Structure Before Collapse: Transient semantic geometry in next-token prediction

Yize Zhao[1], Isabel Papadimitriou[2]* & Christos Thrampoulidis[1]*
[1]Department of Electrical and Computer Engineering, [2]Department of Linguistics
The University of British Columbia
{zhaoyize, cthrampo}@ece.ubc.ca
isabel.papadimitriou@ubc.ca

Abstract

Neural Collapse predicts that balanced one-hot classification pushes model representations to be equally far from each other; a symmetric configuration that depends only on the output label and ignores any semantic similarity in the inputs. This creates a puzzle: next-token prediction language models are trained predominantly (as context length increases) with one-hot labels: the same context is very unlikely to appear twice in training with different labels. However, they clearly learn latent structural features. That is, despite the one-hot training regime, a language model's contextual embeddings represent the fact that the next word in "Mary broke the _____" is likely to be filled by tokens in the latent classes of a) medium-sized, b) rigid, c) inanimate nouns. How does gradient descent find such categorical semantic structure when co-occurrence statistics collapse to one-hot sparsity, eliminating any shared next-tokens among different contexts? To investigate this tension we identify three synthetic controlled settings where inputs have latent semantic factors but are mapped to distinct one-hot labels. We find that semantic geometry emerges early in training, and that representations cluster by shared attributes despite receiving no explicit supervision to do so. This structure is transient: with sufficient capacity and time, the model eventually reaches the predicted symmetric state where all representations are equally separated. We study this phase transition through Gram matrix analysis and propose a preliminary modification to the commonly used unconstrained features model to capture the emergent semantic geometry.

1 Introduction

How do language models learn latent structure in language when, overwhelmingly, the long training contexts that they use to learn next-word-prediction are unique (Goodman, 2001; Buck et al., 2014; Lu et al., 2025)? Acquiring linguistic capabilities involves deducing the multiple levels of latent structure in linguistic input, from semantic similarity relationships to abstract syntactic structure to the subtle rules around rare, semi-fossilized constructions. Though in practice language models encode and use much of this latent information, this is in fact in tension with current theory. For cases of balanced one-hot supervision, (where each context is only paired with one possible label in the training data) the literature on Neural Collapse ($\mathcal{NC}$, Papyan et al., 2020) predicts that learnt representations should form a symmetric Simplex Equiangular Tight Frame (ETF) in which representations are maximally separated and dictated solely by the one-hot labels and retain no trace of the underlying input semantics. [1]

*Equal senior authorship.

[1]Throughout, we use "semantics" as a shorthand to mean structured relationships of similarity between contexts, but we do not take a theoretical stance on whether the structures in our synthetic empirical setups solely reflect linguistic semantics, or if they also apply to syntactic and other relationships. We discuss this in more detail in Experiment 3.

Where, then, does this semantic structure come from? One classical answer locates it in the distribution of co-occurring labels: contexts with similar meanings tend to be followed by similar distributions of words. A long line of distributional methods, from n-gram models to word2vec, recovers semantic relationships by exploiting exactly this label co-occurence signal (Mikolov et al., 2013b; Levy & Goldberg, 2014; Pennington et al., 2014). Recent theoretical work makes an analogous mechanism precise for next-token prediction: When a context is paired not with a single label but with a sparse soft distribution over labels (e.g., "San Francisco is famous for its" followed by "skyscrapers", "beaches", and "restaurants", but not by "sun"), shared labels act as a bridge that aligns the representations of distinct contexts, yielding a meaningful and semantically rich geometry (Zhao et al., 2024; Zhao & Thrampoulidis, 2025). However, the probability of encountering these bridges through repeated contexts vanishes as context length grows. The majority of next-token supervision in modern long-context training therefore falls into the one-hot regime. This brings us back to the central questions motivating our experiments:

How can Gradient Descent find structured semantic geometry when the data is strictly one-hot? If the terminal state of optimization is a symmetric ETF that erases latent relationships between labels, under what conditions does such a semantic geometry emerge?

In this paper, we aim to draw an empirical and theoretical bridge between the theory on Neural Collapse and the realities of language model training. We do this by proposing a set of controlled synthetic toy languages with known latent structuring similarities, and train a suite of transformer models to model them, observing the representations formed during training. For our toy languages, we can accurately describe a representational structure (in terms of a representational similarity, or, Gram matrix) that leverages the semantic similarities, and contrast it with the ETF representational structure that's predicted by the NC theory, that relies only on the one-hot label and not the structure of the context. This way, we can accurately track the training dynamics of structured semantic representation.

Our main finding is that, in the synthetic language learning experiments that we run, **Gradient Descent recovers structured semantic geometry organized by shared latent factors early in training** (despite purely one-hot next-token supervision) before collapsing to a terminal ETF geometry organized only by output labels. Our contributions are summarized as follows:

1. We identify a previously unexamined contradiction: Neural Collapse theory predicts that one-hot supervision erases latent semantic structure from representations, yet language models trained under effectively one-hot regimes demonstrably learn such structure.
2. We introduce an empirical paradigm for tracking whether learned representations reflect latent semantic structure or collapse to the ETF geometry predicted by $\mathcal{NC}$: Representational Similarity Analysis (RSA) to latent semantic categories in synthetic languages.
3. We identify three synthetic languages where optimization recovers structured, semantic geometry, despite purely one-hot next-token supervision, and show that a) this semantic geometry is a transient phase occurring early in training, and b) given sufficient model capacity and training time, it eventually yields its place to a symmetric ETF, reconciling our findings with standard $\mathcal{NC}$ theory.
4. We present a simplified mathematical model that empirically and theoretically successfully reproduces this transient semantic alignment.

2 Background and related work

2.1 Neural Collapse geometries

We use the term NC geometry to refer to the geometric arrangement of last-layer representations of a deep-net in the Terminal Phase of Training, where training loss is driven to zero. Under class-balanced one-hot labeled data in K balanced classes, Papyan et al. (2020) found empirically that cross-entropy training yields a highly-symmetric NC geometry: a simplex equiangular tight frame (ETF) Gram matrix: $G_{\text{ETF}} := (I_K - \frac{1}{K} 1_K 1_K^\top) \otimes 1_{n/K} 1_{n/K}^\top$.

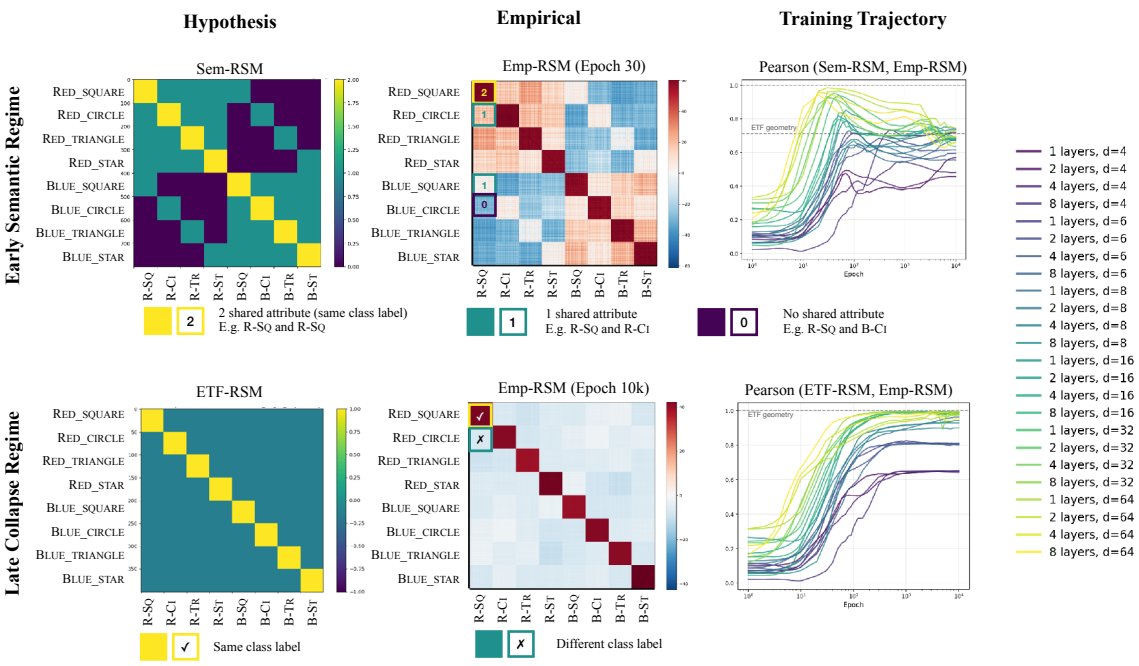


Figure 1: **Transient semantic geometry emerges before collapse** (Experiment 1). Sample-level $n \times n$ RSMs over the full training set, ordered by output label ($n = 800$, $K = 8$). **Left column:** hypothesis RSMs. Sem-RSM encodes latent semantic overlap between class pairs, with values 2, 1, and 0 equal to the number of shared attributes. ETF-RSM encodes the label-driven collapsed geometry predicted by Neural Collapse. **Middle column:** empirical RSMs (Emp-RSM) from the largest model (8 layers, $d = 64$) at epochs 30 and 10k. Early in training, the empirical geometry reflects the semantic hierarchy in Sem-RSM; by the end of training, this graded structure weakens and Emp-RSM moves toward the binary same-label versus different-label pattern of ETF-RSM. **Right column:** Pearson correlation between Emp-RSM and each hypothesis RSM over training, across model depths and widths. Correlation with Sem-RSM rises early and then declines, while correlation with ETF-RSM increases later, especially for larger models, indicating a transition from an early semantic regime to a late collapse regime.

Embeddings of examples belonging to the *same* class collapse to the same representation, and embeddings of examples belonging to *different* classes are equally (in fact, maximally) separated from each other: representations of images of "cars" are equally separated from representations of images of "trucks" as of "dogs".

Two theoretical directions have shown the $\mathcal{NC}$ geometry to have more structure. Firstly, when classes are imbalanced, the geometry converges to a generalization of ETF, the SELI geometry, where similarity relations are structured according to output label frequency similarities (Thrampoulidis et al., 2022). Secondly, in cases of sparse soft-labeled (rather than one-hot) data, the resulting geometry is the spectral components of a context-word co-occurrence matrix (Zhao et al., 2024; Zhao & Thrampoulidis, 2025). In both cases, however, the *sole driver of the similarity structure* continues to be the output labels. A key factor in this limitation is the unconstrained features model (UFM, Mixon et al., 2020; Fang et al., 2021; Zhu et al., 2021), which is predominantly employed in $\mathcal{NC}$ theory and drives the above geometry characterizations but ignores input-data structure by design. Prior work has empirically examined the role of input semantics in $\mathcal{NC}$ in the context of one-hot image classification (Yang et al., 2023; Jiang et al., 2024a), showing that in various empirical settings, deep-net classifiers trained on CIFAR labels (or, artificial super-class labels) lead to semantic representations reflecting class relations, seemingly contradicting the ETF theory.

Although we study language data, our findings shed light on the questions left open by these prior works: semantic geometry does appear in the representations of deep-nets, but as a transient phase early in training, before eventually collapsing to ETF.

2.2 Using synthetic languages to understand language model learning

Our methodology builds on a line of inquiry in language model interpretability of training models on carefully controlled synthetic data in order to understand their language learning capabilities. Representational analyses on models trained on Markov chains (Shai et al., 2024), hierarchical grammars (Murty et al., 2023; Allen-Zhu & Li, 2023; Schulz et al., 2025; Cagnetta & Wyart, 2024), language-like templates (Ahuja et al., 2025; McCoy et al., 2020; Hupkes et al., 2020; Petty & Frank, 2021), and simple synthetic worlds (Li et al., 2021; 2023; Lampinen et al., 2025) show how models form latent generalizations that correspond to the data-generating process, while work that investigates how these latent abstractions connect to natural language trains models on both synthetic and natural languages in controlled contexts (Papadimitriou & Jurafsky, 2020; 2023; Hu et al., 2025; Lee et al., 2026). To understand more linguistically-grounded aspects of learning, Kallini et al. (2024), Misra & Mahowald (2024), and Patil et al. (2024) train models on synthetically altered versions of natural language, to pinpoint specific learning effects. Recent papers have also used synthetic languages to understand the training dynamics of generalization (Qin et al., 2024; Zucchet et al., 2025; Wang et al., 2024; Zhao et al., 2026; Gu et al., 2025; Behnia et al., 2025; Vasudeva et al., 2026). These papers establish that trained language models often contain geometrically organized semantic representations. We ask a complementary question: when such geometry appears during training, and how it evolves under strictly one-hot next-token supervision.

2.3 Representation geometry in language models

That semantic and syntactic structure appears geometrically in language representations was established early in distributional word representations, where such regularities in embedding space are driven by word co-occurrence statistics (Mikolov et al., 2013c; Pennington et al., 2014; Levy & Goldberg, 2014; Arora et al., 2016; Levy et al., 2015); probing and visualization studies carry this into contextual representations (Hewitt & Manning, 2019; Tenney et al., 2019; Reif et al., 2019), and more recent work finds semantic information, latent task or world states, and interpretable features organized geometrically in trained models (Li et al., 2021; Nanda et al., 2023; Shai et al., 2024; Gurnee & Tegmark, 2024; Marks & Tegmark, 2023; Wu et al., 2025; Cunningham et al., 2023; Elhage et al., 2022; Park et al., 2024; 2025; Jiang et al., 2024b; Engels et al., 2024). Throughout this tradition, semantic geometry is tied to token co-occurrence statistics. The line of work most directly connected to ours brings this connection of semantic geometry to co-occurrence statistics into the next-token-prediction setting Thrampoulidis (2024); Zhao et al. (2024), where the soft distribution over next tokens acts as a bridge aligning the representations of distinct contexts into a semantic geometry (Zhao & Thrampoulidis, 2025; Yao & Xu, 2025; Karkada et al., 2026; Noroozizadeh et al., 2025; Behnia & Thrampoulidis, 2026). We instead ask here *when* semantic geometry emerges without it, and how it evolves relative to the label-driven geometry of neural collapse.

Beyond the geometry of trained models, a related line of work asks how representations evolve during training. Studies have examined how the representation geometry changes during different phases in training (Li et al., 2025), and more specifically how linguistic features emerge, are expressed in representation and weight space, and connect to overall model loss (Saphra & Lopez, 2019; Chen et al., 2024; Zhao & Thrampoulidis, 2025). We study this dynamics of representation geometry during training in controlled one-hot settings, where the latent semantic structure is known and can be directly compared against the collapse geometry predicted by neural collapse.

3 Overview of our empirical approach

We develop three controlled settings where diverse input contexts share latent semantic features but are mapped to one-hot targets. All three toy languages share this latent-semantics-yet-one-hot character, but they get progressively more language-inspired (and therefore complex). We train language models of varying widths and depths on our toy languages, and observe their training trajectories. All of our toy languages are posed to the model as a single prediction per example (predicting the next word of a sequence), rather than using a full next-token loss over all sequence positions. In this way, we can ensure that the context-label distribution is strictly one-hot, making the one-hot nature of the supervision as explicit as possible.

3.1 Training setup and independent variables: Model depth, width, and training time

Across all experiments, we train standard transformer decoders on the synthetic languages and track embedding geometry throughout training. For our independent variables, we sweep across different model depths, embedding dimensions, and measure our dependent variables along training. (Specific details in Appendix A.1).

3.2 Dependent variable: Representational similarity to semantic and collapsed geometries

To answer "what similarity structure do the context embeddings of a language model have?," we measure the representational similarity between how the models encode the inputs, and (a) the latent semantic similarities of data versus (b) the surface-level output label class.

Representational Similarity Analysis (RSA, Kriegeskorte et al., 2008; Nili et al., 2014) is a method comparing two spaces that both have a similarity metric defined. RSA works by measuring the correlation between pairwise Representational Similarity Matrices (RSMs) for a number of input samples, n.[2] We track the Pearson correlation between the RSM arrived at empirically by our models (Emp-RSM) with two hypothesis RSMs: a semantic RSM that depends on the language's latent semantic classes (Sem-RSM), and an equiangular tight frame RSM that puts maximal distance between all inputs with different labels (ETF-RSM).

Sem-RSM We define a similarity metric over input context strings based on the number of shared latent classes that a context selects for (eg, if one context selects for small, inanimate nouns and another selects for large, inanimate nouns, they will have a similarity of 1 shared latent class). Since we perfectly know the data-generating process of our toy languages, we can construct an idealized representational similarity matrix where *the only factors* influencing the similarity between two context representations are latent semantic classes.

ETF-RSM We also create an alternative hypothesis RSM: that of the simplex equiangular tight frame geometry (ETF-RSM). In this case, two input contexts have a similarity of 1 if they take the same output label, and a similarity of 0 otherwise, ignoring any latent semantics that contribute to the contexts.[3] In this idealized RSM, the only feature influencing the similarity between two contexts is their output label.[4]

Emp-RSM We compare these two hypothesis RSMs to an empirical RSM of the model context embeddings (Emp-RSM). The RSM of our context embeddings is simply the Gram matrix of embeddings. For each model, at each training epoch, we take the model's last-layer representations of every input in our training data, and construct the Gram matrix of their inner products. The two dependent variables that we measure are the Pearson correlations between each model checkpoint's context representations and the Sem-RSM and ETF-RSM.

Complementary Metrics. In addition to the core metrics listed above, we use the following complementary diagnostics to separate semantic structure from label-driven collapse.

Disentangling semantic and label effects. To separate semantic alignment from same-label collapse, we use two different-label diagnostics. The first compares whether different-label pairs that share a latent factor are more similar than different-label pairs that share none. The second recomputes semantic RSA after removing all same-label pairs. Complementary to our main results, these checks, that are presented in Appendix A.4, confirm that the semantic signal reflects graded structure across labels, rather than only clustering within each exact output label.

[2]RSA is often defined in terms of dissimilarity matrices (rather than similarity matrices), but this is equivalent. When our similarity metric is an inner product, the RSM is a Gram matrix.

[3]Though, of course, output labels do contain some level of semantic information, we contrast this to the Sem-RSM which represents the full latent system underlying the data-generating process

[4]The ETF geometry assumes balanced output labels, i.e. equal number of input contexts per label. In our experiments with imbalanced output labels, we use the SELI geometry generalization of ETF, though we maintain the naming ETF-RSM since the common feature of both geometries distinguishing them from Sem-RSM, is that they are solely driven by output labels.

NC-style collapse diagnostics. We provide direct collapse-style measurements in Appendix A.2 and Appendix A.3. First, we report standard NC1 and NC2 diagnostics, measuring within-label variability and convergence of class means toward the ETF geometry (Papyan et al., 2020). Second, we adapt this idea to semantic attributes, measuring whether examples sharing the same latent attribute form tight clusters. The results support the same training picture as RSA: semantic organization is strongest early, while label-driven collapse strengthens later.

Cross-layer geometry. In Appendix A.8, we repeat the semantic and label-driven measurements across layers, showing that label-driven geometry is generally strongest near the readout, while semantic structure can remain visible in earlier and intermediate layers. A detailed investigation of the geometry evolution across layers is left to future work.

In the following sections, we describe the three synthetic languages that we construct, and report how model embedding geometries correlate with Sem/ETF-RSM.

4 Experiment 1: Color-Shape categories language

Language definition Our first synthetic language is a minimal setting where the output labels depend on the cross of two latent semantic factors. Each input is defined by two factors (in our case, color and shape), and each output label reflects only the cross product of the two (eg, RED_CIRCLE is a one-hot label that only fires for inputs that are in *both* RED and CIRCLE) More formally:

Latent Factors. We define a Color Factor containing attributes $\mathcal{A}_\mathcal{C} = \{\text{RED}, \text{BLUE}\}$ and a Shape Factor ($F = \mathcal{H}$) containing attributes $\mathcal{A}_\mathcal{H} = \{\text{CIRCLE}, \text{SQUARE}, \text{TRIANGLE}, \text{STAR}\}$.

Semantic Class. We define a Semantic Class $\mathcal{S}_a$ as the set of all samples sharing a specific attribute (like, "Red"), regardless of the other factors. These classes are not mutually exclusive: every sample belongs to exactly one Color class and one Shape class simultaneously.

Label Class. The optimization objective is defined by the Label Class $\mathcal{Y}$. A label class y_k corresponds to the unique intersection of semantic factors, $y_k = (a_c, a_h)$, resulting in $K = |\mathcal{A}_\mathcal{C}| \times |\mathcal{A}_\mathcal{H}| = 8$ mutually exclusive one-hot targets. For example, a sample that is in $\mathcal{S}_{\text{RED}}$ and in $\mathcal{S}_{\text{CIRCLE}}$ will always get the label $y = \text{RED\_CIRCLE}$, which in terms of the label one-hot distribution has no relation to other labels like RED_SQUARE.

Context Samples. Each input is a triplet $x = [t_{\text{ID}}, t_c, t_h]$. To ensure the model doesn't simply memorize, we introduce variability within each class in the form of nuisance IDs (t_{ID}) for each sample. The contextual tokens t_c, t_h are drawn from sets of fine-grained variants (e.g., when $a = \text{RED}$, the variants are $\{\text{DARK\_RED}, \text{LIGHT\_RED}, \text{MEDIUM\_RED}\}$), so the model never observes the coarse latent factors directly. The input vocabulary ($V = 818$) comprises of 800 unique Nuisance IDs and 18 Contextual Tokens (3 fine-grained variants for each of the 2 color and 4 shape attributes). The total dataset consists of $n = 800$ samples. Putting this all together, an example of our prediction objective is:

$$x = [768 \quad \text{DARK\_RED} \quad \text{SMALL\_CIRCLE}], \qquad y = \text{RED\_CIRCLE}\,.$$

Results We present our results in Figure 1. We track the models' representational similarity to the two hypothesis RSMs on the left: Sem-RSM and ETF-RSM. Our main result is that models transiently reach maximal similarity to the semantic representation early in training (right, top), and then collapse toward the ETF geometry, which they approach monotonically throughout training (right, bottom). We also observe that the largest models (yellower lines) most effectively find both the semantic geometry early in training and the ETF geometry at the terminal phase, so semantic geometry is not a result of representational bottlenecking. Surprisingly, for some models, RSA with Sem-RSM is close to 1, meaning that the *only* representational structure is our hypothesized semantic structure, with almost no spurious training artifacts.

Remark 1 (Relation to the linear representation hypothesis)**.** *The linear representation hypothesis suggests that changing one latent factor while holding others fixed should correspond to a consistent direction in representation space (Mikolov et al., 2013c; Pennington et al., 2014; Arora et al., 2016; Park et al., 2024; Jiang et al., 2024b). Experiment 1 provides a simple test of this idea*

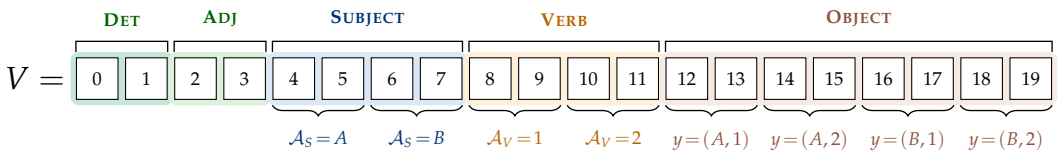


Figure 2: Linear language from Experiment 2: an illustrative example 20-token vocab with $k = 2$ classes, excluding the nuisance IDs. A valid training pair would be $x = [id100 \quad 1 \quad 2 \quad 2 \quad 5 \quad 10 \quad 2]$, $y = 15$. The SUBJECT (5) is in class A, and the VERB (10) is in class 2, so the predicted object would have to be in class $(A, 2)$: $y = 14$ or $y = 15$. We use $|V| = 62$ total non-id tokens, and $k = 10$ categories for subjects and verbs.

because color and shape vary independently. For example, the CIRCLE-to-SQUARE contrast is context-invariant if the corresponding representation differences under RED and BLUE are aligned. The exact Sem-RSM implies that these difference vectors coincide, whereas the exact ETF-RSM implies that they are orthogonal. Thus, the observed transition from Sem-RSM-like to ETF-like geometry suggests that such linear organization is transient: a shared direction for the same latent contrast emerges early and weakens as label-level collapse develops. Relatedly, Yang et al. (2026) study the training dynamics of cross-context concept-direction alignment, which they call abstraction, in a solvable linear-network setting and show that it can be non-monotonic during training.

5 Experiment 2: linear grammar with Zipfian latent categories

Language definition For our second language, we make the setup of Experiment 1 slightly more language-like. We create a synthetic language following a linear template grammar, with categories drawn from a Zipfian distribution. Our vocabulary V is divided into 5 classes, corresponding to parts of speech and grammatical roles: Determiners, Adjectives, Subjects, Verbs, and Objects, plus a nuisance ID to make the contexts strictly one-hot.[5] Training samples follow the template below, where indicates adjective lists of lengths 0-3:

$$x = [\text{ID} \quad \text{DET} \quad \text{ADJ}^* \quad \text{SUBJECT}_{\mathcal{A}_S} \quad \text{VERB}_{\mathcal{A}_V} \quad \text{DET} \quad \text{ADJ}^*] \qquad y = \text{OBJECT}_{(\mathcal{A}_S \mathcal{A}_V)}.$$

Two latent semantic factors are the subject subclass $\mathcal{A}_S$ and verb subclass $\mathcal{A}_V$, each with k sub-classes. The prediction label, the OBJECT, has $k \times k = |\mathcal{A}_S \times \mathcal{A}_V|$ sub-classes. Multiple tokens occur within each sub-class, so the sub-class remains latent in the input; see Figure 2. To make the language more natural, we sample $\mathcal{A}_V$ from a Zipfian distribution (see Piantadosi, 2014, for an analysis of the pervasiveness of Zipfian distributions in natural language). Thus, we use the imbalanced generalization of ETF (SELI geometry) as our ETF-RSM.

In making the object class reliant on a cross-product of the subject class and the verb class, we draw inspiration from how semantic features restrict vocabulary choices in natural language. In a sentence like *The girl ate the _*, the next word is constrained by both the verb and the subject: it should be compatible with eating and plausible for the subject. In *The cat devoured the _*, the verb still selects for an edible object, but the subject changes which objects are most plausible, e.g., "tuna" rather than "cookie".

For greater theoretical simplicity, we make the non-realistic assumption that SUBJECTs and OBJECTs are non-overlapping sets of the vocabulary, and that the elements of every OBJECT subclass are also non-overlapping.

Results Our results in Figure 3 show a similar pattern to Experiment 1: the Empirical RSMs match the hypothesized semantic RSM (left) early in training, before moving away from this structure, more strongly for larger-capacity models. Interestingly, at the terminal phase, though the representation has moved away from the semantic structure, it has not totally collapsed to a diagonal-like label structure: significant fluctuations remain in the off-diagonals of the empirical RSM. Though this is not strictly the ETF-RSM's SELI structure (see Appendix A.6, Figure 9 for convergence to the ETF-RSM), it likely reflects the more unstable learning dynamics of the Zipfian distribution, which we aim to explore in future work.

[5]Though we use "part of speech" as an intuitive term, the language is purely synthetic: the vocabulary tokens are integer token ids, and their mapping to PoS classes is arbitrary; see Figure 2.

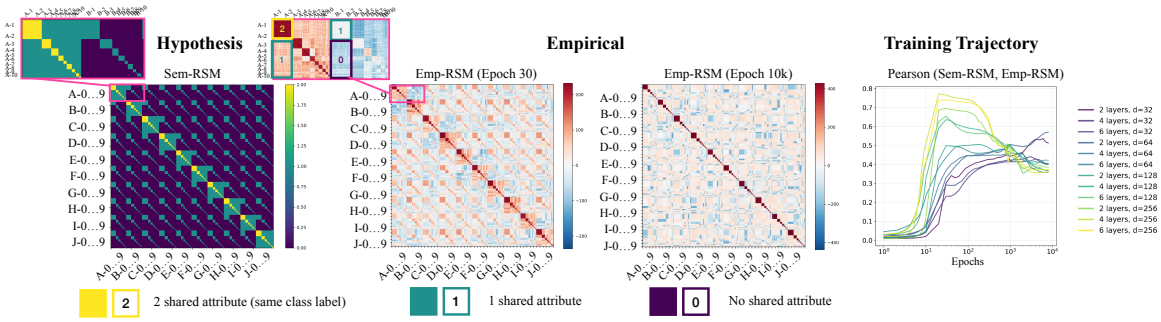


Figure 3: **Transient semantic geometry in the linear template Zipfian language** (Experiment 2). Sample-level $n \times n$ RSMs over the full training set, ordered by semantic and output label ($n = 10$k, $K = 100$). **Left:** Sem-RSM, where similarity is determined by the number of shared latent categories. Because the label distribution is imbalanced, block sizes are unequal (see inset). **Middle:** empirical RSMs (Emp-RSM) at epochs 30 and 10k. Early in training, Emp-RSM reflects the off-diagonal semantic structure of Sem-RSM; later, it weakens. **Right:** Pearson correlation between Emp-RSM and Sem-RSM over training, across model depths and widths. As in Experiment 1, larger models peak higher and earlier.

6 Experiment 3: hierarchical grammar with latent categories

Language definition Our third synthetic language combines a hierarchical grammatical structure with a similar category-selection mechanic as our previous two languages, but where *which* context token it is that defines the category of the label depends on the tree structure (more specifically, the c-command relation). With this language, we aim to capture a controlled toy version of the interactions between grammatical and semantic structure in natural language, where the ways in which meaning restrictions (like tense or negation scope) propagate to future tokens is often dictated by syntactic tree operations.

For our hierarchical grammar every node in our tree can expand either by coordination ($XP \rightarrow XP\ XP$), right-branching ($XP \rightarrow X\ XP$), or trivial projection ($XP \rightarrow X$). Each vocabulary item is also tied to a semantic category, $\mathcal{A}_C$. To sample an x, y pair, we sample a string from the grammar, where the semantic category $\mathcal{A}_C$ is randomly selected for all of the tokens except the last one. The string minus the last token is then our x. The last token, y, is restricted to be the semantic class of the **closest token in x that c-commands it**. C-command is a relation between two nodes in a tree, where A c-commands B if A's direct parent dominates all of B's ancestors (see the red arrow in Figure 4 for an example of c-command), and is an important tree relation in generative linguistics for describing processes like pronoun binding (Reinhart, 1976; Chomsky, 1988), negation scope (Ladusaw, 1979), quantifier scope (May, 1985), and tense (Kratzer, 1998).

To make the model able to induce the tree structure from the language strings, we have two options. The one option would be to have multiple parts of speech and have the model induce the tree structure as in standard grammar induction paradigms. The second option (which we take) is to keep just a single part of speech (X), and mark the internal tree structure in the form of parentheses: interleaving in a Dyck language that has the same internal nodes (see the Figure 4 for an example of such a linearization). We elect to have just a single part of speech in order to not conflate the learning of semantic categories with the grammar induction of parts of speech. Having just one part of speech means that the *only relevant categories* for the model to represent are our semantic categories, and so we can measure the RSA to a ground-truth categorical geometric representation.

In future work, we aim to examine the interplay between jointly inducing grammar and learning category-selection semantics. Doing this would require careful and varied analysis of the representation space, as there is no single ground truth similarity matrix. The interface and interference of syntactic and semantic categories is a fundamental topic of analysis across linguistic traditions, and quantifying "ground-truth" representations to represent different hypotheses is a nuanced and multifaceted process. This type of sweep and analysis is beyond the scope of what the current experiment is contributing to this paper's thesis (a hierarchical processing data point), and so we leave it to future work.

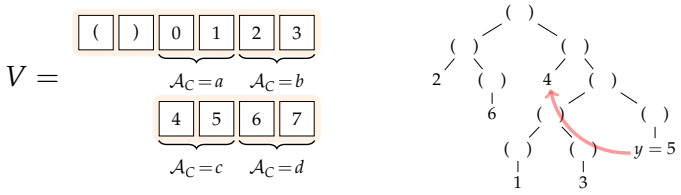


Figure 4: An illustrative example of our Experiment 3 hierarchical language with $|V| = 10, k = 2$. To linearize the tree by interleaving our probabilistic context-free grammar with the Dyck language, we produce an opening or closing bracket when an internal node is first or last encountered. Thus, the training pair for the pictured tree is $x =$ ((2 (6)) (4 ((((1) (3)) (, $y = 5$. Here, y belongs to $\mathcal{A}_c = c$, the class of the token that c-commands it (4). Trailing closing parentheses are excluded.

Results Figure 5 shows the same qualitative pattern as in the first two experiments, now in a hierarchical setting. Early in training, the empirical RSM closely matches the semantic hypothesis: examples controlled by the same latent category form clear blocks, and the correlation with Sem-RSM rises sharply to a high peak. This indicates that the model organizes representations by the latent semantic category relevant for prediction, even though that category is mediated by tree structure rather than a fixed linear position. Note that the RSA to the semantic representation peaks later, and the semantic block-diagonals are still faintly present even in the terminal phase of training. This effect is strongest for the smaller-width models ($d = 8, 16$), which keep higher semantic alignment than the larger-width models. One possible explanation is that the hierarchical structure is both harder to learn and acts as a weak regularizer towards the lower-rank semantic organization, especially when width is limited and the model cannot as easily move to a finer label-level geometry. This is an avenue for exploring in future work on linguistic learning dynamics.

7 Theoretical proxy: spherical Bag-of-Words model

We introduce a minimal Spherical Bag-of-Words model that reproduces the same qualitative transient as the transformer experiments. This gives a simpler setting where we can study the mechanism more directly. In this model, representations first align with input-side latent structure, and later move toward label-driven geometry. Thus, the transient does not seem to require self-attention or depth, but can already appear from the interaction between input structure, optimization, and the geometry of the objective.

Our architecture is a *Spherical continuous bag-of-words (BoW)* model that trains two weight matrices $W \in \mathbb{R}^{K \times d}$ and $H \in \mathbb{R}^{d \times V}$ by minimizing the CE loss between input token encoding matrix $X \in \{0,1\}^{V \times n}$ and one-hot label matrix $Y \in \{0,1\}^{K \times n}$: $\mathcal{L}(W, H) = \mathrm{KL}\left(Y || \mathrm{Softmax}(W \cdot \mathsf{N}(HX))\right)$. Here, K is the number of label classes, V is the token vocabulary of the input contexts, and d the hidden dimension, the KL operator averages over divergence between columns of the two input matrices, and, $\mathsf{N}(\cdot)$ denotes column-wise normalization onto the unit Euclidean sphere. Note this is identical to a continuous BoW (Mikolov et al., 2013a) model except for the normalization, which however we find is crucial to establish terminal convergence to ETF.

We study this Spherical BoW framework on our Experiment 1 language. Firstly, in Appendices B.3 and B.4 we show empirically that when trained with Gradient Descent with small random initialization and $d \geq K$, Spherical BoW possesses the desired transient semantic behavior: early on in training, the analytical Emp-RSM $G_{\text{emp}} := X^\top H^\top H X$ aligns with the data's Sem-RSM $G_{\text{sem}} := B^\top B \otimes 1_m 1_m^\top$ (here, B is the base matrix of shared semantic features; see appendix), before eventually converging to (a scaling of) the ETF-RSM $G_{\text{ETF}} = (I_K - \frac{1}{K} 1_K 1_K^\top) \otimes 1_m 1_m^\top$. Moreover, the larger the hidden dimension, the more pronounced this behavior, consistent with our earlier experiments.

Second, we provide mathematical justification for this optimization behavior. We prove two results (details in Appendix B.2). Theorem 1: Assuming balanced output logits at initialization and initial features that are weakly ϵ-correlated with the classifiers, the change in Emp-RSM after one step of gradient descent is biased towards the Sem-RSM provided $m > 1$ contexts per label class and $\epsilon \ll 1/\sqrt{K}$. Second, Theorem 2: assuming that gradient descent converges to a global minimizer, this minimizer possesses the ETF-RSM structure.

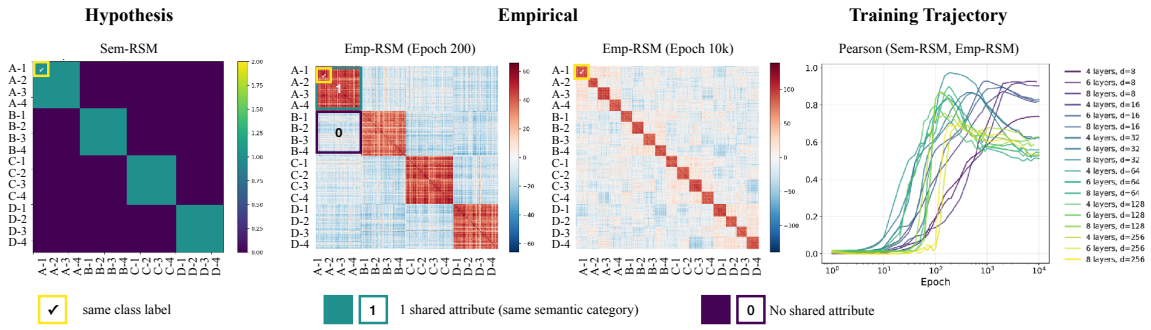


Figure 5: **Transient semantic geometry in the hierarchical grammar** (Experiment 3). Sample-level $n \times n$ RSMs over the full training set, ordered by output label ($n = 2000$, $K = 16$). **Left:** Sem-RSM, where similarity is 1 if two contexts share the same semantic category and 0 otherwise. **Middle:** empirical RSMs (Emp-RSM) at epochs 200 and 10k. Early in training, Emp-RSM exhibits clear category-level block structure matching Sem-RSM, indicating that contexts licensed by the same c-commanding semantic category are grouped together. By the end of training, this coarse semantic structure weakens and finer label-specific diagonal structure becomes more prominent. **Right:** Pearson correlation between Emp-RSM and Sem-RSM over training, across model depths and widths. All models enter a strong semantic regime early, but late-stage behavior differs: narrow models ($d = 8, 16$) retain higher semantic alignment, whereas wider models peak earlier and decline more.

Importantly, the spherical normalization is essential to guarantee this transition to ETF towards convergence; without it (in which case, it becomes vanilla BoW) the parameters converge to a semantic geometry. Thus, our spherical-BoW model can be interpreted as a middle-ground between vanilla BoW favoring only input-semantic geometry and an Unconstrained Feature Model (see Remark 2) favoring only output-label geometry.

8 Conclusion, discussion, and limitations

Our experiments bridge neural-collapse theory and language-model training by showing that semantic geometry can emerge transiently under strictly one-hot supervision. Across three increasingly language-like synthetic settings, including factor-crossed categories, a Zipfian linear grammar, and a hierarchical grammar, representations first organize by latent semantics and later move toward output-label-driven NC geometry. This occurs even in larger, non-bottlenecked models ($d \geq K$), reconciling early semantic structure with terminal collapse.

Fresh-data regime. Our experiments, like most NC studies, use multiple epochs, unlike typical LM training on mostly fresh data. Extending the analysis to this regime will require larger vocabularies and model capacities, and determining whether an analogous undertrained transient appears before convergence.

Mixed context lengths. Autoregressive training is not uniformly one-hot: short contexts may recur and induce soft next-token distributions, while long contexts are effectively unique. Future work should test whether these relatively few short-context predictions bootstrap semantic representation or have negligible effect.

More realistic synthetic languages. Our ground-truth Sem-RSM is possible because the latent structure is simple category restriction. Richer languages jointly involve syntactic and semantic factors, for which no single ground-truth geometry may suffice; studying their interaction is an important next step.

Generalization. ETF geometry has been linked to generalization in image classification (Papyan et al., 2020), with initial extensions to language modeling by Wu & Papyan (2024). It remains to test whether transient semantic geometry similarly correlates with or facilitates generalization.

Acknowledgments

This work was partially funded by the NSERC Discovery Grant No. 2021-03677, the Alliance Grant ALLRP 581098-22, and a gift from Google. YZ was also supported by the UBC 4YF Doctoral Fellowship. The authors also acknowledge the use of the Sockeye cluster by UBC Advanced Research Computing. We also thank the anonymous reviewers from the COLM 2026 review process for their thoughtful feedback.

References

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A Smith, Navin Goyal, and Yulia Tsvetkov. Learning syntax without planting trees: Understanding hierarchical generalization in transformers. *Transactions of the Association for Computational Linguistics*, 13:121–141, 2025.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, learning hierarchical language structures. 2023. URL https://api.semanticscholar.org/CorpusID:270210215.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL https://aclanthology.org/Q16-1028/.

Tina Behnia and Christos Thrampoulidis. How cross-entropy shapes representation geometry: A spectral study on cycle graphs. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026.

Tina Behnia, Ganesh Ramachandra Kini, Vala Vakilian, and Christos Thrampoulidis. On the implicit geometry of cross-entropy parameterizations for label-imbalanced data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10815–10838. PMLR, 2023.

Tina Behnia, Puneesh Deora, and Christos Thrampoulidis. Facts in stats: Impacts of pretraining diversity on language model generalization. *arXiv preprint arXiv:2510.16096*, 2025.

Christian Buck, Kenneth Heafield, and Bas Van Ooyen. N-gram counts and language models from the common crawl. 2014.

Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. *Advances in Neural Information Processing Systems*, 37:83119–83163, 2024.

Yuan Cao, Difan Zou, Yuanzhi Li, and Quanquan Gu. The implicit bias of batch normalization in linear models and two-layer linear convolutional neural networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5699–5753. PMLR, 2023.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. In *International Conference on Learning Representations*, 2024.

Noam Chomsky. *Lectures on Government and Binding: The Pisa Lectures*. Foris Publications, 1988.

Hung-Hsu Chou, Holger Rauhut, and Rachel Ward. Robust implicit regularization via weight normalization. *Information and Inference: A Journal of the IMA*, 13(3):iaae022, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are linear, 2024. URL https://arxiv.org/abs/2405.14860.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

Connall Garrod and Jonathan P Keating. The persistence of neural collapse despite low-rank bias: An analytic perspective through unconstrained features. *arXiv preprint arXiv:2410.23169*, 2024.

Joshua T Goodman. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434, 2001.

Xinran Gu, Kaifeng Lyu, Jiazheng Li, and Jingzhao Zhang. Data mixing can induce phase transitions in knowledge acquisition. *arXiv preprint arXiv:2505.18091*, 2025.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 4124–4135, 2019.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Michael Y Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. Between circuits and chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9691–9709, 2025.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020.

Arthur Jacot, Peter Súkeník, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse. *arXiv preprint arXiv:2410.04887*, 2024.

Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. In *International Conference in Machine Learning (ICML)*. Proceedings of Machine Learning Research, 2024a.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911. PMLR, 2024b. doi: 10.48550/arXiv.2403.03867. URL https://proceedings.mlr.press/v235/jiang24d.html.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible language models. 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. 14691–14714, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.787. URL https://aclanthology.org/2024.acl-long.787/.

Dhruva Karkada, James Simon, Yasaman Bahri, and Michael Deweese. Closed-form training dynamics reveal learned features and linear structure in word2vec-like models. *Advances in Neural Information Processing Systems*, 38:3967–4000, 2026.

Angelika Kratzer. More structural analogies between pronouns and tenses. In *Semantics and linguistic theory*, pp. 92–110, 1998.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

William Allen Ladusaw. *Polarity sensitivity as inherent scope relations.* The University of Texas at Austin, 1979.

Andrew K Lampinen, Arslan Chaudhry, Stephanie CY Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James L McClelland. On the generalization of language models from in-context learning and finetuning: a controlled study. *arXiv preprint arXiv:2505.00661*, 2025.

Dan Lee, Seungwook Han, Akarsh Kumar, and Pulkit Agrawal. Training language models via neural cellular automata. *arXiv preprint arXiv:2603.10055*, 2026.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 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.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (Volume 1: Long papers)*, pp. 1813–1827, 2021.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Melody Zixuan Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Kumar Teru, Adam Santoro, Guillaume Lajoie, and Blake A. Richards. Tracing the representation geometry of language models from pretraining to post-training. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2509.23024. URL https://openreview.net/forum?id=FDruZlKWUb.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Raghavi Chandu, Nouha Dziri, and Yejin Choi. AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ilOEOIqolQ.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023. URL https://arxiv.org/abs/2310.06824.

Robert May. *Logical form: Its structure and derivation*, volume 12. MIT press, 1985.

R Thomas McCoy, Robert Frank, and Tal Linzen. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140, 2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeff Dean. Efficient estimation of word representations in vector space. In *Workshop at ICLR*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013c.

Kanishka Misra and Kyle Mahowald. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 913–929, 2024.

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. Characterizing intrinsic compositionality in transformers with tree projections. In *The Eleventh International Conference on Learning Representations*, 2023.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models, 2023. URL https://arxiv.org/abs/2309.00941.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

Shahriar Noroozizadeh, Vaishnavh Nagarajan, Elan Rosenfeld, and Sanjiv Kumar. Deep sequence models tend to memorize geometrically; it is unclear why. In *Workshop on Foundations of Reasoning in Language Models (FoRLM), NeurIPS 2025*, 2025. doi: 10.48550/arXiv.2510.26745. URL https://openreview.net/forum?id=2NuCrYf8Ap.

Isabel Papadimitriou and Dan Jurafsky. Learning Music Helps You Read: Using transfer to study linguistic structure in 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. 6829–6839, Online, nov 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.554. URL https://aclanthology.org/2020.emnlp-main.554/.

Isabel Papadimitriou and Dan Jurafsky. Injecting structural hints: Using language models to study inductive biases in language learning. *arXiv preprint arXiv: 2304.13060*, 2023. URL https://arxiv.org/abs/2304.13060.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. doi: 10.48550/arXiv.2311.03658. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2406.01506. URL https://openreview.net/forum?id=bVTM2QKYuA.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. Filtered corpus training (fict) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615, 2024.

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

Jackson Petty and Robert Frank. Transformers generalize linearly. *arXiv preprint arXiv:2109.12036*, 2021.

Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

Tian Qin, Naomi Saphra, and David Alvarez-Melis. Sometimes i am a tree: Data drives unstable hierarchical generalization. *arXiv preprint arXiv:2412.04619*, 2024.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, 2019.

Tanya Miriam Reinhart. *The syntactic domain of anaphora*. PhD thesis, Massachusetts Institute of Technology, 1976.

Naomi Saphra and Adam Lopez. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3257–3267, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1329. URL https://aclanthology.org/N19-1329/.

Laura Ying Schulz, Daniel Mitropolsky, and Tomaso Poggio. Unraveling syntax: How language models learn context-free grammars. *arXiv preprint arXiv:2510.02524*, 2025.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 75012–75034. Curran Associates, Inc., 2024. doi: 10.52202/079017-2387. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8936fa1691764912d9519e1b5673ea66-Paper-Conference.pdf.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Peter Súkeník, Christoph H Lampert, and Marco Mondelli. Neural collapse is globally optimal in deep regularized resnets and transformers. *arXiv preprint arXiv:2505.15239*, 2025.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Christos Thrampoulidis. Implicit optimization bias of next-token prediction in linear models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Bhavya Vasudeva, Puneesh Deora, Alberto Bietti, Vatsal Sharan, and Christos Thrampoulidis. Understanding contextual recall in transformers: How finetuning enables in-context reasoning over pretraining knowledge. *arXiv preprint arXiv:2603.20969*, 2026.

Mingze Wang, Ruoxi Yu, Weinan E, and Lei Wu. How transformers get rich: Approximation and dynamics analysis. *arXiv preprint arXiv:2410.11474*, 2024.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. *Advances in Neural Information Processing Systems*, 37:137432–137473, 2024.

Zhaofeng Wu, Xinyan Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/84fb08ae30ab266953d1ff72eb0611b1-Abstract-Conference.html.

William W. Yang, Andrew M. Saxe, and Peter E. Latham. How are linear representations learned? exact solutions to the dynamics of abstraction. *arXiv preprint arXiv:2607.08843*, 2026.

Yongyi Yang, Jacob Steinhardt, and Wei Hu. Are neurons actually collapsed? on the fine-grained structure in neural representations. In *International Conference on Machine Learning*, pp. 39453–39487. PMLR, 2023.

Junjie Yao and Zhi-Qin John Xu. Probability signature: Bridging data semantics and embedding structure in language models. *arXiv preprint arXiv:2509.20124*, 2025.

Xingyu Zhao, Darsh Sharma, Rheeya Uppaal, and Yiqiao Zhong. Shattered compositionality: Counterintuitive learning dynamics of transformers for arithmetic. *arXiv preprint arXiv:2601.22510*, 2026.

Yize Zhao and Christos Thrampoulidis. Geometry of semantics in next-token prediction: How optimization implicitly organizes linguistic representations. *arXiv preprint arXiv:2505.08348*, 2025.

Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=qyilOnIRHI.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*, 2025.

A Additional experimental settings and results

A.1 Experimental details

Experiment 1. We sweep depth $L \in \{1, 2, 4, 8\}$ and embedding dimension $d \in \{4, 6, 8, 16, 32, 64\}$. All models use 2 attention heads. Training is performed with full-batch Adam ($\text{lr} = 5 \times 10^{-4}$) for 10^4 epochs.

Experiment 2. We sweep depth $L \in \{2, 4, 6\}$ and embedding dimension $d \in \{32, 64, 128, 256\}$. All models use 8 attention heads. Training is performed with Adam, batch size $B = 512$, learning rate 10^{-4}, for 10^4 epochs.

Experiment 3. We sweep depth $L \in \{4, 6, 8\}$ and embedding dimension $d \in \{8, 16, 32, 64, 128, 256\}$. All models use 8 attention heads. Training is performed with Adam, batch size $B = 512$, learning rate 5×10^{-4}, for 10^4 epochs.

A.2 Neural Collapse diagnostics

The RSA analyses compare the sample-level empirical RSM to the Sem-RSM and ETF/SELI-RSM. Since the ETF/SELI-RSM correlation is a similarity-structure diagnostic, we also report two standard Neural Collapse diagnostics, NC1 and NC2, which more directly measure within-class collapse and class-mean geometry (Papyan et al., 2020).

Let $H = [h_{1,1}, \ldots, h_{K,n_K}] \in \mathbb{R}^{d \times \sum n_k}$ be the feature matrix (last-layer embeddings) of examples belonging to K classes, such that there is equal number of examples $n_1 = \ldots = n_K$ per class. Let $W = [w_1, \ldots, w_K]^\top \in \mathbb{R}^{K \times d}$ be the classifier weight matrix. Let $\mu_k = \frac{1}{n_k} \sum_{i=1}^{n_k} h_{k,i}$ be the k-th class mean and $\mu_G = \frac{1}{\sum n_k} \sum_{k,i} h_{k,i}$ be the global mean.

$\mathcal{NC}1$ (Variability Collapse): The within-class covariance Σ_W vanishes. $\text{Metric}_{\mathcal{NC}1} = \Sigma_W \to 0$ where $\Sigma_W = \frac{1}{\sum n_k} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (h_{k,i} - \mu_k)(h_{k,i} - \mu_k)^\top$.

$\mathcal{NC}2$ (Convergence to ETF): The centered class means $\tilde{H} = [\mu_1 - \mu_G, \ldots, \mu_K - \mu_G]$ converge to a Simplex Equiangular Tight Frame (ETF).

$$\text{Metric}_{\mathcal{NC}2} = \left\| \frac{\tilde{H}^\top \tilde{H}}{\|\tilde{H}^\top \tilde{H}\|_F} - \frac{1}{\sqrt{K-1}} \left(I_K - \frac{1}{K} 1_K 1_K^\top \right) \right\|_F \to 0$$

NC2 predicts that the representations of examples belonging to different classes are equally (in fact, maximally) separated. Here and throughout, we assume $d \geq K$. This assumption guarantees the existence of an ETF and it is required for NC.

Figure 6 shows that the NC diagnostics are broadly consistent with the RSA results. NC1 decreases over training, indicating increasing within-class collapse. NC2 also tends to decrease, especially in larger models, indicating movement toward the label-driven class-mean geometry. Experiment 2 is the least direct comparison because the label distribution is imbalanced; there we compare against the SELI target rather than the balanced ETF target.

A.3 Numerical characterization of semantic structure

Motivated by the standard NC1 metric in Appendix A.2, we also consider a direct numerical measure of semantic clustering. Instead of grouping examples by output label, this metric groups examples by latent semantic attribute, and asks whether embeddings that share the same attribute form tight clusters.

Semantic Convergence ($\mathcal{NC}1_{\text{sem}}$). We compute the standard $\mathcal{NC}1$ metric with one modification: instead of grouping samples by their label class y_k, we group them by their semantic class (attribute) $a \in \mathcal{A}_F$. Let $\mu^{(a)}$ denote the centroid of the semantic class $\mathcal{S}_a$, and let $\mu_G^{(F)}$ denote the global mean over all embeddings associated with factor F. We define the

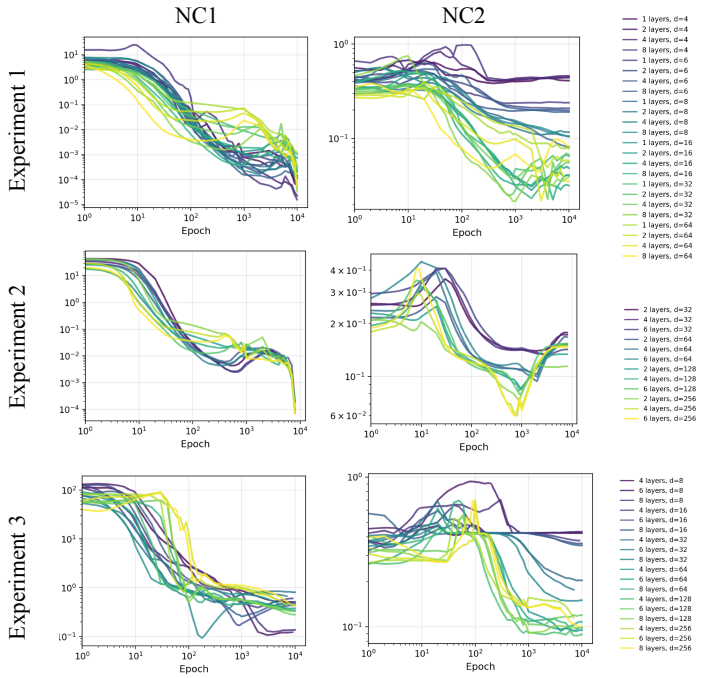


Figure 6: Neural Collapse diagnostics across three experiments. NC1 measures within-class variability relative to class-mean variability, and NC2 measures the deviation of the centered class-mean Gram matrix from the corresponding ETF/SELI target geometry. Across experiments, NC1 decreases sharply during training, while NC2 generally decreases more strongly for larger models, indicating that label-driven structure becomes stronger later in training.

within-class and between-class scatter matrices as

$$\Sigma_W(F) = \frac{1}{n} \sum_{a \in \mathcal{A}_F} \sum_{h \in \mathcal{S}_a} (h - \mu^{(a)})(h - \mu^{(a)})^\top,$$

$$\Sigma_B(F) = \frac{1}{|\mathcal{A}_F|} \sum_{a \in \mathcal{A}_F} (\mu^{(a)} - \mu_G^{(F)})(\mu^{(a)} - \mu_G^{(F)})^\top.$$

We quantify the collapse by the trace ratio $\mathcal{NC}1_{sem}(F) = \mathrm{Tr}(\Sigma_W)/\mathrm{Tr}(\Sigma_B)$. A lower score indicates that sample embeddings with shared attributes have clustered tightly around a common centroid.

The semantic alignment ratio ($\mathcal{NC}1_{\mathbf{sem}}$). We normalize our observation against the expected terminal value of the ETF baseline:

$$\mathcal{R}_{\mathrm{sem}}(F) = \frac{\mathcal{NC}1_{\mathrm{sem}}(F)}{\mathcal{NC}1_{\mathrm{ETF}}(F)} \tag{1}$$

This ratio serves as our primary order parameter. $\mathcal{R}_{\mathrm{sem}} < 1.0$ identifies the Semantic Regime, where geometry is more structured than the UFM baseline. $\mathcal{R}_{\mathrm{sem}} \approx 1.0$ indicates the Collapsed Regime, where geometry is dictated by label class. We also validate these regimes qualitatively by inspecting the Gram matrix $G = H^\top H$ with rows/columns sorted by attribute.

We present the corresponding results in Fig. 7. They show the same qualitative pattern as our RSA analyses: semantic structure peaks early in training, while the label-driven collapsed geometry strengthens later. Its limitation is that it becomes confounded once the embedding dimension is constrained. Small d naturally compresses all embeddings into a low-dimensional subspace, which can reduce within-cluster scatter even when the geometry

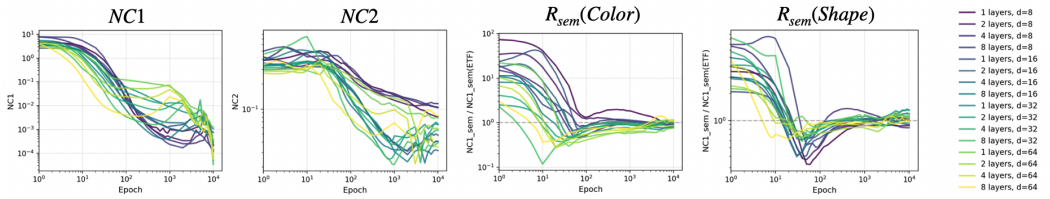


Figure 7: Recreation of Fig. 1, right panel, using our numerical measures of neural collapse and semantic representation. We see very similar learning dynamics to the patterns we report in the main text, with ETF-style collapse increasing monotonically, while semantic representation peaks early in training.

is not specifically organized by the semantic attributes of interest. In other words, $\mathcal{NC}1_{\text{sem}}$ measures whether semantic groups are tight, but not whether this tightness reflects the correct semantic structure rather than generic low-dimensional compression. This makes the metric less reliable for comparing models across different embedding dimensions. For this reason, we do not use it as our main measure, and instead rely on Hypothesis RSA, which better distinguishes true semantic organization from trivial low-rank compression.

A.4 Different-label semantic structure

One possible concern with the Sem-RSM correlation is that it can partly benefit from same-label structure. In our synthetic languages, two examples with the same output label also share all latent factors, so they are maximally similar under the Sem-RSM. Therefore, a representation that mainly clusters examples by exact output label can still have nonzero correlation with the semantic hypothesis. To check that our early semantic signal is not only coming from this effect, we measure semantic structure using only pairs of examples with different output labels.

Let $G^{(t)}_{\text{emp}}$ be the empirical RSM at checkpoint t, G_{sem} be the semantic RSM, and y_i be the output label of example i. We define the set of different-label pairs as

$$\mathcal{D} = \{(i,j) : i < j,\ y_i \neq y_j\}.$$

We also split this set into pairs that share at least one latent factor and pairs that share none:

$$\mathcal{D}_+ = \{(i,j) \in \mathcal{D} : G_{\text{sem},ij} > 0\}, \qquad \mathcal{D}_0 = \{(i,j) \in \mathcal{D} : G_{\text{sem},ij} = 0\}.$$

First, we compute a different-label semantic gap:

$$\Delta_{\text{diff}}(t) = \frac{1}{|\mathcal{D}_+|} \sum_{(i,j)\in\mathcal{D}_+} G^{(t)}_{\text{emp},ij} - \frac{1}{|\mathcal{D}_0|} \sum_{(i,j)\in\mathcal{D}_0} G^{(t)}_{\text{emp},ij}.$$

This asks whether two examples with different labels are still closer when they share some latent semantic factor than when they share none.

Second, we compute semantic RSA after removing same-label pairs. Here, $\text{corr}(\cdot,\cdot)$ denotes Pearson correlation, matching the RSA metric used in the main text, and the computation is restricted to the different-label pair set D:

$$\rho_{\text{diff}}(t) = \text{corr}\left(\left\{G^{(t)}_{\text{emp},ij}\right\}_{(i,j)\in D}, \left\{G_{\text{sem},ij}\right\}_{(i,j)\in D}\right).$$

This is the same semantic-RSM comparison as in the main text, but restricted to different-label pairs.

Figure 8 shows that both different-label diagnostics follow the same qualitative pattern as the main semantic RSA curves. They rise early in training, and then decrease later as the representations become more organized by exact output label. This supports the interpretation that the early semantic regime reflects graded structure across labels, not only collapse within each label.

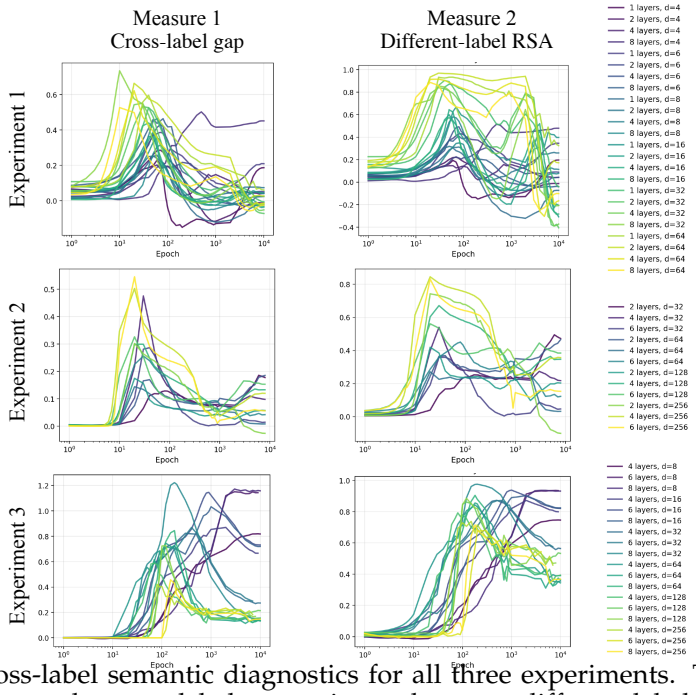


Figure 8: Cross-label semantic diagnostics for all three experiments. The left column (measure 1) shows the cross-label semantic gap between different-label pairs that share at least one latent factor and those that share none. The right column (measure 2) shows semantic RSA computed only over different-label pairs. Lighter colors correspond to larger models. Across experiments, both measures rise early in training and weaken later, especially in larger models, showing that the early semantic alignment is not explained by same-label collapse alone.

A.5 SELI geometry for imbalanced label classes

As mentioned in the main text, in the imbalanced setting, the balanced ETF geometry is no longer the ETF geometry Fang et al. (2021). Instead, we use its imbalanced generalization, the SELI geometry (Thrampoulidis et al., 2022).

For ease of reference we briefly explain here how to obtain this geometry: Let $Y_c = (I_K - \frac{1}{K} 1_K 1_K^\top) Y$ be the centered one-hot encoding label matrix (a simplex-encoded label, or SEL, matrix), and let $Y_c = U \Sigma V^\top$ be its SVD. The SELI-RSM reference geometry is then described by the Gram matrix $G_{\text{SELI}} = V \Sigma V^\top$, arising from the logits interpolating the SEL matrix.

This geometry has been empirically verified and theoretically justified in STEP-imbalanced image classification Thrampoulidis et al. (2022); Behnia et al. (2023); here we adopt it as the best-known proxy in the literature for more general imbalanced settings. For consistency, and to emphasize the label-driven nature of NC geometries, we retain the name ETF-RSM throughout the main body, but in Experiment 2 this reference should be understood as the SELI generalization described above.

A.6 Additional results for Experiment 2

Figure 9 shows the full Experiment 2 trajectories. The clearest comparison is across model size: wider models reach stronger early alignment with Sem-RSM and also move further toward the label-driven geometry later in training, while depth has a smaller effect.

The full figure also clarifies the late-stage geometry. The empirical RSM becomes more label-driven over training, but it does not become diagonal-only. Instead, some off-diagonal structure remains, which is consistent with the Zipfian label imbalance in this experiment. Since the terminal label-driven geometry is the imbalanced SELI geometry rather than

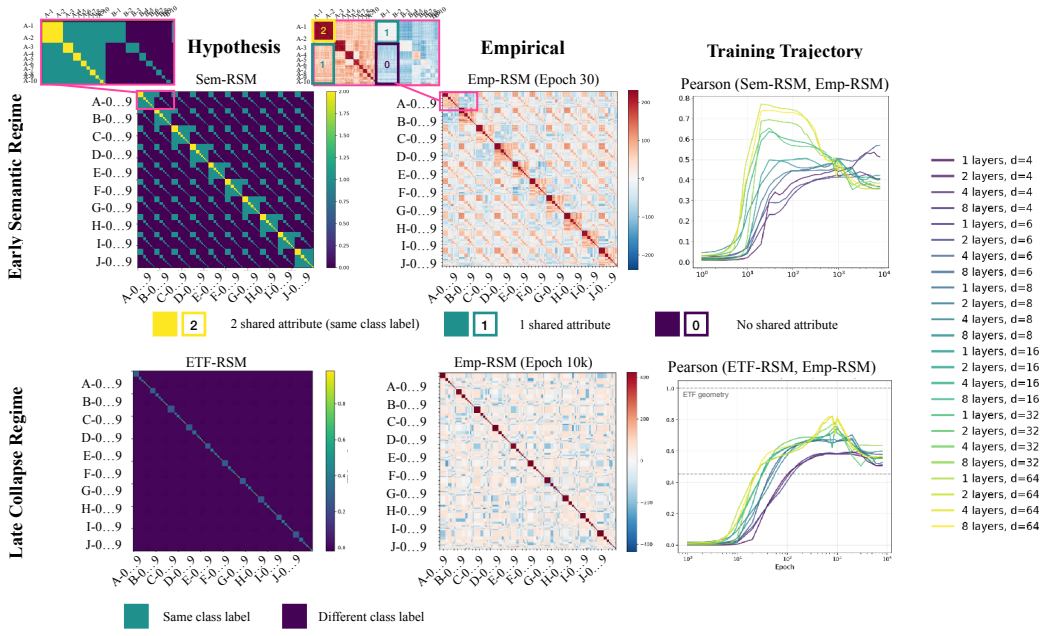


Figure 9: **Full representational similarity results for Experiment 2.** This figure is the Experiment 2 analogue of Fig. 1. **Left column:** hypothesis RSMs. Sem-RSM encodes the semantic hypothesis, where two contexts have similarity 2 if they share both latent attributes, 1 if they share one, and 0 otherwise. ETF-RSM encodes the label-driven reference geometry. In this imbalanced setting, this reference should be understood as the SELI generalization of ETF described in the main text. **Middle column:** empirical RSMs at an early epoch 30 and a late epoch 10k. Early in training, the empirical RSM closely matches the graded block structure of Sem-RSM. Later in training, the matrix becomes more label-driven, but some off-diagonal structure remains. **Right column:** Pearson correlation between Emp-RSM and each hypothesis RSM over training, across model depths and widths. As in Experiment 1, alignment with Sem-RSM rises early and then declines, while alignment with ETF-RSM increases later. Wider models show both stronger early semantic alignment and a stronger later shift toward the label-driven geometry.

the balanced ETF pattern, this residual structure should be understood as part of the late imbalance-driven regime.

A.7 Additional results for Experiment 3

Fig. 10 shows the full Experiment 3 results in the same format as Fig. 1. All models first enter the semantic regime: early in training, the empirical RSM matches Sem-RSM much more closely than ETF-RSM. Later, the empirical RSM becomes more diagonal, Sem-RSM correlation drops, and ETF-RSM correlation increases. But compared with Experiment 1, this later shift is weaker, and semantic blocks remain visible even at epoch 10k.

The clearest comparison in this figure is across widths. Smaller-width models keep higher Sem-RSM correlation and lower ETF-RSM correlation through late training. Models with widths $d = 64$ and $d = 128$ show a more clearly transient semantic phase: they peak higher on Sem-RSM early, then decline more, while also moving further toward ETF-RSM. Depth has a smaller effect than width. Overall, in Experiment 3, width mainly controls how far training moves away from the early semantic geometry.

A.8 Layerwise geometry

The main experiments measure representation geometry at the final layer, since this is the representation directly used by the readout. We additionally ask whether the same semantic-to-label transition appears uniformly across layers, or whether different layers preserve different parts of the geometry.

For each checkpoint t and layer ℓ, we construct an empirical RSM from the layer-ℓ representations, denoted $G^{(t)}_{\mathrm{emp},\ell}$. We then track two diagnostics for each layer: the different-label semantic gap $\Delta^{(t)}_{\mathrm{diff},\ell}$ defined in Appendix A.4 and the Pearson correlation between $G^{(t)}_{\mathrm{emp},\ell}$ and the ETF-RSM.

Figure 11–Fig. 13 show that the transition from semantic structure to label-driven structure is not identical at every layer. Across experiments, the ETF-RSM correlation is generally strongest in the later layers, which is expected because these layers are closest to the output prediction. The semantic gap is more mixed. In Experiment 1, it can remain visible in earlier and intermediate layers even as the final layer becomes more label-driven. In Experiments 2 and 3, upper layers often carry both a strong semantic gap and a strong label-driven signal. Thus, the layerwise picture is not simply that early layers are semantic and late layers are label-driven. Rather, the upper layers appear to first organize examples by task-relevant latent structure, and then become increasingly aligned with exact output-label geometry later in training.

This layerwise analysis does not change the main conclusion, but it makes the transition more precise: the final representation is where label-driven collapse is most visible, while intermediate representations can retain more mixed structure. Future work could study this structure in greater detail.

B Theoretical Analysis and Proofs

Notation. We introduce necessary notation that we use throughout this section. For positive integer N, we write $[N] := \{1, 2, \ldots, N\}$. I_N denotes the identity matrix of size N. 1_N is the vector of all-ones of size N. $\otimes$ denotes Kronecker product. $\mathsf{S}(\cdot) : \mathbb{R}^N \to \Delta^{N-1}$ denotes the softmax map and $\mathsf{S}_c(v) = \exp(v_c)/\sum_{c' \in [n]} \exp(v_{c'})$ is its c-th entry. When $\mathsf{S}(\cdot)$ is applied on a matrix, it acts columnwise. For vector v, we denote its Euclidean (L2) norm as $\|v\|_2$. We let $\mathsf{N}(v) = v/\|v\|_2$ denote spherical normalization. When $\mathsf{N}(\cdot)$ is applied to a matrix, it acts columnwise. For matrix V, we denote its operator norm as $\|V\|_{\mathrm{op}}$ and its Euclidean/Frobenius (L2) norm as $\|V\|_F$. We let e_i denote the standard i-th basis vector.

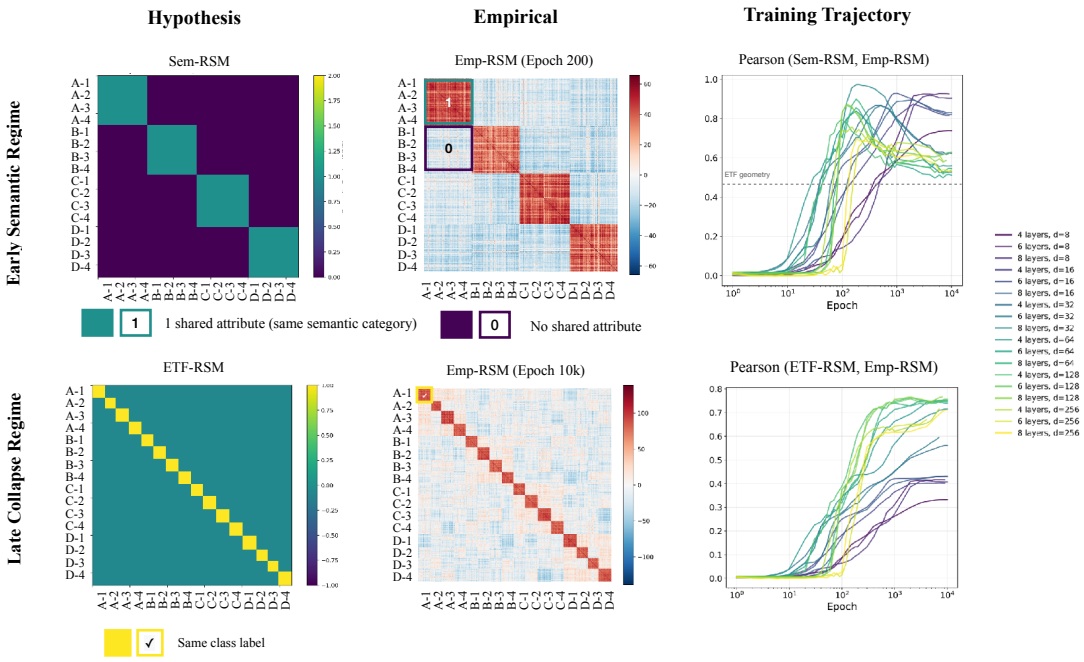


Figure 10: **Full representational similarity results for Experiment 3.** This figure is the Experiment 3 analogue of Fig. 1. **Left column:** hypothesis RSMs. Sem-RSM encodes the semantic hypothesis, where two contexts have similarity 1 if they are associated with the same semantic category and 0 otherwise. ETF-RSM encodes the label-based collapsed geometry, with ✓ marking same-label pairs. **Middle column:** empirical representational similarity matrices (Emp-RSM) at an early epoch 200 and a late epoch 10k. At epoch 200, the empirical RSM shows clear semantic block structure that closely matches Sem-RSM. At epoch 10k, this block structure is weaker and the matrix is more diagonal, although the semantic blocks are still faintly visible. **Right column:** Pearson correlation between Emp-RSM and each hypothesis RSM over training, across model depths and widths. As in the earlier experiments, correlation with Sem-RSM rises early and correlation with ETF-RSM increases later in training. But the dynamics differ from Experiment 1. Models with intermediate and larger widths, especially $d = 64$ and $d = 128$, reach stronger early alignment with the semantic hypothesis and also align more with ETF-RSM later in training. Smaller-width models retain relatively high semantic alignment until the end of training, indicating a less complete shift away from the semantic geometry.

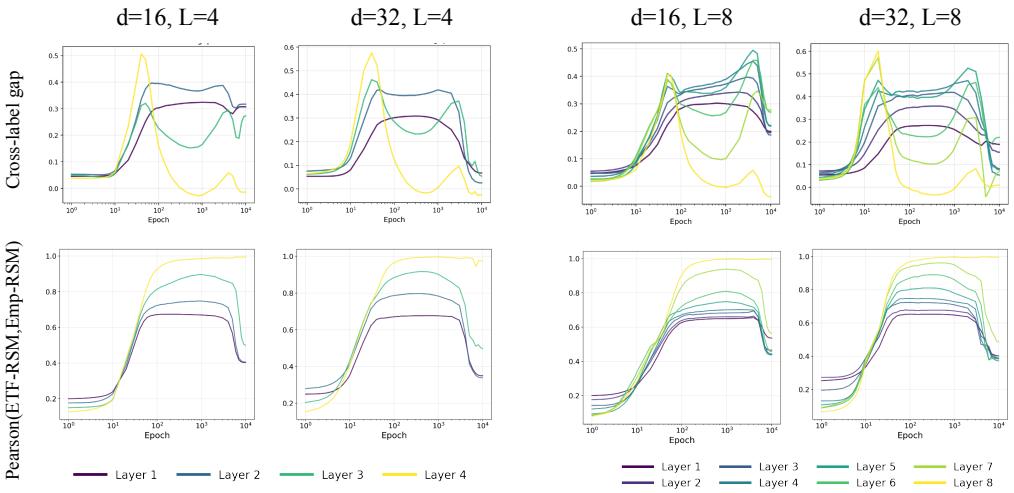


Figure 11: Layerwise RSA dynamics for Experiment 1. For four representative models, we track the cross-label semantic gap (top row), which compares the similarity of different-label pairs that share at least one latent factor to those that share none, and the Pearson correlation between each layer's empirical RSM and the ETF-RSM (bottom row) over training. Colors indicate layer depth, with darker curves corresponding to earlier layers and lighter curves to later layers. ETF-like output-label geometry is concentrated near the readout, while semantic structure can persist in earlier and intermediate layers.

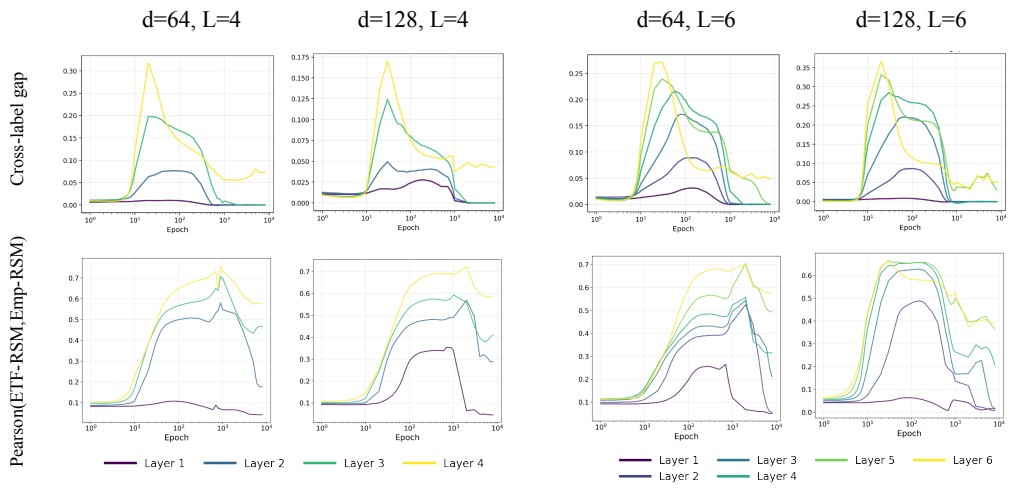


Figure 12: Layerwise RSA dynamics for Experiment 2.

B.1 Details on our Spherical BoW

Let K be the number of classes and m be the number of samples per class. The total number of samples is $n = K \cdot m$. Let s be the number of semantic features. Denote $V = n + s$ the total number of features. Let $B \in \mathbb{R}^{s \times K}$ be the base matrix of shared semantic features. For example, for $s = 6$ semantic features corresponding, in order, to red, blue, circle, square, triangle, star, and $K = 8$ classes corresponding in order to red-circle, red-square, and so on,

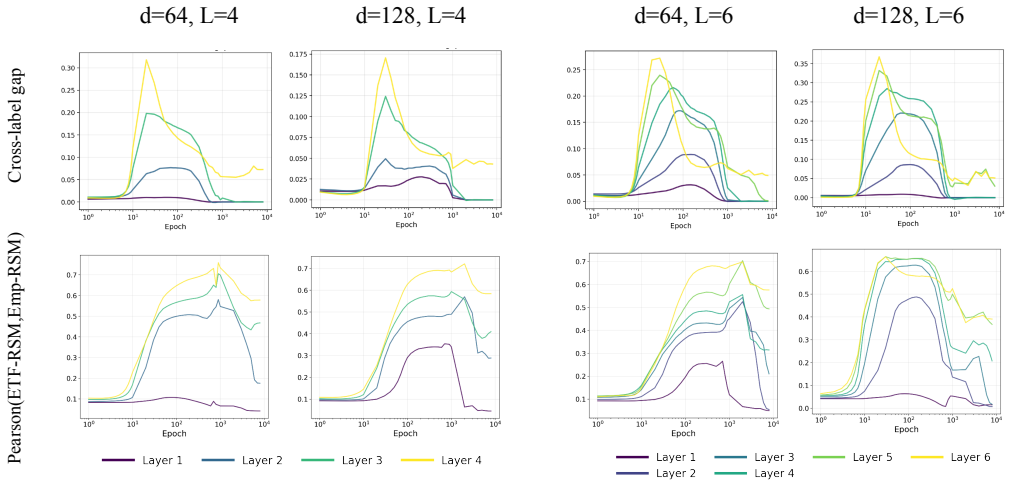


Figure 13: Layerwise RSA dynamics for Experiment 3.

the semantic base matrix $B \in \mathbb{R}^{6\times 8}$ reads

$$B = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

With this, the data matrix $X \in \mathbb{R}^{V\times n}$ is constructed by combining unique one-hot features for every sample with m repeated copies of the shared class features, i.e.,

$$X = \begin{bmatrix} I_n \\ B \otimes 1_m^\top \end{bmatrix} \in \mathbb{R}^{V\times n},$$

where I_n is the $n \times n$ identity matrix and 1_m is the all-ones vector of length m. Let $x_i \in \mathbb{R}^V$ denote the i-th column of X.

Let $y_i \in [K]$ be the true class of sample i. Specifically, let the one-hot encoding matrix $Y = I_K \otimes 1_m^\top$. We optimize a feature weight matrix $H \in \mathbb{R}^{d\times V}$ and let a classifier weight matrix $W \in \mathbb{R}^{K\times d}$. Unless otherwise stated, we set $d = K$.

We consider the following training objective

$$\bar{\mathcal{L}}(H, W) = \mathrm{KL}\left(Y||\mathrm{Softmax}(W \cdot \mathsf{N}(HX))\right) = -\sum_{i=1}^{n} \log\left(\mathbb{S}_{y_i}(W\frac{Hx_i}{\|Hx_i\|_2})\right). \tag{3}$$

This is essentially a continuous Bag-of-words (BoW) model with an additional projection $\mathsf{N}(Hx_i) = Hx_i/\|Hx_i\|_2$ of the learned features Hx_i onto the unit hypersphere before classification. The normalization can be interpreted as modeling the layer-norm at the last layer of the architecture. Also, consistent with the CBoW setup (Mikolov et al., 2013a), note that optimization over the parameters W and H imposes no further architectural constraints. We call the model in Eq. (3) the Spherical BoW model.

Remark 2 (Relation to UFM). *Our model is inspired by, but differs from the unconstrained features model (UFM) (Mixon et al., 2020; Fang et al., 2021) which is the go-to model in the NC literature. In the UFM, the penultimate layer features of a deep-net are treated as free optimization*

variables rather than the output of a specific architecture. In the language setting that we study here, this means that embedding of each context is a free optimization variable (Zhao et al., 2024). To explicitly relate it to Eq. (3), the UFM corresponds to minimizing jointly over W and H the loss $\mathcal{L}(H, W) = \mathrm{KL}\left(Y||\mathsf{S}\left(WH\right)\right) = -\sum_{i=1}^{n} \log(\mathsf{S}_{y_i}(WHe_i))$. Because the UFM treats embeddings as free optimization variables, we see that it abstracts away the input data. This abstraction precludes any analysis of how specific structure of the input data influences the resulting representations. Instead, our BoW-style model explicitly accounts for such structures by replacing the basis input vector e_i above with a one-hot encoding x_i over multiple features *present in the i-th example.*

Remark 3 (Normalization)**.** *The spherical normalization is essential in Eq. (3) to guarantee that in the TPT the geometry converges to an ETF. In the absence of it, we can show that the geometry retains semantic structure even at convergence. Thus, the vanilla CBoW model does not exhibit the transient semantic geometry seen in our transformer experiments.*

B.2 Analysis

To aid the analysis, we henceforth fix the classifier W to an ETF, i.e. $W = I_K - \frac{1}{K} 1_K 1_K^\top$ and train H. We numerically verify in Figure 14 that fixing W maintains the transient semantic-geometry behavior. The theorems below provide theoretical justification for this. Note that despite fixing the classifier, the training objective remains non-convex due to the spherical normalization, and so the gradient-descent analysis remains challenging. We therefore resort to certain simplifying conditions under which our formal results hold. Relaxing these conditions is left to future work; we note that gradient-descent dynamics and implicit bias in linear models are substantially less understood in the presence of normalization (e.g., Hoffer et al. (2018); Cao et al. (2023); Chou et al. (2024)) than without it Soudry et al. (2018).

Let $Z = HX \in \mathbb{R}^{K \times n}$ be the unnormalized feature matrix, where $z_i \in \mathbb{R}^K$ is its i-th column. The normalized features are denoted as $u_i = \mathsf{N}(z_i) = \frac{z_i}{\|z_i\|_2} \in \mathbb{S}^{K-1}$. Also, let $w_c, c \in [K]$ denote the c-th row of W.

B.2.1 Early Phase

Theorem 1 (Semantic bias in the first gradient step)**.** *Consider one gradient descent step on the Spherical Bag-of-Words objective (3) with fixed ETF classifier $W = I_K - \frac{1}{K} 1_K 1_K^\top$. Let $Z^{(1)} = H^{(1)} X \in \mathbb{R}^{K \times n}$ be the feature matrix after this step with learning rate η. Assume at initialization there exists a small $\varepsilon > 0$ and a scale $\alpha > 0$ such that for all $i \in [n]$:*

1. *the initial logits are approximately balanced across samples, so that the corresponding softmax vectors satisfy: $\left\| \mathsf{S}\left(W u_i^{(0)}\right) - \frac{1}{K} 1_K \right\|_2 \leq \varepsilon$*
2. *the initial feature norms are approximately constant: $\left| \frac{1}{\|z_i^{(0)}\|_2} - \alpha \right| \leq \varepsilon$;*
3. *the initial normalized features are weakly aligned with their true classifiers: $\left| u_i^{(0)\top} w_{y_i} \right| \leq \varepsilon$.*

Then, the change $Z^{(1)} - Z^{(0)}$ in the feature matrix after one gradient step is given by

$$Z^{(1)} - Z^{(0)} = \alpha\eta W(I_K + mB^\top B) \otimes 1_m^\top + \eta R\,,$$

where the residual is bounded by $\|R\|_{\mathrm{op}} \leq \mathcal{O}(\varepsilon\sqrt{K}) \cdot \sqrt{m}(1 + m\|B\|_{\mathrm{op}}^2)$. Thus, if $m > 1$ and $\varepsilon \ll \frac{1}{\sqrt{K}}$, the relative error vanishes and the gradient update exhibits an explicit semantic bias through the term $B^\top B$.

Proof. Let $\ell_i(z_i) = -\log(\mathsf{S}_{y_i}(W\mathsf{N}(z_i))) = -\log(\mathsf{S}_{y_i}(Wu_i))$ denote the contribution to the loss of the i-th example. For each $i \in [n]$,

$$\nabla_{z_i} \ell_i(z_i) = \frac{1}{\|z_i^{(0)}\|_2} \left(I_K - u_i^{(0)} u_i^{(0)\top}\right) W(\mathsf{S}\left(W u_i^{(0)}\right) - e_{y_i}).$$

By Assumption (1) and using $W1_K = 0$, this becomes:

$$\nabla_{z_i} \ell_i(z_i) = \frac{1}{\|z_i^{(0)}\|_2} \left(I_K - u_i^{(0)} u_i^{(0)\top} \right) (-w_{y_i} + W r_i) .$$

By Assumption (2), $\alpha_i := \frac{1}{\|z_i^{(0)}\|_2} = \alpha + \epsilon_{\alpha,i}$ with $|\epsilon_{\alpha,i}| \le \varepsilon$. Thus,

$$\nabla_{z_i} \ell_i(z_i) = -\alpha w_{y_i} + \xi_i ,$$

where

$$\xi_i = -\epsilon_{\alpha,i} w_{y_i} + \alpha_i (u_i^{(0)\top} w_{y_i}) u_i^{(0)} + \alpha_i (I_K - u_i^{(0)} u_i^{(0)\top}) W r_i .$$

Noting that $\|w_{y_i}\|_2 < 1$, $\|u_i^{(0)}\|_2 = 1$, and both W and the projection matrix have operator norms of 1, applying the triangle inequality alongside Assumptions (1)–(3) yields

$$\|\xi_i\|_2 \le \varepsilon + (\alpha + \varepsilon)\varepsilon + (\alpha + \varepsilon)\varepsilon = \varepsilon(1 + 2\alpha + 2\varepsilon) := \varepsilon' .$$

Therefore, recalling $Y = I_K \otimes 1_m^\top$, the full gradient with respect to Z is:

$$\nabla_Z \bar{\mathcal{L}}(Z^{(0)}) = -\alpha W Y + \Xi, \qquad \|\Xi\|_{\text{op}} \le \sqrt{n} \max_{i \in [n]} \|\xi_i\|_2 \le \sqrt{n}\varepsilon' .$$

By the chain rule, $\nabla_H \bar{\mathcal{L}}(H^{(0)}) = \nabla_Z \bar{\mathcal{L}}(Z^{(0)}) X^\top$. Thus, the updated feature matrix is:

$$Z^{(1)} = \left(H^{(0)} - \eta \nabla_H \bar{\mathcal{L}}(H^{(0)}) \right) X = Z^{(0)} - \eta \nabla_Z \bar{\mathcal{L}}(Z^{(0)}) X^\top X \tag{4}$$

$$= Z^{(0)} + \alpha\eta W Y X^\top X - \eta \Xi X^\top X . \tag{5}$$

Using $X^\top X = I_n + (B^\top B) \otimes (1_m 1_m^\top)$ and some algebra steps yields

$$Y X^\top X = (I_K + m B^\top B) \otimes 1_m^\top .$$

Finally, let the residual matrix be $R = -\Xi X^\top X$. Because the non-zero eigenvalue of $1_m 1_m^\top$ is m, the maximum eigenvalue of $X^\top X$ is $1 + m\|B\|_{\text{op}}^2$. Thus, using $n = Km$:

$$\|R\|_{\text{op}} \le \|\Xi\|_{\text{op}} \|X^\top X\|_{\text{op}} \le \sqrt{K}\sqrt{m}\varepsilon' \left(1 + m\|B\|_{\text{op}}^2 \right) .$$

This completes the proof. □

B.2.2 Terminal Phase

Theorem 2. *Consider training with gradient descent the Spherical Bag-of-Words model (3) with a fixed ETF classifier $W = I_K - \frac{1}{K} 1_K 1_K^\top$. Assume that gradient descent converges to a global minimizer H_∞, and let $Z_\infty = H_\infty X$. Let $U_\infty = N(Z_\infty)$ be the corresponding matrix of normalized features. Then, the normalized Gram matrix satisfies:*

$$U_\infty^\top U_\infty = \frac{K}{K-1} \left(I_K - \frac{1}{K} 1_K 1_K^\top \right) \otimes 1_m 1_m^\top = \frac{K}{K-1} G_{ETF}$$

Proof. The proof follows from two main steps.

First, by Lemma 1, optimizing over H is equivalent to optimizing over the free feature matrix $Z = HX \in \mathbb{R}^{K \times n}$. Because H_∞ is assumed to be a global minimizer of $\bar{\mathcal{L}}$, the corresponding unnormalized feature matrix Z_∞ must globally minimize the objective over all possible matrices in $\mathbb{R}^{K \times n}$.

Note that the objective decouples over the samples as $\bar{\mathcal{L}}(Z) = \sum_{i=1}^n \ell_i(u_i)$. Because Z_∞ is a free matrix, its columns z_i can be varied independently. Suppose, for the sake of contradiction, that for some sample i, the normalized feature $u_{i,\infty}$ is not a global minimizer of $\ell_i(u_i)$ over the sphere $\mathbb{S}^{K-1}$. Then there exists some $u_i^* \in \mathbb{S}^{K-1}$ such that $\ell_i(u_i^*) < \ell_i(u_{i,\infty})$. By constructing a new matrix $\tilde{Z}$ identical to Z_∞ but with the i-th column replaced by u_i^*,

we would achieve a strictly lower total loss $\bar{\mathcal{L}}(\tilde{Z}) < \bar{\mathcal{L}}(Z_\infty)$, which contradicts the global optimality of Z_∞.

Consequently, for every sample i, $u_{i,\infty}$ must strictly be a global minimizer of the per-sample spherical cross-entropy objective. By Lemma 2, this minimizer is unique and satisfies:

$$u_{i,\infty} = \frac{w_{y_i}}{\|w_{y_i}\|_2} \qquad \text{for all } i \in [n].$$

Collecting the normalized columns by class, and using $\|w_c\|_2 = \sqrt{\frac{K-1}{K}}$, gives the normalized feature matrix:

$$U_\infty = \sqrt{\frac{K}{K-1}} W (I_K \otimes 1_m^\top).$$

Further using $W^\top W = W = I_K - \frac{1}{K} 1_K 1_K^\top$, we find the advertised expression for the normalized Gram matrix:

$$U_\infty^\top U_\infty = \frac{K}{K-1} (I_K \otimes 1_m) \left(I_K - \frac{1}{K} 1_K 1_K^\top \right) (I_K \otimes 1_m^\top) = \frac{K}{K-1} \left(I_K - \frac{1}{K} 1_K 1_K^\top \right) \otimes 1_m 1_m^\top .$$

□

Remark 4 (Global Convergence Assumption)**.** *The assumption that gradient flow converges to a global minimizer is a common simplifying assumption in the UFM and Neural Collapse literature Fang et al. (2021); Zhu et al. (2021); Thrampoulidis et al. (2022). Since the objective $\bar{\mathcal{L}}$ is non-convex, establishing global convergence from random initialization is in general difficult. This issue remains in our setting: even with W fixed, the optimization over H is still non-convex because of the normalization. Accordingly, we adopt this assumption to separate the analysis of the terminal representation from the analysis of the optimization dynamics, and focus here on the geometry of the limiting features.*

Supporting Lemmas

Lemma 1 (Equivalence to the Unconstrained Feature Model)**.** *Optimizing the objective $\bar{\mathcal{L}}$ over $H \in \mathbb{R}^{K \times V}$ is equivalent to optimizing over the free variables $Z \in \mathbb{R}^{K \times n}$.*

Proof. It suffices to show that for arbitrary matrix $Z \in \mathbb{R}^{K \times n}$ there exists a weight matrix H such that $HX = Z$. This is possible thanks to the structure of the data matrix $X = \begin{bmatrix} I_n \\ B \otimes 1_m^\top \end{bmatrix}$. Indeed, the desired is true by setting $H = [Z \quad 0_{K \times s}]$. □

Lemma 2 (Optimal Directional Alignment)**.** *Suppose $W = I_K - \frac{1}{K} 1_K 1_K^\top$ is fixed to an ETF. For any sample i with class label $y_i \in [K]$, the minimization of the cross-entropy loss over the sphere*

$$\min_{u_i \in \mathbb{S}^{K-1}} \left\{ \ell_i(u_i) = -\log \left(\mathbb{S}_{y_i}(W u_i) \right) \right\} ,$$

has a unique global minimizer given by $u_i^* = \frac{w_{y_i}}{\|w_{y_i}\|_2}$, where $w_c, c \in [K]$ denotes the c-th row of W.

Proof. Denote $c = w_{y_i}^\top u_i$. Because $u_i \in \mathbb{S}^{K-1}$, by Cauchy-Schwarz inequality dictates that:

$$c \leq \|w_{y_i}\|_2 ,$$

with equality holding if and only if u_i is perfectly aligned with w_{y_i}, i.e., $u_i = \frac{w_{y_i}}{\|w_{y_i}\|_2}$.

Moreover, since $\sum_{c=1}^K w_c = 0$:

$$\sum_{c \neq y_i} w_c^\top u_i = -w_{y_i}^\top u_i = -c .$$

From this, and Jensen's inequality applied to the strictly convex exponential function:

$$\frac{1}{K-1}\sum_{j\neq y_i}\exp(w_j^\top u_i) \geq \exp\left(\frac{1}{K-1}\sum_{j\neq y_i} w_j^\top u_i\right) = \exp\left(\frac{-c}{K-1}\right),$$

with equality if and only if $w_c^\top u_i$ is identical for all $c \neq y_i$.

Substituting this lower bound back into the loss function yields a lower bound on $\ell(u_i)$ as a function of c:

$$\ell(u_i) \geq \log\left(1+(K-1)\exp\left(\frac{-c}{K-1}-c\right)\right) = \log\left(1+(K-1)\exp\left(\frac{-K}{K-1}c\right)\right).$$

The lower bound is strictly decreasing function of c. Therefore, to globally minimize the loss, we must strictly maximize c.

But c attains its maximum uniquely at $u_i^* = \frac{w_{y_i}}{\|w_{y_i}\|_2}$. At this point it holds that

$$w_c^\top u_i^* = \frac{w_c^\top w_{y_i}}{\|w_{y_i}\|_2} = \frac{-1/K}{\sqrt{1-1/K}} = \frac{-1}{\sqrt{K(K-1)}}.$$

Thus, the equality condition for Jensen's inequality is satisfied. We conclude that the lower bound is achieved and $u_i^* = \frac{w_{y_i}}{\|w_{y_i}\|_2}$ is the unique global minimizer. □

B.3 Direct simulation of the Spherical BoW dynamics

To complement the theoretical analysis, we run simulation with Spherical BoW model and Color-Shape data with semantic basis matrix B as in Eq. (2). We use $K = 8$ label classes, $m = 5$ samples per class, so $n = Km = 40$ total samples, and $s = 6$ shared semantic features corresponding to the two colors and four shapes. This input matrix is:

$$X = \begin{bmatrix} I_n \\ B \otimes 1_m^\top \end{bmatrix} \in \mathbb{R}^{V\times n}, \qquad V = n + s = 46,$$

where $B \in \mathbb{R}^{6\times 8}$ is the semantic base matrix. We sweep hidden dimension $d \in \{8, 16, 32, 64\}$, initialize all trainable parameters with small zero-mean Gaussians of standard deviation 0.05, and run full-batch gradient descent for 10^5 steps with learning rate 10^{-3}. We consider two settings. In the first, the classifier is fixed to an ETF readout and only H is trained (aligned with theoretical analysis of Sec. B.2). In the second, both W and H are trained jointly. At each step we compute the empirical RSM

$$G_{\text{emp}} = X^\top H^\top H X$$

and compute its Pearson correlation with semantic and ETF hypothesis RSMs. The results in Figure 14 show the same two-stage behavior as in our Transformer experiments. In both settings, G_{emp} first moves toward the semantic geometry: the early empirical RSM exhibits the same block structure as Sem-RSM, and correlation with Sem-RSM rises sharply. Later in training, this semantic alignment declines while correlation with ETF-RSM increases, and the final empirical RSM approaches the label-driven ETF structure.

Two points are notable. First, the transient semantic phase already appears when W is fixed to ETF, which is the setting analyzed in our theory. This shows that a trainable classifier is not necessary for the emergence of the semantic phase. Second, increasing d makes both phases more pronounced: larger hidden dimensions reach higher early alignment with Sem-RSM and later converge more fully toward ETF-RSM. Allowing W to be trainable mainly shifts the transition to later times, but does not change the qualitative picture.

B.4 Spherical BoW on the Color-Shape Categories language

We next test the Spherical BoW model on the full Color-Shape Categories language from Experiment 1, described in Section 4, rather than on the simplified analytic design matrix

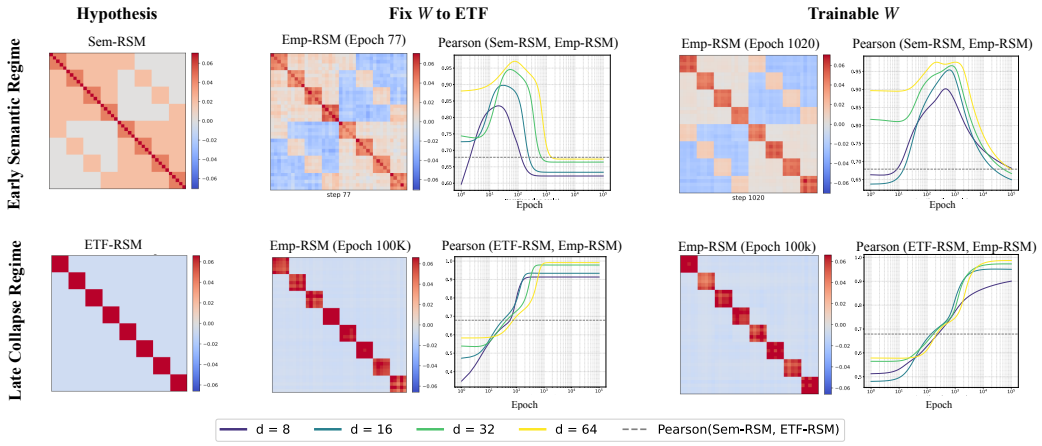


Figure 14: **Numerical simulation of the Spherical BoW dynamics on the color-shape setup. Left column:** hypothesis RSMs. Sem-RSM encodes the semantic similarity structure, while ETF-RSM encodes the label-driven collapsed geometry. **Middle and Right columns:** Simulations of the Spherical BoW objective under two settings: fixing W to an ETF readout, and training W jointly with H. In each setting, we show a representative empirical RSM in the early semantic regime and at the end of training, together with the Pearson correlation of $G_{\text{emp}} = X^\top H^\top H X$ with Sem-RSM and ETF-RSM over training, for $d \in \{8, 16, 32, 64\}$. In each setting, the early empirical RSM is taken from the epoch at which the correlation with Sem-RSM is maximal, and the late empirical RSM is taken from the final training epoch. In both settings, the empirical geometry first aligns with the semantic structure and later approaches the ETF geometry. Larger d makes both the early semantic peak and the late ETF convergence more pronounced. The same transient semantic phase appears even when W is fixed to ETF, showing that joint training of the classifier is not necessary for the effect.

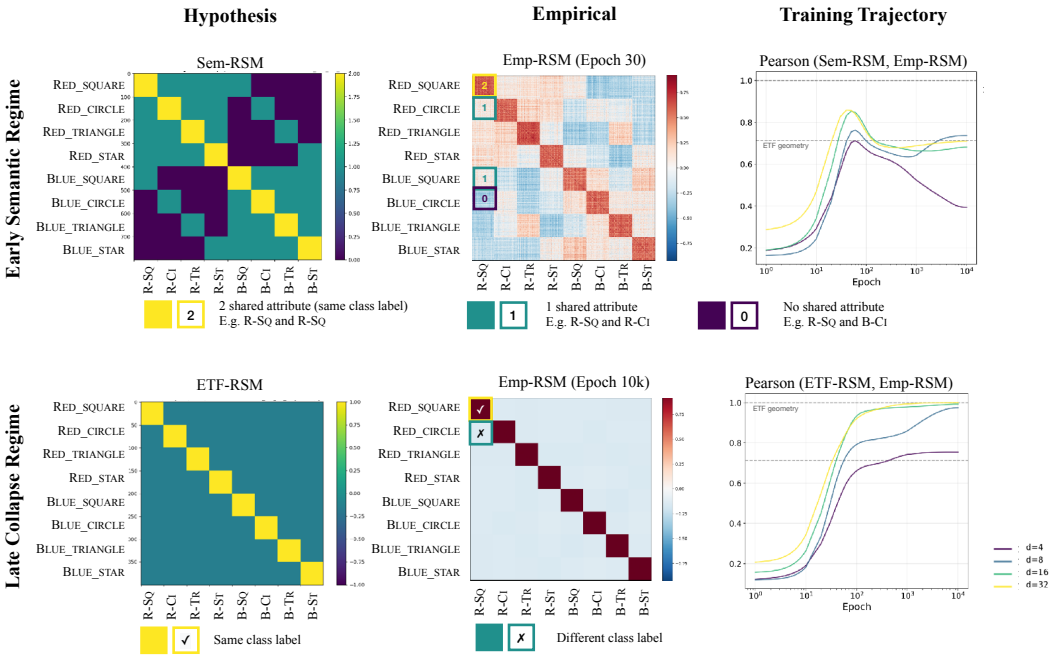


Figure 15: **Spherical BoW on the full Color-Shape Categories language.** This figure is the Spherical BoW analogue of Fig. 1. **Left column:** hypothesis RSMs from Experiment 1. Sem-RSM encodes semantic overlap between classes, with values 2, 1, and 0 for pairs sharing two, one, or no latent attributes, while ETF-RSM encodes the label-driven collapsed geometry. **Middle column:** empirical RSMs at an early epoch (30) and a late epoch (10k). Early in training, the empirical geometry shows the graded semantic block structure of Sem-RSM. By the end of training, this structure weakens and the empirical RSM becomes close to the ETF geometry. **Right column:** Pearson correlation between Emp-RSM and each hypothesis RSM over training, for $d \in \{4, 8, 16, 32\}$. As in the Transformer experiments, alignment with Sem-RSM rises early and then declines, while alignment with ETF-RSM increases later. Larger hidden dimensions show both stronger early semantic alignment and more complete late convergence to the label-driven geometry.

used in the direct simulation above. We use the same dataset as in the main experiment: $n = 800$ samples across $K = 8$ label classes, with one nuisance ID per sample and contextual tokens drawn from fine-grained color and shape variants. Thus, the training labels remain strictly one-hot, while the inputs still contain the latent semantic overlap structure of Experiment 1.

We train the Spherical BoW model on multi-hot encodings of these input contexts and track the empirical representational similarity matrix throughout training. We sweep hidden dimension $d \in \{4, 8, 16, 32\}$, and compare the resulting empirical RSM to the same two hypothesis RSMs used in the main experiment: Sem-RSM, which reflects semantic overlap between classes, and ETF-RSM, which reflects the label-driven collapsed geometry.

The results in Figure 15 match the Transformer dynamics from Experiment 1. For models with unconstrained dimension ($d > K$), early in training, the empirical RSM exhibits the block structure of Sem-RSM, and correlation with Sem-RSM rises. Later in training, this semantic structure weakens, correlation with Sem-RSM declines, and correlation with ETF-RSM increases monotonically toward the label-driven geometry. As in Experiment 1, the hidden dimension controls how strongly both phases appear. Larger d leads to stronger early semantic alignment and more complete late convergence to ETF-RSM, while the smallest model ($d = 4$) shows both a weaker semantic peak and a less complete late collapse. This again supports the view that the semantic phase is not caused by a representational bottleneck. Rather, even in this much simpler architecture, larger models recover the semantic geometry more strongly before later training shifts them toward the label-driven regime.

B.5 Limitations

We view our Spherical CBoW as a first step toward establishing analytical proxies for studying semantic-driven geometries, analogous to the role the UFM plays for label-driven NC geometries. In the NC literature, the UFM is valuable not only because it explains already-observed geometries, but even more so because it forms the basis for *predicting* geometry in new settings (e.g. imbalanced data or soft labels). An analogous framework for semantic geometry would be highly desirable, but the challenge is substantial. The UFM captures the large-model regime by treating embeddings as free optimization variables — an assumption that is intuitive for sufficiently expressive models and has been rigorized in certain settings (Jacot et al., 2024; Súkeník et al., 2025; Garrod & Keating, 2024). However, this very assumption erases input semantic structure by design. At the same time, we have seen that architectural bottlenecking alone cannot explain the strong semantic geometry that emerges in larger models, which fully reflects the input data structure. Our Spherical BoW preserves this structure and reveals the transient semantic phase, but it lacks the intuitive expressiveness of the UFM. This is perhaps to be expected: for instance, its bag-of-words assumption yields an order-independent model that cannot distinguish two input contexts differing only in the positional arrangement of their semantic tokens. Extending the Spherical BoW framework to address these limitations is an exciting direction for future work, and we view both this model and our experiments as a starting point.