TRANSITION AFFINITY: A UNIFIED NOTION OF GEOMETRY FOR LLMS AND TEXT CORPORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer next-token predictions induce an argmax partition of the embedding space, where each cell consists of representations for which a given vocabulary token has maximal logit. We first argue that this representation is a useful way to interpret a transformer. We then show that the geometry of this partition is governed by a phenomenon we call transition affinity: the size of the shared region between two cells is highly correlated to the probability of the two corresponding tokens following one another. This yields a common geometric representation, allowing us to compute correlations between two LLMs, two corpora, or an LLM and a corpus. We conclude by showing how transition affinity can be used to detect that a model has been distilled.

1 INTRODUCTION

Transformers (Vaswani et al., 2017), driven by their success in natural language processing (Wolf et al., 2020), have become the dominant architecture in modern deep learning. They rely on mapping units of text, known as tokens, to a high dimensional space called the *embedding space*. As such, this space is thought to carry meaningful structure about language and about the world it describes, and its characteristics have been an important object of study in the last years. Indeed, the embedding spaces of LLMs have been shown to encode semantic concepts such as space, time, and truth along directions (Gurnee & Tegmark, 2024; Park et al., 2024; Marks & Tegmark, 2024; Park et al., 2025), but also more complex features such as sarcasm, deception or incompetence (Templeton et al., 2026), and grammatical structure shared across languages (Brinkmann et al., 2025). In this work we take a more general approach. Rather than extracting specific meanings from the embedding space, we study how its geometry is shaped by the statistical features of the training data.

Our object of study is what we call the *argmax partition* of the embedding space (Section 2). To sample the next token, a transformer maps a vector of its embedding space to a probability distribution over the vocabulary. We can therefore partition the embedding space according to the most likely next token. This leads to one cell per token, containing the vectors for which that token is the top prediction. Seen through this partition, generation is a walk from cell to cell. The model reads a token, its embedding lands in some cell, it emits the token of that cell, and moves on. Intuitively, a transformer should mostly hop between neighbouring cells, as illustrated in Section 3.

Our analysis relies on the intuition that the probability of a hop should be correlated with what cell is closest. This motivates our definition of the geometry of a transformer. For any two tokens $\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{V}$ with cells $\mathcal{C}_1$ and $\mathcal{C}_2$, we can compute the proportion $\tau_{\mathbf{t}_1 \to \mathbf{t}_2}$ of $\mathcal{C}_1$ that is closest with $\mathcal{C}_2$. Accordingly, we define the geometry of a transformer $\mathcal{T}$ as the matrix $\mathcal{G}_\mathcal{T} = \left(\tau_{\mathbf{t}_1 \to \mathbf{t}_2}\right)_{\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{V}}$. This same geometry can easily be computed for any corpus of texts or document $\mathcal{D}$ as the matrix $\mathcal{G}_\mathcal{D} = \left(\pi_{\mathbf{t}_1 \to \mathbf{t}_2}\right)_{\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{V}}$, where $\pi_{\mathbf{t}_1 \to \mathbf{t}_2}$ denotes the probability that $\mathbf{t}_2$ follows $\mathbf{t}_1$. We call transition affinity the correlation between both geometries and show that it does indeed carry meaning in Section 4.

Finally, we show that transition affinity allows to detect distillation and fine-tuning in Section 5. Notably, as the argmax partition is fixed by the unembedding matrix alone, transition affinity can be computed for any open model, at any scale, without prompts or probes.

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.

1.1 Related Work

Transition affinity studies the bigram statistics of the transformer, that is the frequency and probability of pairs of adjacent items. The intuition that those statistics are largely encoded in the embedding and unembedding matrices stems from Elhage et al. (2021). More generally, studying the embedding space through the lens of these matrices has become standard in interpretability. The logit lens (Nostalgebraist, 2020) and the tuned lens (Belrose et al., 2023) project intermediate representations onto the vocabulary, Geva et al. (2022) and Dar et al. (2023) interpret weight matrices in vocabulary space, and Cancedda (2024) studies the spectrum of the unembedding matrix and its degenerate directions. These work motivate the argmax partition as a visual tool to represent how a transformer navigates its embedding space.

The structure of the embedding space of transformers has been extensively studied both empirically and theoretically. It has been shown to satisfy several characteristics, four of which we mention in the following. This paper adds a new one to the list, namely transition affinity. All four of those existing constraints are compatible with the transition affinity constraint. It is clear for all of them but the Linear Representation Hypothesis, which we discuss further in Appendix B.

The Mixed-State Presentation. The most closely related work to ours is that of Shai et al. (2024), who also study how the structure of the training data influences that of the embedding space at the last layer. More specifically, they show that a transformer trained on data sampled from a Hidden Markov Model exhibits its mixed-state presentation as the embedding space. In contrast, we only consider the successive outputs of the model and their general transition probabilities rather then the underlying hidden states. By considering this simpler setting, we are able to scale our analysis to bigger models. We discuss this relation further in Remark 6.

The Linear Representation Hypothesis. (Mikolov et al., 2013b; Pennington et al., 2014; Nanda et al., 2023; Park et al., 2024; Tigges et al., 2024) show that semantic concepts are represented as vectors in the embedding space of NLP models (*e.g.* "king - man + woman $\simeq$ queen" (Mikolov et al., 2013a)).

The Manifold Hypothesis. The hypothesis that only a lower dimensional submanifold of the embedding space is being used by a transformer has been discussed in the past years (Robinson et al., 2025).

On Anisotropy. Other work has showed that embeddings tend to occupy a narrow cone in high-dimensional space (Mimno & Thompson, 2017; Mu & Viswanath, 2018; Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020; Cai et al., 2021).

2 The Argmax Partition of a Transformer

We start by introducing the definition of a transformer, and the notations we will be using throughout the paper.

Definition 1 (Transformer)**.** Let $\mathcal{V} = \{\mathbf{t}_1 \dots, \mathbf{t}_{|\mathcal{V}|}\}$ be the vocabulary, that is the set of all possible tokens. A transformer of embedding dimension d sends any sequence of arbitrary length n, $\mathbf{t}_1 \dots \mathbf{t}_n \in \mathcal{V}^n$, to a vector $\mathbf{h} \in \mathbb{R}^d$ (the hidden representation of the last token at the last layer). A logit $\ell_i(\mathbf{h}) = \langle \mathbf{e}_i, \mathbf{h} \rangle$ is then associated to each token $\mathbf{t}_i$. This is mapped to a probability distribution over $\mathcal{V}$ using $\mathrm{softmax}$, and the next token is sampled accordingly.

We can now define the argmax partition that is our main object of study in this paper. This partition was introduced by Meyer et al. (2026) who use it to justify the intuition behind their proofs, but without conducting any analysis.

Definition 2 (Argmax partition)**.** We partition the embedding space such that each cell consists of representations for which a given vocabulary token has maximal logit. Formally, the cell associated with token i is

$$\mathcal{C}_i = \left\{ \mathbf{h} \in \mathbb{R}^d : \ell_i(\mathbf{h}) \geq \ell_j(\mathbf{h}) \text{ for all } j \in [|\mathcal{V}|] \right\}.$$

We illustrate the argmax partition in Figure 1. A few important geometric properties of the argmax cells are the following.

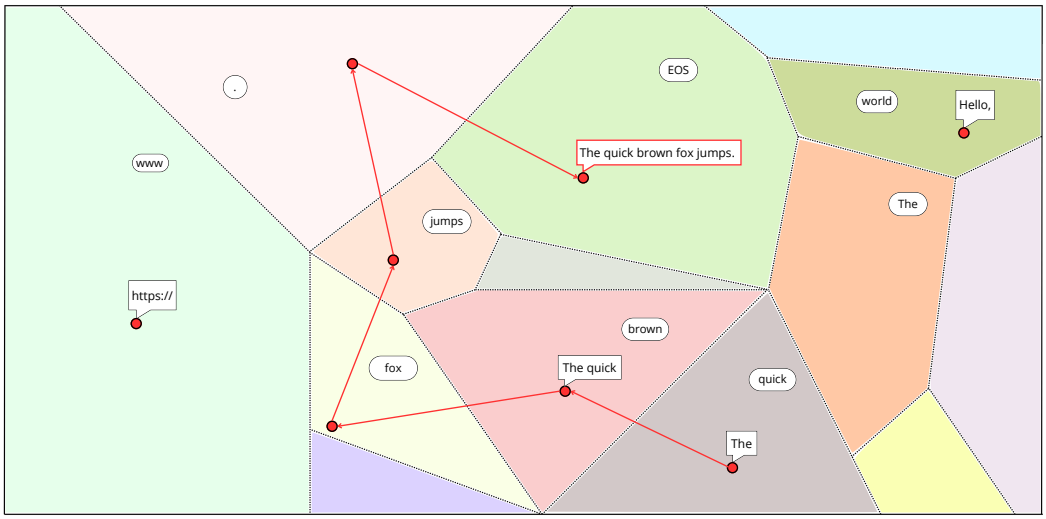

Figure 1: **Plane cut of the argmax partition.** The transformer can be viewed as following trajectories in the argmax partition. Each successive final embedding is indicated by a red dot. The next word is sampled according to the position of that embedding, which leads to a new final embedding, *etc*. Note that this argmax partition is for illustration purposes and doesn't correspond to an actual transformer, as any trajectory of length greater than 2 will almost surely not be along a 2-dimensional plane.

Property 3 (Geometric characteristics of the argmax cells)**.** Clearly, every $\mathcal{C}_i$ corresponds to a convex cone with apex at the origin. This is illustrated in Figure 3. However, since the embedding dimension d is typically greater than 2, we represent the argmax partition by taking a plane cut of the embedding space throughout this paper (as illustrated in Figure 1). The cells then correspond to convex polygons.

We prove this in Appendix A.1. Let $P = \text{conv}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be the convex hull of the token embeddings and a_{ij} denote the area of the border between two cells. An important thing to not is that some tokens $\mathbf{t}$ might correspond to en ampty cell. No matter the hidden state, the argmax never corresponds to $\mathbf{t}$.

Property 4 (On the number of different argmax cells)**.**

(i) If $\mathbf{e}_i$ is an interior point of P, then $\mathcal{C}_i = \{\mathbf{0}\}$ and token i is never predicted by a transformer with greedy decoding (that is, an argmax transformer). If $\mathbf{e}_i$ is a vertex of P, then $\mathcal{C}_i$ is the normal cone of P at $\mathbf{e}_i$ and is d-dimensional. We call such a token *active*.

(ii) If i and j are active, then $a_{ij} > 0$ if and only if the segment $[\mathbf{e}_i, \mathbf{e}_j]$ is an edge of P.

We prove this in Appendix A.2. The first half of Property 4, that a token whose unembedding lies inside the convex hull of the vocabulary can never receive the highest logit, was observed in Neural Network Language Models by Demeter et al. (2020), and Grivas et al. (2022) showed how to certify which tokens of a trained model are *unargmaxable* by solving one linear program per token, finding that such tokens are rare in practice and are mostly low-frequency subwords.

3 The Argmax Partition Preserves the Structure of the Training Data

In this section we give a simple illustration of the usefullness of the argmax partition. We show that when trained on synthetic data sampled following a Markov chain, the argmax partition is exactly the same as the Markov graph.

Definition 5 (Markov chain)**.** A (time-homogeneous) Markov chain on a finite state space $\mathcal{V}$ is specified by a transition probability $\pi : \mathcal{V} \to \Delta(\mathcal{V})$, where $\Delta(\mathcal{V})$ denotes the simplex of distributions on $\mathcal{V}$: from a state $\mathbf{t}$, the next state is sampled from $\pi(\mathbf{t})$, independently of the past. Its *Markov graph* is the directed graph on $\mathcal{V}$ with one node per state and an edge $\mathbf{t} \to \mathbf{t}'$ whenever $\pi(\mathbf{t})$ puts positive mass on $\mathbf{t}'$.

Experimental setting. We take the state space to be the vocabulary $\mathcal{V}$, so that each state is a token. A sequence is generated by drawing an initial state $\mathbf{t}_1$ uniformly at random and then sampling $\mathbf{t}_{n+1} \sim \pi(\mathbf{t}_n)$. We train a transformer with one layer, one attention head, an MLP of width 8, learned positional embeddings, and no normalization layers. The embedding dimension d is between 2 and 7, depending on the experiment. Input and output embeddings are tied. Weights are initialized i.i.d. from $\mathcal{N}(0, 0.2^2)$ and trained by plain SGD (learning rate 0.3, batches of 64 sequences of length 16) on the next-token cross-entropy. We generate synthetic data using Markov chains. That is the next token of a sentence $\mathbf{t}_1 \dots \mathbf{t}_n$ is sampled from a distribution $\pi(\mathbf{t}_n)$ depending only on $\mathbf{t}_n$. For simplicity, we take $\pi(.)$ to be the uniform distribution on a subset of the vocabulary set. A data set is then fully described by a directed graph, with one node for every token, and an arrow for every possible next token[1].

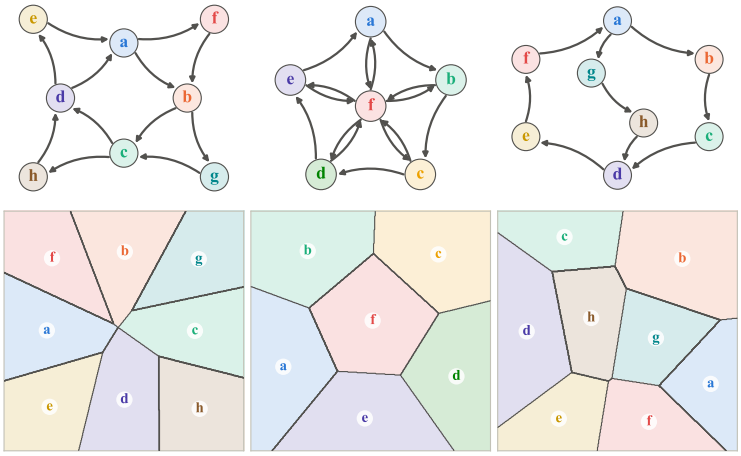

Figure 2: **The argmax partition perfectly matches the Markov graph of the data is was trained on.** We train a transformer on three different Markov chains, represented on the *top row*. We show a plane cut of the argmax partition of the trained transformer on the *bottom row*.

Remark 6 (Related work). This view of the embedding space is closely related to the mixed-state presentation (MSP) of Shai et al. (2024). The difference is that we study Markov chains, where the emitted token is the state itself, instead of hidden Markov models, where a state $\mathbf{s}$ emits a token $\mathbf{t} \sim \nu_{\mathbf{s}}$ drawn from an emission distribution. While the MSP of a Markov chain is exactly its Markov graph, it is a much richer object for a general hidden Markov model. Indeed, some finite models have an infinite MSP. Their framework therefore is a generalization of ours, at the price of an analysis that stays tractable only at small scale. Indeed, Shai et al. (2024) study transformers with four attention heads in total on three-state hidden Markov models. In contrast, our simpler setting lets us measure the transition affinity of state-of-the-art transformers.

3.1 Testing Different Markov Chains

Since the agmax partition is in d dimensions, we need to choose a plane cut to represent. We therefore search for a plane whose cut meets every cell: a candidate plane is scored by the area of the smallest region it shows, and we keep the best of forty thousand random planes, refined by hill-

[1]The only exception is the last of the three chains of Figure 2, where token a is followed by b with probability 0.9 and by g with probability 0.1.

climbing. We show the planes of transformers of dimensions $d = 4, 3, 7$ respectively in Figure 2. Each is trained for 4000 steps, and every run ends within 0.024 nats[2] of its respective Markov chain.

3.2 Showing the Impact of Noise

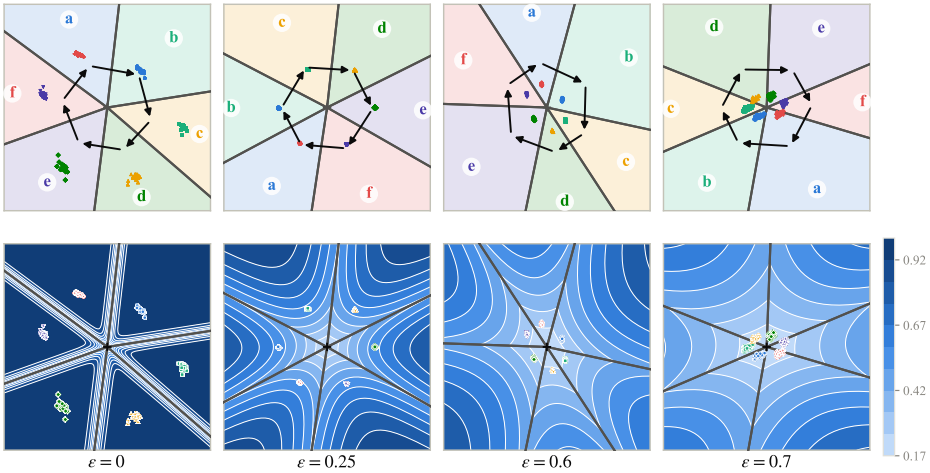

Figure 3: **Argmax partition of a two-dimensional transformer on the 6-state ordered cycle Markov chain with increasing noise.** *Top:* the argmax cells, labeled by token. The dots are the final hidden states of the 16-token sequences, colored by the token just read; the black arrows join the cones of consecutive tokens. *Bottom:* the largest predicted probability $\max_i p_\theta(i \mid h)$. *Main observation:* the cones sit in the order of the cycle. The state after token i lies in the cone of $i + 1$, with it becoming closer to 0 (and hence less confident) as the noise increases.

We now study a two-dimensional transformer, so that the argmax partition can be represented entirely. Note that by definition, every cell of the argmax partition is a cone with apex at the origin, so this is exactly what we observe in 2 dimensions. We train a transformer for 900 steps on the simple 6-cycle $a \to b \to c \to d \to e \to f \to a$, with added noise ε: the next token follows the cycle with probability $1 - \varepsilon$, and is sampled uniformly over the other five tokens with probability ε.

Figure 3 shows the outcome for four different seeds with $\varepsilon \in \{0, 0.25, 0.60, 0.70\}$. The first thing we notice is that no matter the initialisation, the final argmax partition follows the Markov graph. Remarkably, this holds even for $\varepsilon = 0.7$, where the cycle is barely a preference: the successor has probability 0.3, and each other token 0.14. Second, as the noise increases, the representations become closer to the origin. Indeed, if the representation lies at the origin, the next token is sampled totally uniformly. So the closer the representation is to the origin, the less it matters which cell it lies in.

4 Transition Affinity

In this section we show that the embedding space of a transformer is indeed constrained by transition affinity. That is the cells of two tokens likely to follow one another share a large border in the argmax partition.

4.1 Definition and Validation

We first define transition affinity formally, and then evaluate it by comparing three pretrained models with English Wikipedia.

[2]Here, we compute the distance between two distributions p and q over a finite alphabet using KL divergence: $D_{\mathrm{KL}}(p \,\|\, q) = \sum_i p_i \ln \frac{p_i}{q_i}$ nats.

Measurement. We measure geometry by the relative sizes of regions inside each argmax cell. Within $\mathcal{C}_i$, token i has the largest logit; the runner-up token labels a further subdivision of that cell. To measure these regions, fix a box $B = \prod_{k=1}^{d}[L_k, U_k] \subset \mathbb{R}^d$, with $L_k < U_k$, and draw N independent vectors $\mathbf{h}_r \sim \mathrm{Unif}(B)$. Evaluate the logits over the full vocabulary, and let $a(\mathbf{h})$ and $b(\mathbf{h})$ index the largest and second-largest logits, respectively, with a fixed rule for ties. For $i \neq j$, define

$$n_{ij} = \sum_{r=1}^{N} \mathbf{1}\{a(\mathbf{h}_r) = i,\ b(\mathbf{h}_r) = j\}, \qquad n_i = \sum_{j \neq i} n_{ij}, \qquad \widehat{\tau}_{ij}^{B} = \frac{n_{ij}}{n_i} \quad (n_i > 0). \tag{1}$$

Normalizing by n_i compares runner-up identities within the same cell. The ratio estimates $\Pr[b(\mathbf{h}) = j \mid a(\mathbf{h}) = i]$ under uniform sampling in B. We use this cell-volume fraction as the transition affinity $\tau_{\mathbf{t}_i \to \mathbf{t}_j}$.

The corpus supplies a second conditional distribution, now over the token immediately following i. For a corpus $\mathcal{D}$, let $c_{ij}^{\mathcal{D}}$ count adjacent occurrences of $\mathbf{t}_i\mathbf{t}_j$ within an article, and define

$$c_i^{\mathcal{D}} = \sum_{k} c_{ik}^{\mathcal{D}}, \qquad \widehat{\pi}_{ij}^{\mathcal{D}} = \frac{c_{ij}^{\mathcal{D}}}{c_i^{\mathcal{D}}} \quad (c_i^{\mathcal{D}} > 0). \tag{2}$$

Thus $\widehat{\pi}_{ij}^{\mathcal{D}}$ is the empirical probability of j immediately after i. Every adjacent pair is counted, so three consecutive tokens contribute two bigrams. Both scores are directed and conditional on the first token, giving empirical versions of $\mathcal{G}_{\mathcal{T}}$ and $\mathcal{G}_{\mathcal{D}}$. Self-transitions contribute to $c_i^{\mathcal{D}}$, although winner and runner-up are necessarily distinct.

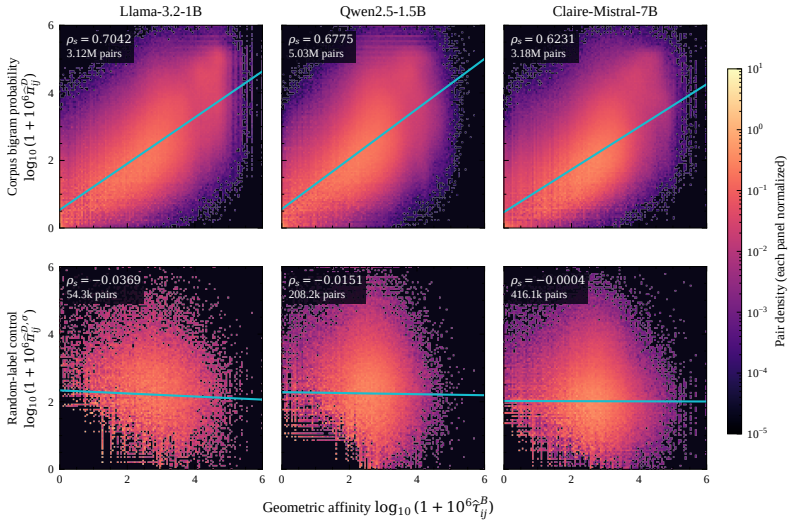

Figure 4: **Corpus followability aligns with transition affinity under the actual token labels.** Top: conditional bigram frequency against transition affinity. Bottom: the same comparison after randomly relabeling corpus tokens. Each panel includes all positive-overlap pairs for its labeling and reports the unbinned Spearman correlation. Axes, bins, and normalized density scale are shared. Cyan lines are affine least-squares guides in the displayed coordinates.

Transition Affinity. The scores describe different events, so we compare their ordering rather than requiring numerical agreement. We use the pairs observed in both geometry and text,

$$\mathcal{P}_{\mathcal{T},\mathcal{D}}^{+} = \{(i,j) : i \neq j,\ n_{ij} > 0,\ c_{ij}^{\mathcal{D}} > 0\}. \tag{3}$$

For a pair set $\mathcal{S}$ on which both scores are defined, transition affinity is their Spearman correlation,

$$\rho_s(\mathcal{S}) = \mathrm{Corr}_{(i,j)\in\mathcal{S}}\big(\mathrm{rank}_{\mathcal{S}}(\widehat{\tau}_{ij}^{B}), \mathrm{rank}_{\mathcal{S}}(\widehat{\pi}_{ij}^{\mathcal{D}})\big), \tag{4}$$

where Corr is Pearson correlation and ranks are computed within $\mathcal{S}$, with average ranks for ties. A positive value means that pairs with higher transition affinity tend to have higher conditional bigram frequency. Each distinct directed pair contributes once, and its corpus frequency enters through $\widehat{\pi}_{ij}^{\mathcal{D}}$.

English Wikipedia experiment. We use a deterministic subset of the 2023-11-01 English Wikipedia snapshot.[3] The measurement articles are identical for Llama-3.2-1B (Grattafiori et al., 2024b), Qwen2.5-1.5B (Yang et al., 2024), and Claire-Mistral-7B-0.1 (Jiang et al., 2023), and are tokenized separately for each model.[4] We count adjacent tokens within articles, without adding special tokens or transitions across article boundaries. For each model, $N = 10^9$ samples estimate $\widehat{\tau}^B$ in its fixed sampling box. Once the box is fixed, geometric scores require only the output embeddings and sampled vectors. Every pair in $\mathcal{P}^+_{T,\mathcal{D}}$ enters both the correlation and the figure; no pair is subsampled.

Random-label control. To test whether the alignment depends on token identities, we retain the geometry and randomly relabel the corpus. Let σ be a uniformly sampled permutation of all vocabulary indices. The same permutation acts on both tokens in every bigram:

$$\widehat{\pi}^{\mathcal{D},\sigma}_{ij} = \frac{c^{\mathcal{D}}_{\sigma(i)\sigma(j)}}{c^{\mathcal{D}}_{\sigma(i)}}, \qquad \mathcal{P}^+_\sigma = \{(i,j) : i \neq j,\ n_{ij} > 0,\ c^{\mathcal{D}}_{\sigma(i)\sigma(j)} > 0\}. \tag{5}$$

The denominator is positive on $\mathcal{P}^+_\sigma$. Relabeling preserves the corpus bigram table but randomizes its correspondence with the geometry. We apply Eq. (4) to $\widehat{\tau}^B$ and $\widehat{\pi}^{\mathcal{D},\sigma}$ on $\mathcal{P}^+_\sigma$. The selection rule is unchanged; the selected pairs can differ because overlap is recomputed after relabeling. We repeat this comparison for 199 random permutations per model. Figure 4 displays the first, specified before examining the results.

Results and analysis. Across 3.12M, 5.03M, and 3.18M directed pairs, the observed correlations are 0.704, 0.677, and 0.623 for Llama, Qwen, and Claire, respectively. In each model, the dense band rises with transition affinity: pairs that receive larger geometric scores also tend to have larger conditional bigram frequencies. Random relabeling gives mean correlations of -0.012, -0.001, and -0.002, respectively. The largest values among the 199 random assignments are 0.532, 0.249, and 0.107, all below the corresponding observed correlations. Thus the actual token labels give stronger alignment than every tested random assignment. The nearly horizontal trends in the displayed controls make the contrast visible.

The output-embedding geometry is correlated with the conditional bigram structure of English Wikipedia: among the observed token pairs, higher corpus followability tends to correspond to higher transition affinity. The relationship observed on synthetic Markov chains also appears in natural text. Section 4.2 next asks whether this correlation distinguishes the corpora on which different models were trained.

4.2 Transition Affinity Increases Through Training

To obtain models whose embedding spaces reflect different training distributions, we pre-train the same architecture independently on datasets from four domains: Chess[5], Chinese (CN)[6], Code[7], and French (FR)[8]. All runs start from the same randomly initialized checkpoint of OLMo-2-0425-1B[9], so that differences between the resulting models can be attributed to the data rather than to the initialization. Each dataset is split 80:20 into training and validation sets, each model is trained for up to 10 epochs on its training split, and we keep the checkpoint with the lowest validation loss. This yields four models, M_{Chess}, M_{CN}, M_{Code}, and M_{FR}, together with the untrained checkpoint M_{Original}, which serves as a reference.

We then compute the correlation between the transition affinity of every model M_i and the bigram statistics of every dataset D_j. The result is the 5×4 matrix of Figure 5: one row per model, including the reference, and one column per dataset. For the four pre-trained models, the diagonal

[3] Wikimedia/Wikipedia, configuration `20231101.en`. The subset contains 1,181,146 distinct articles, including 945,025 measurement articles: 675.1M, 710.8M, and 801.2M tokens under the respective tokenizers.

[4] Model cards: Llama-3.2-1B, Qwen2.5-1.5B, and Claire-Mistral-7B-0.1.

[5] https://huggingface.co/datasets/bingbangboom/chess-puzzles-images-large

[6] https://huggingface.co/datasets/0xDing/wikipedia-cn-20230720-filtered

[7] We use the R-language subset of https://huggingface.co/datasets/bigcode/the-stack

[8] We use `gallica_presse_1.parquet` from https://huggingface.co/datasets/PleIAs/French-PD-Newspapers

[9] https://huggingface.co/allenai/OLMo-2-0425-1B/tree/stage1-step0-tokens0B

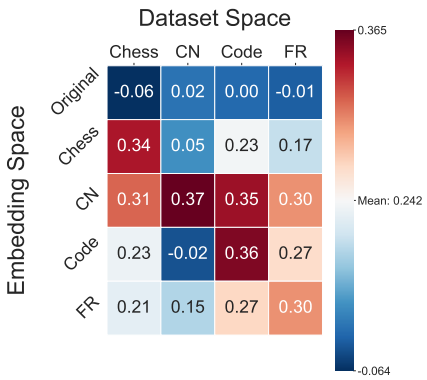

Figure 5: **Each model's argmax partition is most correlated with the dataset it was trained on.** Correlation between the transition affinity of each model (rows) and the bigram statistics of each dataset (columns). For all four pre-trained models the diagonal entry is the largest of its row, rising from at most 0.02 at initialization (Original row) to 0.30–0.37 after pre-training, while the off-diagonal entries stay closer to the values of the untrained model.

entries compare a model with its own pre-training data and the off-diagonal entries compare it with the other three datasets. If transition affinity records what the model learned during pre-training, the diagonal entry of each row should rise well above its value in the Original row, and should exceed the off-diagonal entries of the same row.

The off-diagonal entries also rise above the Original row, though by less than the diagonal. This is expected: the four corpora are tokenized with the same vocabulary and share a large part of their bigram structure, such as the transitions involving punctuation, digits, whitespace, and subword continuations, and the datasets themselves are correlated with each other. A model that learns the transitions of its own corpus therefore also learns the transitions common to all four, which raises its correlation with every dataset. The diagonal measures the shared part plus the corpus-specific part, which is why it is the largest entry of each row. We confirm this phenomenon in the next Section, where we show that removing the most frequent tokens allows to capture more relevant information with transition affinity.

5 Using transition affinity to detect Distillation

Section 4 shows that transition affinity captures information about a model's training distribution. We now ask whether it also identifies relationships between models: whether models connected through distillation or subsequent adaptation correlate more strongly than models from other families.

Experimental setting. We use Llama-3.1-8B (Grattafiori et al., 2024a) as the reference model and compare it with the eight models in Table 1. Llama-3.2-3B[10] is obtained from Llama-3.1-8B by pruning and knowledge distillation, using logits from Llama-3.1-8B and Llama-3.1-70B as token-level targets. Swallow (Fujii et al., 2024; Okazaki et al., 2024), Tulu 3 (Lambert et al., 2024), and Hermes 3 (Teknium et al., 2024) are Llama-3.1-based variants obtained through continued pre-training or post-training. Phi-4-mini (Microsoft et al., 2025), OLMo-2-0425-1B (OLMo et al., 2024), stable-code-3B (Pinnaparaju et al.) , and Hunyuan-7B-Instruct [11] are unrelated models used for cross-family comparison.

Measurement. As in Section 4, we compare the transition-affinity matrices with Spearman's rank correlation over directed token pairs observed in both models. We report two settings: all eligible tokens, and excluding the 20 most frequent tokens (whitespace, commas, periods, newlines, and similar). Both settings keep the positive-overlap restriction; "all tokens" means no additional frequency filtering.

[10] https://huggingface.co/meta-llama/Llama-3.2-3B
[11] https://huggingface.co/tencent/Hunyuan-7B-Instruct

Table 1: **Llama-3.1-8B shares a greater transition affinity with models that are either distilled or fine-tuned from it.** Transition affinity is computed over corresponding token pairs observed in both models, without frequency filtering and after excluding the 20 most frequent tokens. A dash indicates a result not reported.

Model	All tokens	**Excluding top-20 tokens**
Distilled or fine-tuned from Llama-3.1-8B		
Llama-3.2-3B	0.84	**0.75**
Swallow	0.83	**0.78**
Tulu 3	0.80	**0.75**
Hermes 3	0.85	**0.81**
Not distilled or fine-tuned from Llama-3.1-8B		
Phi-4-mini	0.17	**0.12**
OLMo-2-0425-1B	0.78	**0.59**
stable-code-3B	0.37	**0.28**
Hunyuan-7B-Instruct	0.74	**0.62**

Results. With all tokens, the four Llama-related models have the highest correlations with the reference (0.84 for Llama-3.2-3B, 0.83 for Swallow, 0.80 for Tulu 3, 0.85 for Hermes 3), but cross-family models also score high, notably OLMo-2-0425-1B (0.78) and Hunyuan-7B-Instruct (0.74). Excluding the 20 most frequent tokens, Llama-3.2-3B, Swallow, and Tulu 3 retain 0.75, 0.78, and 0.75, while OLMo-2-0425-1B, stable-code-3B, and Hunyuan-7B-Instruct drop to 0.59, 0.28, and 0.62. Among the six models with results in both settings, the gap between the lowest Llama-related score and the highest cross-family score grows from 0.02 to 0.13. Filtering thus lowers absolute correlations but separates the two groups clearly.

This is expected: a very frequent token receives high affinity with almost every neighbor in every model, so it contributes structure shared across families rather than lineage-specific structure, consistent with the common punctuation and whitespace transitions in Section 4.2. Removing these tokens reduces this shared bias. Transition affinity therefore identifies models related by distillation and subsequent adaptation.

6 Conclusion

We introduced the argmax partition, the division of the embedding space into one convex cone per token of the unembedding matrix, and showed that it carries meaningful information about the structure of the training data. We then introduced transition affinity as a tool to extract bigram statistics from it. This tool identifies the training corpus of a model when the candidate corpora are varied enough, and more importantly it detects distillation and fine-tuning between models.

Limitations. The signal is noisy, because border sizes are shaped by constraints other than bigram statistics. In the embedding space, token frequency and confusability indirectly change the size of the borders. Indeed, frequent tokens have large cells that touch many others, and confusable tokens must sit close together, which enlarges their common border whatever the corpus. Moreover, transition affinity does not distinguish directions: the border between a and b is one object, while the bigram probabilities of ab and ba differ. Finally, transition affinity only considers bigram statistics. A more refined tool could take more characteristics of the partition into account, such as its cell sizes, its distances to the origin, or the paths through several cells that a generation follows.

Future Work. Despite the noise described in the previous paragraph and only leveraging the unembedding matrix, a tool as simple as transition affinity detects relevant correlations, such as information about the training corpus and the lineage of a model. We find this promising, and we believe the tool is worth investigating further, both by correcting for the biases above and by extending it beyond bigram statistics. More broadly, we have suggested the argmax partition as a relevant object for studying transformers.

AI USE STATEMENT

In this work, we used generative AI tools for checking typos, writing proofs in latex, and helping with writing code. The proof of Section B.3 was written with the help of AI tools. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, summarize or analyse existing literature, brainstorming, sourcing/searching for information, identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.312. URL https://aclanthology.org/2025.naacl-long.312/.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xYGNO86OWDH.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.263. URL https://aclanthology.org/2024.acl-long.263/.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893/.

David Demeter, Gregory Kimmel, and Doug Downey. Stolen probability: A structural weakness of neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2191–2197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.198. URL https://aclanthology.org/2020.acl-main.198/.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, pp. (to appear), University of Pennsylvania, USA, October 2024.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkEYojRqtm.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, An-

nie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024a. URL https://arxiv.org/abs/2407.21783.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024b. URL https://arxiv.org/abs/2407.21783.

Andreas Grivas, Nikolay Bogoychev, and Adam Lopez. Low-rank softmax can have unargmaxable classes in theory but rarely in practice. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6738–6758, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.465. URL https://aclanthology.org/2022.acl-long.465/.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10 (3), 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. 2024.

John M Lee. Smooth manifolds. In *Introduction to smooth manifolds*, pp. 1–31. Springer, 2013.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119–9130, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.733. URL https://aclanthology.org/2020.emnlp-main.733/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Maxime Meyer, Mario Michelessa, Caroline Chaux, and Vincent Y. F. Tan. How many different outputs can a transformer generate? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=x4xOtzoYa6.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013a. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013b. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2873–2878, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1308. URL https://aclanthology.org/D17-1308/.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

Nostalgebraist. Interpreting gpt: the logit lens., 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, pp. (to appear), University of Pennsylvania, USA, October 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2024. URL https://arxiv.org/abs/2501.00656.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Nikhil Pinnaparaju, Reshinth Adithyan, Duy Phung, Jonathan Tow, James Baicoianu, and Nathan Cooper. Stable code 3b. URL https://huggingface.co/stabilityai/stable-code-3b.

Michael Robinson, Sourya Dey, and Tony Chiang. Token embeddings violate the manifold hypothesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=fiXCcxGysZ.

Adam Shai, Paul M. Riechers, Lucas Teixeira, Alexander Gietelink Oldenziel, and Sarah Marzen. Transformers represent belief state geometry in their residual stream. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YIB7REL8UC.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. Hermes 3 technical report, 2024. URL https://arxiv.org/abs/2408.11857.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2026. URL https://arxiv.org/abs/2605.29358.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 58–87, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.5. URL https://aclanthology.org/2024.blackboxnlp-1.5/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report, 2024. *URL https://arxiv.org/abs/2407.10671*, 7:8, 2024.

Günter M. Ziegler. *Lectures on polytopes*. Springer-Verlag, New York, 1995. URL http://www.worldcat.org/search?qt=worldcat_org_all&q=9780387943657.

A PROOFS OF SECTION 2

A.1 PROOF OF PROPERTY 3

We prove the following.

Property 7 (Geometric characteristics of the argmax cells)**.** Every cell $\mathcal{C}_i$ of the argmax partition is a convex polyhedral cone with apex at the origin. Consequently, the intersection of $\mathcal{C}_i$ with any affine plane is a convex polygon.

Proof. **The cells are cones.** Since the logits are linear, for any $\mathbf{h} \in \mathbb{R}^d$,

$$\ell_i(\mathbf{h}) \;\geq\; \ell_j(\mathbf{h}) \iff \langle \mathbf{e}_i - \mathbf{e}_j,\, \mathbf{h} \rangle \;\geq\; 0,$$

so that

$$\mathcal{C}_i \;=\; \bigcap_{j \neq i} H_{ij}, \qquad H_{ij} := \big\{ \mathbf{h} \in \mathbb{R}^d :\; \langle \mathbf{e}_i - \mathbf{e}_j,\, \mathbf{h} \rangle \geq 0 \big\}.$$

Each H_{ij} is a closed half-space whose bounding hyperplane passes through the origin. In particular H_{ij} is convex, contains $\mathbf{0}$, and is stable under multiplication by any $\lambda \geq 0$, since $\langle \mathbf{e}_i - \mathbf{e}_j, \lambda \mathbf{h} \rangle = \lambda \langle \mathbf{e}_i - \mathbf{e}_j, \mathbf{h} \rangle$. These three properties are preserved by intersection, so $\mathcal{C}_i$ is a convex cone with apex at the origin; it is polyhedral because it is cut out by the finitely many $(|\mathcal{V}| - 1)$ linear inequalities above.

Plane cuts are polygons. Let $A = \mathbf{h}_0 + \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$ be an affine plane and parametrise it by $\mathbf{s} \in \mathbb{R}^2 \mapsto \mathbf{h}_0 + U\mathbf{s}$, where $U = [\, \mathbf{u}_1 \; \mathbf{u}_2 \,]$. Reading the inequalities above in this parametrisation,

$$\mathbf{h}_0 + U\mathbf{s} \in \mathcal{C}_i \iff \big\langle U^\top(\mathbf{e}_i - \mathbf{e}_j),\, \mathbf{s} \big\rangle \;\geq\; \big\langle \mathbf{e}_j - \mathbf{e}_i,\, \mathbf{h}_0 \big\rangle \qquad \text{for all } j \neq i.$$

Each of these is a closed half-plane of $\mathbb{R}^2$ (or, in the degenerate case $U^\top(\mathbf{e}_i - \mathbf{e}_j) = \mathbf{0}$, the empty set or all of $\mathbb{R}^2$). Hence $\mathcal{C}_i \cap A$ is a finite intersection of half-planes, i.e. a convex polygon. □

A.2 PROOF OF PROPERTY 4

We prove the following

Lemma 8 (Cells and the convex hull)**.**

(i) If $\mathbf{e}_i$ is an interior point of P, then $\mathcal{C}_i = \{\mathbf{0}\}$ and token i is never predicted. If $\mathbf{e}_i$ is a vertex of P, then $\mathcal{C}_i$ is the normal cone of P at $\mathbf{e}_i$ and is d-dimensional. We call such a token active.

(ii) If i and j are active, then $a_{ij} > 0$ if and only if the segment $[\mathbf{e}_i, \mathbf{e}_j]$ is an edge of P.

Proof. The polytope P is full-dimensional, because its vertices span $\mathbb{R}^d$ and have mean $\mathbf{0}$. For $\mathbf{h} \in \mathbb{R}^d$, the set $F(\mathbf{h})$ of points of P at which $\mathbf{p} \mapsto \langle \mathbf{p}, \mathbf{h} \rangle$ attains its maximum is called a face of P. The maximum equals $\max_j \ell_j(\mathbf{h})$, because a linear function on a convex hull is maximized at one of the generating points. Hence $\mathbf{h} \in \mathcal{C}_i$ if and only if $\mathbf{e}_i \in F(\mathbf{h})$.

If $\mathbf{e}_i$ is in the interior of P, a nonzero linear function cannot attain its maximum over P there, so $\mathcal{C}_i = \{\mathbf{0}\}$. If $\mathbf{e}_i$ is a vertex, $\mathcal{C}_i = \{\mathbf{h} : \mathbf{e}_i \in F(\mathbf{h})\}$ is by definition the normal cone of P at $\mathbf{e}_i$, and for a full-dimensional polytope the normal cone of a face of dimension s has dimension $d - s$ (Ziegler, 1995, Section 7.1), here d. This proves (i).

For (ii), $\mathbf{h}$ lies in $\mathcal{C}_i \cap \mathcal{C}_j$ exactly when $F(\mathbf{h})$ contains both vertices, so $\mathcal{C}_i \cap \mathcal{C}_j$ is the normal cone of the smallest face F_{ij} containing both. Its dimension $d - \dim F_{ij}$ equals $d - 1$ exactly when F_{ij} is an edge, which is then $[\mathbf{e}_i, \mathbf{e}_j]$. A $(d-1)$-dimensional cone with apex at the origin meets the open ball in a set of positive $(d-1)$-dimensional volume, so $a_{ij} > 0$ in this case. Otherwise $\dim(\mathcal{F}_{ij}) \leq d-2$ and $a_{ij} = 0$ by definition. □

Remark 9. In practice, each $\mathbf{e}_i$ is clearly either a vertex of P or an interior point of P. Indeed, the randomness of initialization and training implies that an embedding lying exactly on a proper face of P never occurs in practice.

B Compatibility of Geometric Constraints

We study in this section the compatibility between transition affinity and the linear representation hypothesis. In particular, we provide a lower bound on the embedding dimension, that depends on the number of constraints, under which they cannot be satisfied. Section B.1 fixes the setting. Section B.2 counts the coordinates left free by r linear relations. Section B.3 shows that each of k independent border-area constraints removes one more, so that k constraints and r relations force the embedding dimension to be at least a number $d_{\min}(n, r, k)$ given in closed form. Section B.4 shows that some combinations of constraints are not encodable, independently of the embedding dimension.

B.1 Setting

Notation. Cells $\mathcal{C}_i$, logits ℓ_i, the embedding dimension d, and the convex hull P of the embeddings are as in Section 2. We write $n = |\mathcal{V}|$ for the vocabulary size and $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_n] \in \mathbb{R}^{d\times n}$ for the matrix of output embeddings. Since the cells are cones, they extend to infinity, and we measure borders inside the ball $\mathbb{B}_R^d = \{\mathbf{h} \in \mathbb{R}^d : \|\mathbf{h}\|_2 \le R\}$, whose radius R only sets the unit of area. For $i \neq j$, the border between the cells of i and j is the part of $\mathcal{C}_i \cap \mathcal{C}_j$ inside the open ball,

$$\mathcal{F}_{ij} = \mathcal{C}_i \cap \mathcal{C}_j \cap \operatorname{int}(\mathbb{B}_R^d),$$

and its area is

$$a_{ij} = \begin{cases} \mathcal{H}^{d-1}(\mathcal{F}_{ij}), & \dim(\mathcal{F}_{ij}) = d-1, \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathcal{H}^s$ denotes s-dimensional volume (Hausdorff measure). An intersection of dimension below $d-1$ is not a border and gets area zero.

Invariances and normalization. Moving all embeddings by the same vector, or scaling them all by the same positive factor, leaves the partition unchanged, since every pairwise logit comparison is preserved. The partition is thus invariant under $\mathbf{e}_i \mapsto c\,\mathbf{e}_i + \mathbf{b}$ with $c > 0$. Rotating or reflecting all embeddings, $\mathbf{e}_i \mapsto \mathbf{Q}\mathbf{e}_i$ with $\mathbf{Q} \in \Omega(d)$, where $\Omega(d) = \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_d\}$ is the orthogonal group, rotates or reflects the partition without changing any distance or area. Distances between embeddings are therefore meaningful only once the scale is fixed. We impose

$$\mathbf{E}\mathbf{1} = \mathbf{0}, \qquad \|\mathbf{E}\|_F = 1, \tag{6}$$

and regard configurations that differ by an element of $\Omega(d)$ as the same geometry. We also assume throughout that $\operatorname{rank}\mathbf{E} = d$, since the vocabulary is far larger than the embedding dimension and the output embeddings of a trained model span the whole embedding space.

Linear relations. Exact linear relational hypotheses have the form $\mathbf{E}\mathbf{A} = \mathbf{0}$ with $\mathbf{A} \in \mathbb{R}^{n\times r'}$, one relation per column, such as the analogy $\mathbf{e}_{\text{man}} - \mathbf{e}_{\text{woman}} = \mathbf{e}_{\text{king}} - \mathbf{e}_{\text{queen}}$ or the additive composition $\mathbf{e}_{\text{king}} = \mathbf{e}_{\text{man}} + \mathbf{e}_{\text{royal}}$. They are understood for the normalized embeddings of (6). The centering $\mathbf{E}\mathbf{1} = \mathbf{0}$ is already imposed by that normalization and is not listed among the relations, so $\mathbf{1} \notin \operatorname{col}(\mathbf{A})$. By Lemma 10 below, we may replace $\mathbf{A}$ by $\mathbf{\Pi}\mathbf{A}$ and drop dependent columns without changing the relations or their number $r = \operatorname{rank}\mathbf{A}$. We therefore assume

$$\mathbf{E}\mathbf{A} = \mathbf{0}, \qquad \mathbf{1}^\top\mathbf{A} = \mathbf{0}, \qquad \mathbf{A} \in \mathbb{R}^{n\times r} \text{ of rank } r. \tag{7}$$

Requiring two tokens to be close, $\|\mathbf{e}_i - \mathbf{e}_j\|_2 < \delta$, is a strict inequality. It restricts the feasible region without lowering its dimension, so such constraints are allowed in Theorem 12 and cost nothing in the count.

Lemma 10 (Relations can be taken with zero coefficient sum)**.** *Let $\mathbf{\Pi} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ be the orthogonal projection onto $\mathbf{1}^\perp$. If $\mathbf{E}\mathbf{1} = \mathbf{0}$, then $\mathbf{E}\mathbf{A} = \mathbf{E}\mathbf{\Pi}\mathbf{A}$ for every $\mathbf{A}$. Hence $\mathbf{E}\mathbf{A} = \mathbf{0}$ if and only if $\mathbf{E}\mathbf{\Pi}\mathbf{A} = \mathbf{0}$, and the columns of $\mathbf{\Pi}\mathbf{A}$ sum to zero. If moreover $\mathbf{1} \notin \operatorname{col}(\mathbf{A})$, then* $\operatorname{rank}(\mathbf{\Pi}\mathbf{A}) = \operatorname{rank}\mathbf{A}$.

Proof. $\mathbf{E}\mathbf{\Pi} = \mathbf{E} - \frac{1}{n}(\mathbf{E}\mathbf{1})\mathbf{1}^\top = \mathbf{E}$, and $\mathbf{1}^\top\mathbf{\Pi} = \mathbf{1}^\top - \frac{1}{n}(\mathbf{1}^\top\mathbf{1})\mathbf{1}^\top = \mathbf{0}^\top$. For the rank, the kernel of $\mathbf{\Pi}$ is $\operatorname{span}\{\mathbf{1}\}$, so $\mathbf{\Pi}\mathbf{A}\mathbf{c} = \mathbf{0}$ forces $\mathbf{A}\mathbf{c} \in \operatorname{span}\{\mathbf{1}\} \cap \operatorname{col}(\mathbf{A}) = \{\mathbf{0}\}$. Thus $\mathbf{\Pi}\mathbf{A}$ and $\mathbf{A}$ have the same kernel, hence the same rank. □

B.2 Coordinates Left Free by the Linear Representation Hypothesis

The embeddings have dn coordinates. Centering removes d of them, the unit norm one more, the rotations and reflections $\binom{d}{2}$, and each of the r relations d more. What remains is the number of free parameters of a geometry satisfying the relations. The relations and the centering leave only $q = n - 1 - r$ directions in which the rows of $\mathbf{E}$ can lie, so a configuration of rank d exists only if $d \leq q$.

Proposition 11 (Dimension of the geometries satisfying the relations)**.** *Let $q = n - 1 - r$, and let $d \leq q$; for $d > q$ no configuration of rank d satisfies (7) and (6). The configurations $\mathbf{E} \in \mathbb{R}^{d\times n}$ of rank d satisfying (7) and (6), taken modulo $\Omega(d)$, form a smooth manifold $\mathcal{M}_{A,d}$ of dimension*

$$D_{n,r}(d) = dq - \binom{d}{2} - 1 = dn - d - dr - \binom{d}{2} - 1. \tag{8}$$

Proof. Let $\mathcal{K} = \operatorname{span}\{\mathbf{1}, \operatorname{col}(\mathbf{A})\}^{\perp} \subseteq \mathbb{R}^n$. Since $\mathbf{1} \perp \operatorname{col}(\mathbf{A})$ and $\operatorname{rank}\mathbf{A} = r$, the span has dimension $r+1$ and $\dim\mathcal{K} = n-1-r = q$. The constraints $\mathbf{E1} = \mathbf{0}$ and $\mathbf{EA} = \mathbf{0}$ say exactly that every row of $\mathbf{E}$ lies in $\mathcal{K}$. Let $\mathbf{B} \in \mathbb{R}^{n\times q}$ have orthonormal columns spanning $\mathcal{K}$, so $\mathbf{B}^\top\mathbf{B} = \mathbf{I}_q$. The map $\mathbf{X} \mapsto \mathbf{X}\mathbf{B}^\top$ is a linear bijection from $\mathbb{R}^{d\times q}$ onto the matrices whose rows lie in $\mathcal{K}$, with inverse $\mathbf{E} \mapsto \mathbf{E}\mathbf{B}$. It preserves rank, because $\mathbf{B}$ has full column rank, and it preserves the Frobenius norm, because $\|\mathbf{X}\mathbf{B}^\top\|_F^2 = \operatorname{tr}(\mathbf{X}\mathbf{B}^\top\mathbf{B}\mathbf{X}^\top) = \|\mathbf{X}\|_F^2$. In particular $\operatorname{rank}\mathbf{E} \leq q$, which proves the first claim, and the configurations in question correspond to

$$S = \{\mathbf{X} \in \mathbb{R}^{d\times q} : \operatorname{rank}\mathbf{X} = d,\ \|\mathbf{X}\|_F = 1\}.$$

The unit sphere of $\mathbb{R}^{d\times q} \cong \mathbb{R}^{dq}$ is a smooth manifold of dimension $dq - 1$, and $\operatorname{rank}\mathbf{X} = d$ is the open condition $\det(\mathbf{X}\mathbf{X}^\top) \neq 0$. Hence S is an open subset of the sphere, nonempty because $d \leq q$, and a smooth manifold of dimension $dq - 1$.

The group $\Omega(d)$ is compact, and it is a Lie group of dimension $\binom{d}{2}$, because it is the preimage of $\mathbf{I}_d$ under the map $\mathbf{Q} \mapsto \mathbf{Q}^\top\mathbf{Q}$ from $\mathbb{R}^{d\times d}$ to the $\binom{d+1}{2}$-dimensional space of symmetric matrices, and $\mathbf{I}_d$ is a regular value of this map (Lee, 2013, Chapter 7). It acts smoothly on S by $\mathbf{X} \mapsto \mathbf{Q}\mathbf{X}$, which preserves rank and Frobenius norm, and $\mathbf{Q}\mathbf{X}$ and $\mathbf{X}$ describe the same geometry because $\mathbf{Q}\mathbf{X}\mathbf{B}^\top = \mathbf{Q}\mathbf{E}$. The action is free, because if $\mathbf{Q}\mathbf{X} = \mathbf{X}$, then $\mathbf{Q}$ fixes every column of $\mathbf{X}$, and these span $\mathbb{R}^d$, so $\mathbf{Q} = \mathbf{I}_d$. A free smooth action of a compact Lie group is proper, so by the quotient manifold theorem (Lee, 2013, Theorem 21.10) the quotient $\mathcal{M}_{A,d} = S/\Omega(d)$ is a smooth manifold of dimension $\dim S - \dim\Omega(d) = dq - 1 - \binom{d}{2}$. □

B.3 The Minimal Embedding Dimension

Let $\mathbf{\Phi} = (\phi_1, \ldots, \phi_k)$ be scalar functions of the geometry, such as border areas, their ratios, or distances between embeddings. Such functions are unchanged by $\Omega(d)$, so they are functions on $\mathcal{M}_{A,d}$; we write $[\mathbf{E}]$ for the point of $\mathcal{M}_{A,d}$ represented by $\mathbf{E}$. We call a configuration $\mathbf{E}_0$ regular for $\mathbf{\Phi}$ if $\mathbf{\Phi}$ is well defined and continuously differentiable on a neighborhood of $[\mathbf{E}_0]$ in $\mathcal{M}_{A,d}$. For border areas this holds whenever small perturbations of $\mathbf{E}$ do not change which sets of vertices form faces of P, because each border is then cut out, inside a hyperplane through the origin, by inequalities whose normals depend smoothly on $\mathbf{E}$. This is the case when every $\mathbf{e}_i$ is a vertex and every facet of P is a simplex. A collection of constraints $\mathbf{\Phi} = \boldsymbol{\phi}^\star$, together with relations and strict inequalities, is compatible if some configuration satisfies all of them.

Fixing the values of k such functions removes k free parameters, provided the constraints are independent, in the sense that none of them is, to first order, a consequence of the others. This is the rank condition (9) below, which says that the k gradients are linearly independent.

Theorem 12 (Dimension of the compatible geometries)**.** *Let $\mathbf{E}_0$ satisfy (7), (6), and finitely many strict inequalities between continuous functions of the geometry, and let it be regular for $\mathbf{\Phi}$ with*

$$\operatorname{rank} D\mathbf{\Phi}([\mathbf{E}_0]) = k. \tag{9}$$

Then the configurations satisfying the same relations and inequalities and $\mathbf{\Phi} = \mathbf{\Phi}([\mathbf{E}_0])$ form, near $[\mathbf{E}_0]$, a submanifold of $\mathcal{M}_{A,d}$ of dimension $D_{n,r}(d) - k$, and every target vector sufficiently close to $\mathbf{\Phi}([\mathbf{E}_0])$ is realized by a nearby configuration. In particular,

$$k \leq D_{n,r}(d) = d(n - 1 - r) - \binom{d}{2} - 1. \tag{10}$$

Proof. By Proposition 11, $\mathcal{M}_{A,d}$ is a manifold of dimension $D_{n,r}(d)$ near $[\mathbf{E}_0]$, and (9) says that $\mathbf{\Phi}$ is a submersion at $[\mathbf{E}_0]$, hence on a neighborhood of it, because maximal rank is an open condition. The level set of a submersion through the point is a submanifold of codimension k, and a submersion maps a neighborhood of the point onto a neighborhood of its image (Lee, 2013, Chapters 4 and 5). The strict inequalities involve continuous functions, so they still hold on a neighborhood of $[\mathbf{E}_0]$, and we shrink the neighborhood accordingly. Finally, (10) holds because the rank of a differential is at most the dimension of its domain. □

The theorem says that, near a compatible configuration, each constraint can be tuned independently of the others and removes one free parameter. Since $D_{n,r}(d)$ grows with d for $d \leq q$, the bound (10) is a lower bound on the embedding dimension.

Corollary 13 (Minimal embedding dimension)**.** *If r relations and k constraints satisfying (9) hold together in dimension d, then, with $q = n - 1 - r$,*

$$d \geq d_{\min}(n, r, k) = \left\lceil \frac{2q + 1 - \sqrt{(2q+1)^2 - 8(k+1)}}{2} \right\rceil. \tag{11}$$

If $k > \frac{q(q+1)}{2} - 1$, no dimension suffices.

Proof. By Proposition 11 and Theorem 12, $d \leq q$ and $k \leq dq - \binom{d}{2} - 1$. The second inequality reads $d^2 - (2q+1)d + 2(k+1) \leq 0$, which holds exactly when d lies between the roots $\frac{2q+1 \mp \sqrt{(2q+1)^2 - 8(k+1)}}{2}$ of the quadratic. The smaller root is at most q exactly when $(2q+1)^2 - 8(k+1) \geq 1$, that is, when $k \leq \frac{q(q+1)}{2} - 1$, and the larger root always exceeds q. Hence, for $k \leq \frac{q(q+1)}{2} - 1$, the admissible dimensions are the integers from $d_{\min}$ to q, and for larger k there are none. □

B.4 Incompatibility Example

We provide an example where one linear constraint is incompatible with a transition affinity constraint, independent of the dimension,

Proposition 14 (Borders forbidden by an exact analogy)**.** *Let four distinct active tokens satisfy $\mathbf{e}_a - \mathbf{e}_b = \mathbf{e}_c - \mathbf{e}_d$. Then*

$$\mathcal{C}_a \cap \mathcal{C}_d = \mathcal{C}_b \cap \mathcal{C}_c,$$

and this set has dimension at most $d - 2$. In particular $a_{ad} = a_{bc} = 0$, so the exact relation $\mathbf{e}_{\text{man}} - \mathbf{e}_{\text{woman}} = \mathbf{e}_{\text{king}} - \mathbf{e}_{\text{queen}}$ precludes borders between man *and* queen *and between* woman *and* king*, and these cells meet only where all four meet.*

Proof. Let $\mathbf{h} \in \mathcal{C}_a \cap \mathcal{C}_d$ and $M = \max_j \ell_j(\mathbf{h}) = \ell_a(\mathbf{h}) = \ell_d(\mathbf{h})$. The relation gives $\ell_b(\mathbf{h}) + \ell_c(\mathbf{h}) = \ell_a(\mathbf{h}) + \ell_d(\mathbf{h}) = 2M$, and neither term exceeds M, so both equal M and $\mathbf{h} \in \mathcal{C}_b \cap \mathcal{C}_c$. The reverse inclusion is symmetric. The common set lies in $\{\mathbf{h} : \ell_a(\mathbf{h}) = \ell_b(\mathbf{h}) = \ell_d(\mathbf{h})\}$, which has codimension two because three distinct vertices of P are not collinear. □

Thus no embedding dimension reconciles this analogy with a transition affinity constraint asking for a positive border between either pair.

C Additional Experiments

To examine how the proposed correlation metric evolves as the model is progressively exposed to training data, we start from the OLMo-2-0425-1B stage1-step0-tokens0B checkpoint and pretrain the model on Wiki-CN[12] for a single epoch. We study the behavior of the correlation metric under a single-pass setting, where each example is used for training only once. This avoids repeated exposure to the same samples and allows us to more directly track how the metric changes as training

[12] https://huggingface.co/datasets/0xDing/wikipedia-cn-20230720-filtered

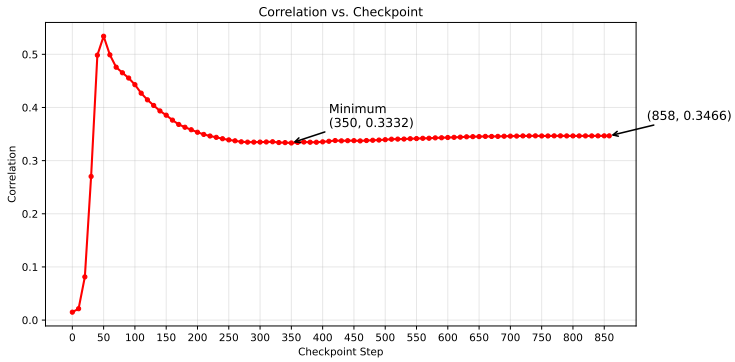

Figure 6: Evolution of the proposed correlation metric during single-pass pre-training on Wiki-CN. Starting from the OLMo-2-0425-1B stage1-step0-tokens0B checkpoint, we save checkpoints every 10 optimization steps and evaluate the correlation metric throughout one epoch of training. The correlation rises sharply early in training, then decreases and gradually stabilizes.

progresses. We save a checkpoint every 10 optimization steps and compute the proposed correlation metric at each checkpoint.

As shown in Figure 6, the correlation increases rapidly during the early stage of training, followed by a gradual decline. After reaching a post-peak minimum around step 350, the metric remains relatively stable for the remainder of the epoch. This suggests that the proposed correlation metric changes substantially during the initial phase of training before converging to a relatively stable regime.

This could be explained by the fact that the transformer first focuses on learning the bigram statistics of the training data, before learning more nuanced statistics such as the Mixed-State Presentation (Shai et al., 2024), which lowers the correlation.