6 billion tokens) and the validation cohort 2.4 million (4.2 billion tokens); the cohorts share no individual and span the same birth years, so the split is not temporal. Validation loss, and with it checkpoint selection, is computed on the validation cohort. All runs use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with weight decay 0.1 on weight matrices and the embedding table, a peak learning rate of 3×10^{-4} reached by linear warmup over the first 1.5 epochs and then held constant, dropout 0.1 and 16-bit mixed precision.

Life-model scale. Supplementary controls use text embedders, models trained to embed descriptions rather than predict text. The five core embedders are MiniLM-L6, nomic-embed-text-v1.5,

Table 7: Primary life-model checkpoints. Updates per epoch are derived from the step count of the selected checkpoint; the 3M model used 180k tokens per update and the other runs 120k.

Size	Width	Blocks	Heads	Non-emb. params	Epochs (selected)	Updates/epoch	Val. loss
10k	16	3	2	10k	5 (5)	107k	2.591
100k	64	2	4	100k	5 (4)	107k	1.726
1M	128	5	4	1M	5 (5)	107k	1.425
3M	256	4	4	3M	5 (4)	71k	1.330
10M	256	12	4	10M	5 (5)	107k	1.252
100M	768	14	12	100M	5 (4)	109k	1.176
500M	1,600	16	25	500M	3 (3)	114k	1.156

gte-multilingual-base, Qwen3-Embedding-0.6B and gte-Qwen2-7B-instruct (hereafter 7B). Contextual mKNN increases at every size step against nine text embedders: the five core models plus paraphrase-MiniLM-L3-v2, gte-Qwen2-1.5B and Qwen3-Embedding-4B/8B. The trend survives selection-aware calibration for all five core embedders (App. B). Against 7B, raw contextual local alignment on the 15,933-code subset rises from 0.144 at 10k to 0.288 at 100M. Static local alignment on the full 23,657-code set instead peaks at 3M (0.203), then reaches 0.187 and 0.188 at 10M and 100M. The matched-code comparison below separates representation choice from code coverage. In the primary comparison, training replicates at five sizes preserve the increasing trend in their means (App. Fig. 6); supplementary seed controls are described in App. B.

Supporting evidence across training populations. At the best block, life–life mKNN rises from 0.19–0.27 for pairs involving the 10k model to 0.41–0.50 among the 1M to 100M models; random models agree at ≤ 0.004. Two 3M models trained on disjoint population halves reach 0.43 static and 0.55 in context, close to full-data versus half-data agreement (0.43 and 0.54). Shared training individuals are therefore unnecessary for this within-modality agreement. The models share an architecture, objective and population source (App. B).

Seed variation and training. The completed 100k seed pair differs by 0.0012 in static mKNN and 0.0010 in context on the full set and contextual subset. On the 1,024-code sample of the primary comparison, the same two runs differ by 0.006 on average across the 12 LLMs (App. A), so seed variation grows as the gallery shrinks. These figures come from one size and one training recipe. The 0.033 gap between the original 100k checkpoints coincides with their batch-size difference.

Two 100k seed replicates share 35% of their static neighbours, yet alignment with each of eight text embedders (the five core models, bge-small-en-v1.5 and Qwen3-Embedding-4B/8B) differs by at most 0.0014 across both representations (median 0.0005). As in earlier embedding studies (Wendlandt et al., 2018; Antoniak & Mimno, 2018; Mehrer et al., 2020), reproducible aggregate scores need not imply reproducible individual neighbours.

Across the six sizes of the supplementary series and the second 100k checkpoint, best-block alignment orders perfectly with predictive performance for all nine embedders (Spearman 1.00 with negative validation loss; 0.71 for the static table). Within a single training run, the pattern is weaker. Across 15 checkpoints saved every 20,000 steps of the 3M run, static mKNN approaches its peak by step 120,000 and then falls by 0.0009 while validation loss continues to improve. Training progress and cross-modal agreement therefore need not move together throughout a run.

Width and depth at a fixed budget. Both width and depth vary across the size series. To separate them, we trained three shapes with the same 3M non-embedding parameters, each with three seeds: the reference 3M model (width 256, 4 blocks), a narrow and deep model (128, 16) and a wide and shallow model (512, 1). All runs share the data, optimiser, 120k-token updates, 8,192-token context and three training epochs. They match non-embedding parameters and training tokens, but not total parameters (the tied embedding table adds $23{,}657\times$ width) or compute.

At the 3M budget, depth has little effect: the 16-block model matches the reference in calibrated mKNN (difference -0.0015, against run ranges of 0.007) at a slightly lower validation loss (Table 8). The single-block model has a higher loss ($+0.110$) and aligns less (-0.034); for every one of the 12 LLMs, each of its runs scores below each reference run. Width does not explain the gap.

Table 8: Width and depth controls under the primary protocol (PRH cohort, 1,024 codes, best layer pair, 200 permutations). Calibrated scores are means over LLMs and runs; parentheses give the range of run means; r is the Pearson correlation with language performance, averaged over runs. Total parameters count the tied embedding table once. All runs train for three epochs on identical data.

					mKNN		RBF CKA	
Non-emb.	Width×blocks	Runs	Total	Val. loss	Calibrated	r	Calibrated	r
3M	256×4 (reference)	3	9M	1.342	.267 (.263–.270)	.85	.518	.31
	128×16	3	6M	1.334	.265 (.262–.269)	.87	.491	.01
	512×1	3	15M	1.452	.232 (.227–.235)	.75	.459	.08
13M	256×16	3	18M	1.248	.273 (.272–.273)	.90	.497	.54
	512×4	1	24M	1.263	.274	.85	.522	.62
0.8M	128×4	1	3M	1.456	.254	.72	.486	−.14
	256×1	1	6M	1.553	.224	.73	.444	.08

The widest model aligns least; projecting every life candidate onto 128 principal components leaves the three shapes at 0.270, 0.270 and 0.241; and restricting the reference to single blocks without concatenation leaves a raw gap of 0.026. A different 20,000-life extraction sample lowers every run's score by 0.006–0.017 and preserves these comparisons (single-block gap -0.038, deep minus reference $+0.001$). In the one-axis variants, 16 blocks of width 256 and 4 blocks of width 512 align equally at 13M, whereas at 0.8M the single-block model again aligns less. The association with language performance remains positive for every shape ($r = 0.72$–0.90).

Across all 25 runs trained for three epochs, calibrated mKNN increases as validation loss falls (Spearman 0.92 with negative validation loss), so local alignment tracks predictive performance more closely than parameter count or shape. Single-block models are the exception: at validation losses of 1.44–1.46, the single-block 3M model aligns 0.02–0.03 below the 1M and 128×4 models. Because most one-axis shapes have a single seed, we treat this depth effect as suggestive. Global alignment (RBF CKA) depends more strongly on shape. Calibrated RBF CKA ranks the reference above the deep and wide models, even after the 128-component projection (0.512, 0.494 and 0.464). Its correlation with language performance falls from 0.31 for the reference to 0.01 for the deep model at nearly equal loss, so the sign changes of this association across the size series (Section 4.4) are not a function of size alone.

Layer selection and calibration. Text embedders receive each description with their published prefix and pooling conventions (for example the instruction prefix for nomic, and mean or last-token pooling otherwise), without truncation and in half precision. For the supplementary best-block comparisons, each metric scores all life-model blocks against a fixed text representation and selects its own maximum. Final layer normalisation is applied to each block output before averaging over occurrences in the same 20,000 lives. The cohort comparisons additionally search language-model layers and concatenated candidates. mKNN selects the first life-model block at all six sizes for each of the nine embedders; CKA can select a different block.

Calibration repeats the selection inside every permutation: the same 200 permutations are applied to every block, and the observed maximum is calibrated against the null maximum. Relative to calibrating only the selected block, this changes mKNN by at most 0.00005 and CKA by 0.0012. Calibrated mKNN increases at every size step for all five core embedders. Table 9 gives the raw and calibrated scores against 7B; calibration changes them little. Calibrated CKA against 7B ranges from 0.47 to 0.59 and peaks at 10M.

Table 10 compares mKNN and linear CKA, holding the 15,933 codes fixed when comparing static embeddings with contextual vectors. Static embeddings align more closely in mKNN for every core text embedder (7B: 0.292 static vs. 0.266 contextual); contextual vectors align more closely in CKA.

Table 9: Best-block local alignment (mKNN) with gte-Qwen2-7B-instruct on the 15,933-code contextual subset. Calibration repeats life-block selection in each of 200 permutations.

Life model	Raw mKNN	Calibrated mKNN
10k	0.144	0.143
100k	0.205	0.204
1M	0.244	0.244
3M	0.266	0.265
10M	0.268	0.268
100M	0.288	0.288

Table 10: Contextual vectors vs. static embeddings of the 3M model (raw mKNN and linear CKA per embedder on the identical 15,933-code subset; contextual values at each metric's best block).

Text embedder	$\text{mKNN}_{k=10}$ (raw)		CKA (raw)	
	Static	Contextual	Static	Contextual
MiniLM 33M	0.248	0.226	0.297	0.472
nomic 137M	0.265	0.241	0.316	0.528
gte-multi 305M	0.257	0.233	0.329	0.551
Qwen3 0.6B	0.262	0.236	0.348	0.566
gte-Qwen2 7B	0.292	0.266	0.373	0.576

C WORDING AND CODING-CATEGORY CONTROLS

These controls test how much agreement can be explained by description wording and coding categories. We compare lexical baselines with learned text representations, examine neighbours that the baselines miss, and change the descriptions and allowed neighbour categories.

Neighbours that wording does not predict. Each code enters a life model as an atomic token with its own embedding row. The model cannot read its description or code prefix. In the 18.9 million events of the local registry sample, no event records a code together with a parent code, and none mixes domains. Agreement with wording or code hierarchies must therefore be learned from histories. Trained life models align with both the character TF-IDF baseline and the identifier hierarchy (Table 13); untrained models stay near the null (0.003–0.013).

Language models read the descriptions directly, so wording can explain part of their neighbourhoods. Averaged over layers, a quarter to two fifths of an LLM's neighbours are also TF-IDF neighbours. This fraction reaches about half in the most lexical layer; untrained LLMs with the same tokenisers already reach 34–37%. For each query, we split the LLM's ten neighbours into those among its ten TF-IDF neighbours and those outside that set. We then measure the fraction of each group also found among the life model's ten neighbours. Calibration repeats the layer search in each of 200 permutations, as in the primary comparison. Because the groups depend on the LLM, each model receives its own recovery score.

Recovery increases with life-model size in both groups (Table 11). TF-IDF neighbours are recovered more often, while neighbours that TF-IDF misses show larger proportional gains. For the latter group, raw recovery in the PRH cohort rises from 14.9% at 10k to 20.6% at 100M, compared with 1.2% under the permutation null. The corresponding recovery rate for untrained LLMs is about 7% at every life-model size.

We further exclude pairs sharing a content-word stem, or find neighbours after regressing LLM similarities on character and word TF-IDF and code-hierarchy indicators. These stricter controls retain positive associations with language performance in both cohorts. In the HellaSwag cohort, recovery of TF-IDF neighbours instead correlates negatively with performance at every life-model size. The positive HellaSwag association thus comes from neighbours that TF-IDF does not predict.

Table 12 illustrates shared neighbours with no content word in common: drugs with the same indication, related diagnoses, and an occupation linked to its industry. There are 308 such pairs that

TF-IDF misses but that are neighbours in both the 10k model and at least 9 of the 12 LLMs of the PRH cohort; at 100M there are 457.

Table 11: Life-model agreement with LLM neighbourhoods, by how predictable the neighbours are from wording (1,024 codes, best layer pair, 200 permutations repeating the search). Calibrated scores averaged over the LLMs of each cohort; r: range from 10k to 100M of the pooled Pearson correlation with language performance (OpenWebText $1-$bits per byte, HellaSwag accuracy). Neighbours not identified by TF-IDF make up 59–76% of an LLM's neighbours; those that also share no content word with the query make up 41–61%. Residual: neighbours after regressing each LLM's cosine similarities, query by query, on character and word TF-IDF similarity (and on indicators of the code hierarchy).

	PRH cohort (12 LLMs)				HellaSwag cohort (25 LLMs)			
LLM neighbours scored	10k	100M	Growth	r	10k	100M	Growth	r
TF-IDF neighbours	0.337	0.442	+31%	0.69–0.88	0.362	0.466	+29%	−0.35 to −0.09
All neighbours	0.209	0.282	+35%	0.72–0.89	0.196	0.258	+32%	0.88–0.93
Not identified by TF-IDF	0.137	0.194	+42%	0.77–0.92	0.117	0.160	+36%	0.89–0.92
and no shared content word	0.116	0.172	+48%	0.83–0.94	0.092	0.133	+43%	0.90–0.92
Residual, TF-IDF	0.086	0.127	+48%	0.96–0.98	0.070	0.099	+41%	0.96–0.97
Residual, TF-IDF and hierarchy	0.023	0.048	+110%	0.95–0.96	0.014	0.031	+121%	0.89–0.92

Table 12: Shared neighbours that wording does not predict: the two descriptions share no content word and are not TF-IDF neighbours. Upper block: the codes are neighbours both in the life model and in at least 9 of the 12 PRH-cohort LLMs; From gives the smallest life model from which this holds at every larger size up to 100M, and none of the listed pairs meets this criterion for the 10k model. Lower block: further pairs from different coding systems that are neighbours in the 100M life model and LLaMA-65B; the pairs marked † also meet this criterion. Descriptions from Statistics Denmark's documentation.

Code	Shared neighbour	From
Life model and at least 9 of the 12 PRH-cohort LLMs		
Proton pump inhibitors, lansoprazole	H2-receptor antagonists, ranitidine	100k
Microvascular angina	Atherosclerotic heart disease	100k
Clinically assessed angina pectoris	ST-elevation acute myocardial infarction	100k
Early retirement pension (labour-market status)†	Integration benefit (social benefit)	100k
Migraine with complication	Cluster headache syndrome	1M
Primary hyperparathyroidism	Hypocalcemia NOS	1M
Ovulation stimulants, synthetic, clomifene	Gonadotropins, urofollitropin	1M
Type 2 diabetes with coma	Hyperglycemia NOS	1M
Brickworks (industry, earlier classification)†	Concrete-goods factories (industry)	1M
Pyonephrosis	Postrenal uremia	10M
Monoclonal gammopathy of undetermined significance	Hypogammaglobulinemia NOS	100M
Weaving- and knitting-machine operators (occupation)†	Production of pullovers, cardigans etc. (industry)	100M
Across coding systems, 100M life model and LLaMA-65B		
Retinal detachment without treatment (diagnosis)	Ophthalmology service, fluorescein angiography (hospital service)	
Magnesium, magnesium citrate (prescription)	Hypocalcemia NOS (diagnosis)	
Shipping industry, ferry and passenger transport (industry)	Travel and tourist agencies (industry, earlier classification)	
Transport clerks (occupation)	Travel agency work (occupation, later classification)	

Scaling beyond the lexical baseline. We subtract life–TF-IDF local alignment from life–text local alignment on the 15,933 codes of the contextual subset, using description-only inputs (character TF-IDF fitted on the same descriptions) and identical neighbour restrictions in every space: all neighbours, or neighbours outside the query's coding system (43 systems; Fig. 9). At the best mKNN block, the increment is larger at 100M than at 10k and at 10M for every core embedder under both restrictions, and it rises at every step from 10k to 100M for seven of the ten embedder–restriction curves. Averaged across embedders, it grows from 0.016 to 0.045 with all neighbours and from 0.008 to 0.031 outside the coding system between 10k and 100M; at 500M it is 0.045 and 0.036. We estimate uncertainty by resampling query groups defined by coding system and 3-character code

prefix (2,000 paired bootstrap resamples). The neighbour graphs and trained models remain fixed, so the intervals describe variation across code groups only. The 10k-to-100M changes are 0.029 with a 95% interval of $[0.024, 0.034]$ for all neighbours and 0.023 $[0.020, 0.027]$ outside the coding system. The smaller 10M-to-100M changes are 0.004 $[0.002, 0.007]$ and 0.004 $[0.002, 0.006]$, respectively. Static increments instead peak at 3M for seven of the ten curves.

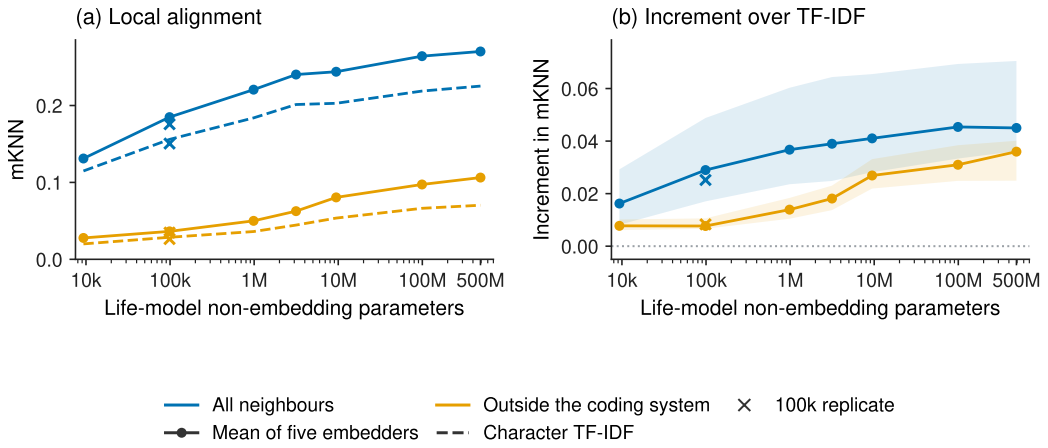


Figure 9: Scaling of contextual vectors at the best block beyond character TF-IDF on the 15,933 codes of the contextual subset ($k = 10$). (a) Mean life–text local alignment across five core embedders (solid) and life–TF-IDF local alignment (dashed), with all neighbours or with neighbours restricted to other coding systems. (b) Their difference; shading spans the five embedder-specific increments. Crosses mark the 100k batch-size replicate.

Results at each life-model size. Table 13 gives the baseline and neighbourhood-recovery scores at all seven sizes on the primary 1,024 codes. The identifier hierarchy uses domain, coding system and shared code prefix. The LLM rows average over the PRH cohort, using the neighbour groups and calibration defined above. Scores improve little from 100M to 500M. The lower block of Table 12 gives further examples shared by the 100M life model and LLaMA-65B across coding systems, with no shared content word and no TF-IDF neighbour match.

Table 13: Structure learned beyond wording on the 1,024 codes of the primary comparison (calibrated scores; LLM rows are means over the 12 PRH-cohort LLMs). The first two rows compare life-model neighbours with neighbours defined by the descriptions' character TF-IDF and by the identifier hierarchy; the other rows give, for each LLM's ten neighbours of a code, the fraction the life model also has, split by whether the TF-IDF baseline predicts them.

	10k	100k	1M	3M	10M	100M	500M
Life model vs. description TF-IDF	.159	.169	.190	.193	.203	.209	.210
Life model vs. identifier hierarchy	.256	.289	.319	.335	.334	.340	.346
LLM neighbours predicted by TF-IDF	.337	.367	.416	.417	.431	.442	.446
LLM neighbours not predicted by TF-IDF	.137	.148	.174	.184	.187	.194	.197
and sharing no content word	.116	.135	.156	.163	.167	.172	.174
Similarity after removing TF-IDF	.086	.101	.112	.120	.125	.127	.130

Word and character patterns. Table 14 separates surface-pattern controls from learned embeddings. Random 300-dimensional word-vector averages encode word overlap; word and character TF-IDF use exact sparse cosine neighbours (SVD-300 for CKA). GloVe and Model2Vec are learned static baselines. Description-only character n-grams reach mKNN 0.157 with the static life table and 0.148 with final-block contextual vectors. The core contrastive embedders reach 0.176–0.203 and 0.165–0.193, respectively. This input-form control uses the final block throughout; the scaling control above selects the best mKNN block. Identifier-only character patterns reach static mKNN

0.187, illustrating the alignment available directly from code prefixes. Adding identifiers raises scores for both learned embedders and surface-pattern controls, without increasing the learned models' advantage.

Across-category agreement. Restricting candidate neighbours among the 7,206 ICD-10 codes tests how much agreement extends across the specified taxonomy groups. Excluding the query's 3-character category lowers description-based life–text mKNN from 0.17–0.20 to 0.09–0.11; excluding its chapter lowers it to 0.06–0.08. Character TF-IDF falls from 0.146 to 0.068 and 0.039. The learned embedders retain an advantage under both restrictions. Identifier-only agreement falls much more strongly, to 0.003–0.04 outside categories and 0.001–0.002 outside chapters.

Alternative wording and language. On 5,076 codes whose unique official WHO, ISCO or NACE title differs from the registry translation, the official titles align at least as well for all seven tested embedders and both life representations, and stripping ICD boilerplate changes static mKNN by at most 0.006. Two multilingual embedders also align above chance with the original Danish descriptions. Calibrating against within-ICD shuffles that preserve chapters retains 97–98% of ICD alignment for MiniLM and 7B, and shuffles that preserve 3-character categories retain 44–47%, so coding categories account for substantial agreement even though cross-category local alignment remains.

Table 14: Local alignment (mKNN) of the 3M life model with each text input: surface-pattern controls (random word-vector averages and exact sparse TF-IDF neighbours), learned static embedders and five core contrastive embedders. "Static" and "contextual" refer to the life table and fixed final-block contextual vectors, using the full set and contextual subset respectively; the text features are fixed. Description-only inputs are the primary comparison. The strong identifier-only character baseline motivates the category restrictions in the text. Random floors: 0.000 (static) / 0.003 (contextual).

Text representation	Identifier		Description		Identifier + description	
	Static	Contextual	Static	Contextual	Static	Contextual
word overlap (random word vectors)	0.014	0.031	0.133	0.133	0.134	0.141
TF-IDF word	0.017	0.031	0.141	0.134	0.150	0.150
TF-IDF char 3–5-gram	0.187	0.171	0.157	0.148	0.194	0.186
GloVe-300d average	0.021	0.044	0.146	0.152	0.152	0.167
Model2Vec potion-8M	0.145	0.155	0.160	0.153	0.193	0.191
MiniLM 33M	0.182	0.191	0.176	0.165	0.193	0.191
nomic 137M	0.179	0.195	0.187	0.176	0.207	0.206
gte-multi 305M	0.097	0.120	0.180	0.169	0.208	0.199
Qwen3 0.6B	0.192	0.198	0.183	0.175	0.215	0.211
gte-Qwen2 7B	0.171	0.182	0.203	0.193	0.226	0.224

D CHANCE CALIBRATION AND MEASUREMENT SENSITIVITY

Calibration asks how much alignment exceeds a specified null; changing the gallery or kernel asks how sensitive the result is to the measurement choice.

For the five core embedders, calibrated mKNN with the reference 3M model is 0.175–0.203 in the static full-set comparison. On the contextual subset it is 0.225–0.265 at the selected block and 0.164–0.192 at the final block. The latter values belong to the fixed-layer controls, not the best-block comparison in Table 10.

Kernel alignment. Global alignment uses centred kernel alignment with a Gaussian kernel (Kornblith et al., 2019). It converts distances between pairs of events into similarities, $K_{ij} = \exp[-\|x_i - x_j\|^2/(2\sigma^2)]$ in one space and L_{ij} analogously in the other. With the centred matrices $K_c = HKH$ and $L_c = HLH$, where $H = I_N - \mathbf{1}\mathbf{1}^\top/N$ and $\mathbf{1}$ is the vector of N ones,

$$\text{RBF CKA}(X, Y) = \frac{\sum_{i,j}(K_c)_{ij}(L_c)_{ij}}{\sqrt{\sum_{i,j}(K_c)_{ij}^2}\sqrt{\sum_{i,j}(L_c)_{ij}^2}}. \tag{2}$$

The numerator measures agreement between the centred similarities, and the denominator normalises for the magnitude of each matrix. We use $\sigma = 1$ on row-normalised features and retain the kernel diagonals. The bandwidth sets which distances contribute most; linear CKA and smaller bandwidths are examined below.

Calibration and interpretation. For K row permutations, we use the $\lceil (1-\alpha)(K+1) \rceil$th order statistic of the observed score together with the K null scores, with $\alpha = 0.05$. The add-one right-tail permutation value is $p = (1 + \#\{s^{(k)} \geq s\})/(K+1)$; full-set analyses use $K = 1{,}000$ permutations, so its minimum is $1/1001$ there, and $1/201$ for the layer and bandwidth controls. These are per-comparison values. The permutation scheme specifies the null: unrestricted shuffles test correspondence across the whole vocabulary, whereas within-category shuffles test residual correspondence conditional on the chosen categories. Validity requires exchangeability under that scheme; neither null isolates learning from shared wording or taxonomy by itself.

For the 7B text representation ($d = 3584$), calibrated linear CKA is 0.281 at $n = 1{,}000$ and 0.280 at $n = 23{,}657$, while the null mean falls from 0.098 to 0.003; calibrated mKNN instead increases with gallery size at fixed $k = 10$, which is why size comparisons hold the gallery and k fixed. CKA therefore stays above the null across the tested sample-size range, including $n < d$; Section 4.4 examines its trend with capability.

Why width and layer search matter. The confounders in Gröger et al. (2026) can arise without learned structure. For fixed n, independent Gaussian representations with row-normalised embeddings have Gram matrices $G_X, G_Y \to I_n$ as both widths increase. Writing $H = I_n - \mathbf{1}\mathbf{1}^\top/n$, their centred matrices both approach H, so biased linear CKA approaches one. In contrast, independent exchangeable neighbour sets satisfy $\mathbb{E}[\mathrm{mKNN}] = k/(n-1)$, irrespective of width. With 128 observations, 200 independent repetitions and widths 16–1,024, mean raw linear CKA rises from 0.113 to 0.890; RBF CKA ($\sigma = 1$) rises from 0.250 to 0.960, while mKNN stays near $10/127$. Mean calibrated scores remain below 0.001 (199 permutations per repetition).

Layer search introduces a separate effect: the maximum cannot decrease when candidate pairs are added. At width 64, searching eight Gaussian representations per space gives 64 pairs; within-space correlation is 0.5 before normalisation, and the two spaces are independent. Calibrating only the winning pair yields positive scores in 73.5%, 64.0% and 88.0% of repetitions for linear CKA, RBF CKA and mKNN. Repeating the search inside each permutation reduces these fractions to 4.5%, 4.5% and 5.5%. These controls explain the calibration procedure; they do not measure how much architecture contributes to the trained-model trend, which the trained width and depth controls address (App. B).

Bandwidth and local distances. A median-bandwidth RBF kernel retains calibrated CKA near 0.28 for static life embeddings versus 7B, but does not isolate very close neighbours. At 0.2 times each space's median pairwise distance, RBF CKA falls from a raw score of 0.521 to a calibrated score of 0.0008 on a 4,000-token subsample of the full set (200 permutations; text–text pairs retain 0.1–0.3 and random tables 0 at the same bandwidth). This effect is very small, although its per-comparison permutation value is $p = 1/201$. On a matched subsample of the contextual subset, static embeddings and contextual vectors retain 0.055 and 0.084, respectively. Fine-scale distance agreement is thus strongly attenuated and depends on coverage and representation. These bandwidths are defined relative to each space's median distance and differ from the fixed bandwidth in the primary RBF comparison. The sensitivity to code coverage prevents a general claim that local-distance agreement vanishes.

Sweeping k from 1 to 500 on matched contextual-subset codes preserves the static embeddings' advantage over contextual vectors, although absolute local alignment and the random baseline change with k.

Random-initialisation floor. Untrained transformers are run over the same lives and pooled by code, so this control includes structure introduced by the input sequences and the averaging procedure. Across the three tested shapes, their mean mKNN with the eight text embedders of App. B is 0.001–0.003, compared with 0.121–0.181 for the corresponding trained models at the final block. Random life models share at most 0.004 mKNN with one another and 0.006 with trained life models. Training therefore adds substantial local alignment beyond this architecture-and-input baseline.

E REGISTRY CO-OCCURRENCE

The PRH proposes that models converge by capturing related co-occurrence statistics. Factorising the registry's positive pointwise mutual information (PPMI) matrix for the 3,000 most frequent codes into 256 dimensions gives a representation that shares neighbours with both modalities (mKNN 0.30 with the reference life model and 0.21–0.23 with text embedders). Whole-history PMI correlates with cosine similarity at 0.28 for the static table, 0.12 for contextual vectors and 0.12–0.21 for text embedders; contextual vectors track one- or two-event windows and next-event transitions more closely than whole histories, increasingly so with model size. Removing the association of ranked pairwise distances with PMI retains 79–96% of life–text agreement, whereas adjusting for the taxonomy retains 31–49% for static embeddings and at most a third in context. Life representations track local co-occurrence, but this measured PMI proxy does not explain most of their shared distance structure with text.

This comparison uses the 3,000 most frequent codes and text embedders. A matched comparison uses the primary protocol: the same 1,024 codes, both LLM cohorts, the best layer pair and 200 permutations repeating the search. The counts cover every life of the training cohort, visited in a fixed random order and cut to the first 4,096 tokens as for the contextual vectors (12.2 billion tokens); representations are exported after 20,000, 100,000 and 1 million lives and after the full cohort. Each of the 1,024 codes is counted against the full vocabulary in three ways, with each token counted once per record. Codes found in fewer than five people of a sample receive a zero vector (198 codes at 20,000 lives, 34 at 100,000, none beyond).

All three baselines score a pair of tokens with the pointwise mutual information (PMI; Church & Hanks, 1990), $\mathrm{PMI}(c,t) = \log[\hat{P}(c,t)/(\hat{P}(c)\hat{P}(t))]$: the logarithm of how much more often the two occur together than if they were independent. Negative values, pairs rarer than chance, are set to zero, giving positive PMI (PPMI; Bullinaria & Levy, 2007), and the context frequencies $\hat{P}(t)$ are smoothed with exponent $\alpha = 0.75$ (Levy et al., 2015). The baselines differ in what counts as occurring together.

First-order Markov chain. The consecutive-record counts form the transition table of a first-order Markov chain over records. With $C[c,t]$ the number of consecutive record pairs (e_i, e_{i+1}) in which $c \in e_i$ and $t \in e_{i+1}$, and $n(c)$ the number of records that contain c,

$$\hat{P}(t \in e_{i+1} \mid c \in e_i) = \frac{C[c,t]}{n(c)}. \tag{3}$$

The representation stores each entry as a lift rather than a probability, $\mathrm{PMI}(c,t) \approx \log[\hat{P}(t \in e_{i+1} \mid c \in e_i)/\hat{P}(t)]$ with $\hat{P}(t)$ the share of records containing t: the logarithm of how much more often t follows c than chance predicts. A code's vector concatenates its outgoing row with its incoming row, the same statistic for the reversed chain.

± 5-record window. With $B[c,t]$ the number of record pairs (e_i, e_j) with $|i-j| \le 5$ (including $j = i$) in which $c \in e_i$ and $t \in e_j$, and W the number of such pairs, $\mathrm{PMI}(c,t) = \log[(B[c,t]/W)/(\hat{P}(c)\hat{P}(t))]$ asks whether t occurs within five records of c, in either direction, more often than chance. This is the record-level counterpart of the symmetric context window of word vectors, whose skip-gram training implicitly factorises a shifted PMI matrix (Levy & Goldberg, 2014). It keeps proximity but not order.

Bag of events. Each life is treated as a document and its codes as words, ignoring time. With $A[c,t]$ the number of lives that contain both c and t, $d(c)$ the number of lives that contain c and M the number of lives, $\mathrm{PMI}(c,t) = \log[(A[c,t]/M)/((d(c)/M)(d(t)/M))]$. This is the bag-of-words co-occurrence of a document collection applied to lives, and it keeps neither order nor proximity.

For each baseline, the 1,024 PPMI rows are embedded exactly through the eigendecomposition of their Gram matrix. A fixed random rotation preserves every cosine but keeps the protocol's 95th-percentile clipping from truncating the leading components, which otherwise costs the window and Markov vectors 0.03–0.04.

Each baseline is a table over pairs of the 23,657 training tokens. The Markov table is ordered, with $23{,}657^2 \approx 5.6 \times 10^8$ entries; the window and bag-of-events tables are symmetric, with approximately 2.8×10^8 distinct entries each. The vectors of the 1,024 codes read approximately 4.8×10^7

Markov entries (their rows and columns) and 2.4×10^7 entries of each symmetric table; on the full cohort 1.7×10^6 (Markov), 1.5×10^6 (window) and 4.9×10^6 (bag of events) of them are positive. For comparison with these counted entries, the 10k, 100k and 1M life models have about 0.4M, 1.6M and 4M trained parameters, including the tied embedding table.

All count representations exceed their permutation thresholds. Table 15 reports their mean calibrated mKNN as the number of training lives increases. The gains of the two record-level baselines shrink with data, while the bag of events is still rising.

Table 15: Mean calibrated local alignment (mKNN) of the count baselines with each LLM cohort. Columns give the number of training lives used to construct the counts; all comparisons use the primary 1,024 codes.

	PRH cohort				HellaSwag cohort
Baseline	20,000	100,000	1 million	Full cohort	Full cohort
Markov	0.138	0.190	0.221	0.233	0.213
Window	0.130	0.178	0.210	0.221	0.203
Bag of events	0.043	0.059	0.085	0.111	0.101

On the full cohort, the Markov vectors exceed the 10k life model (0.209 and 0.196 in the PRH and HellaSwag cohorts, respectively) for all 12 PRH LLMs and 22 of the 25 HellaSwag LLMs. They exceed the 100k model (0.226 and 0.210) for 11 of 12 and 18 of 25 LLMs, respectively. The 1M model (0.261 and 0.240) exceeds the Markov vectors for every LLM; the 100M model reaches 0.282 and 0.258. The window vectors exceed the 10k model for 12 of 12 and 19 of 25 LLMs, and the 100k model for 3 of 12 and none of 25. The bag of events stays below the 10k model for every LLM.

The local alignment of the count vectors also correlates with language performance ($r = 0.60$–0.93; 0.82 and 0.93 for the Markov vectors). So does their global alignment with the HellaSwag cohort: calibrated RBF CKA of 0.29 (Markov), 0.37 (window) and 0.19 (bag of events), against 0.42–0.48 for the life models, correlates with HellaSwag accuracy at $r = 0.88$–0.91.

The margin of the life models over the Markov vectors grows with size: -0.023, -0.007, 0.028, 0.035, 0.042 and 0.050 from 10k to 100M in the PRH cohort, and -0.018, -0.003, 0.027, 0.032, 0.037 and 0.045 in the HellaSwag cohort. Across LLMs, its correlation with language performance turns from negative for the 10k and 100k models ($r = -0.94$ and -0.89 with HellaSwag accuracy), where the counts account for the agreement, to positive from 3M upwards (0.62, 0.67 and 0.74 for 3M, 10M and 100M); in the PRH cohort it is 0.10, 0.23 and 0.73.

Next-token prediction also learns relations within individual records, not only across events separated in time. A model trained on chronologically shuffled histories at one size would test how much of the remaining gap sequence modelling explains.

F LIFE-MODEL SCALING AT FIXED DATA

All life models are trained on the same registry corpus, 12.6 billion tokens per epoch (the 100M and 500M runs count slightly longer epochs; Table 7), so the size series tests larger models on fixed data. The primary 3M run uses 180k tokens per update; the other primary runs use 120k. Figure 10 relates their validation loss to training compute and size; the diminishing gains motivate stopping the series at 500M non-embedding parameters.

Best validation loss falls from 2.591 at 10k to 1.252 at 10M and 1.176 at 100M, but only to 1.156 at 500M: the last step gains 0.020 for 3.3 times the training compute, against 0.076 from 10M to 100M. The 100M model stops improving after four epochs, and its validation loss rises slightly in the fifth (1.1761 to 1.1771) while its training loss keeps falling, so further passes over the same histories no longer help at this size. A saturating fit $L(N) = E + AN^{-\alpha}$ to the best losses of the seven sizes gives $E = 1.14$ and $\alpha = 0.36$; the 500M model lies 0.02 above this asymptote, and the fit predicts 1.149 at 5B parameters. Dropping the 10k point lowers the asymptote to 1.09 and the 5B prediction to 1.12, still within 0.04 of the 500M model. For reference, about 20 training tokens per parameter are compute-optimal for language models (Hoffmann et al., 2022), which would match the 12.6 billion tokens of the training cohort to about 0.6B parameters, and repeated data loses value

after a few epochs (Muennighoff et al., 2023); this text-derived heuristic is not a limit on useful capacity for this task. Local alignment follows validation loss across the size series (Fig. 3) and at a fixed parameter budget (App. B), and from 100M to 500M both the loss gain and the alignment gain shrink (Fig. 3a).

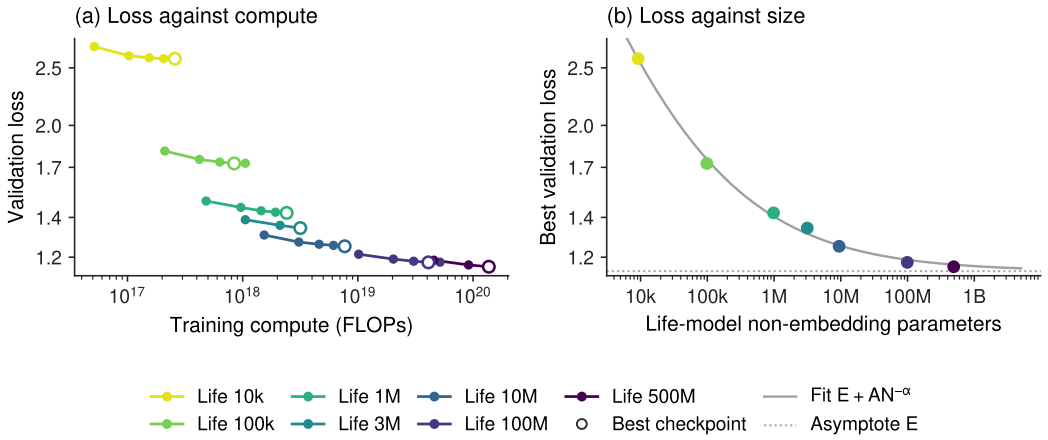


Figure 10: Life-model validation loss at fixed data. (a) Validation loss after every epoch against training compute, estimated as $6ND$ (Kaplan et al., 2020) with N counting non-embedding and output-layer parameters and D the tokens seen (attention over the context is omitted); open circles mark the best checkpoints. (b) Best validation loss against non-embedding parameters with the saturating fit $E + AN^{-\alpha}$ (solid) and its asymptote E (dotted). The 3M curve is the three-epoch seed run with the recipe of the other sizes; the 3M checkpoint of the alignment comparisons has no exported training curve.