Adversarial Concept Search: Predicting Compositional Errors From Feature Geometry

Jennifer Meng Lu
Brown University
meng_lu@brown.edu

Ruochen Zhang
Brown University
ruochen_zhang@brown.edu

Isabelle Lee
University of Southern California
lee.isabelle.g@gmail.com

David Alvarez-Melis
Harvard University
dam@seas.harvard.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

Naomi Saphra
Boston University
nsaphra@bu.edu

Abstract

Humans cannot always intuit what scenarios are most challenging to LLMs. Hoping to capture challenging edge cases, developers either design problems to be difficult for humans or curate extensive benchmarks. What if we could instead anticipate which scenarios a model will fail on? In this paper, we use an LLM's representational geometry to predict which concept combinations it will fail on. We attribute this compositional failure to interference between salient features. In tasks that require systematic composition—toy programmatic settings, multihop reasoning, multilingual factual recall—we find that when a pair of concepts is encoded near-orthogonally, the model reliably composes them. When their linear encodings are close, producing interference, the model fails to compose them. Our method reliably anticipates failure modes across different compositional tasks, *without* evaluating specific inputs. These results lay the groundwork to use representational geometry to identify high-risk examples, construct targeted stress tests, and provide a scalable foundation for active learning in real-world deployment.

1 Introduction

As large language models (LLMs) improve across a wide range of tasks and domains, it is increasingly difficult to identify remaining challenges. Since humans cannot reliably predict which concept combinations will be challenging for LLMs, this gap in understanding makes dataset curation inefficient and limits our ability to anticipate LLM failure modes. Developers either design problems to challenge humans—but not LLMs—-or they curate broad benchmarks, hoping to cover informative edge cases. What if we could instead anticipate which scenarios an individual model will fail on?

We call this objective *Adversarial Concept Search* (ACS): the task of identifying meaningful conceptual scenarios that are likely to induce model failure without evaluating the model on specific inputs that instantiate those scenarios. In this paper, we efficiently identify adversarial scenarios by predicting failures of *compositional generalization*, specifically systematicity [1, 2]. A system generalizes compositionally when it successfully recombines known atomic concepts in novel configurations that were not observed during training. For example, if the training distribution includes "black dog" and "white cat", a compositional model will reliably process "black cat". This capacity is crucial for robust performance in out-of-distribution (OOD) settings.

By mapping these compositional capabilities, we could systematically construct custom challenge sets for efficient, targeted stress testing. As a practical demonstration, we successfully leverage a model's internal representational geometry to predict its failures in untested scenarios. Unlike existing

Preprint.

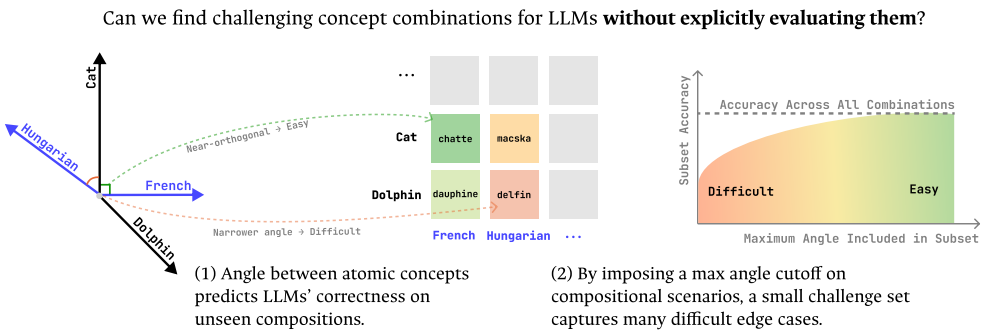


Figure 1: The angle between atomic concept representations identifies the most challenging compositions, enabling failure prediction across the combinatorial space without evaluating specific inputs.

error prediction methods such as used for active learning [3], we do not require a specific input processed by the model. When curating datasets for semi-supervised settings like language modeling, such methods are of limited practical use. In these settings, we lack diverse existing inputs and are constrained by the expense of generating and processing every possible input. Our approach instead predicts errors from only a *description* of the concepts involved and their atomic representations. This tool allows a developer to prioritize generating or collecting coherent, challenging inputs. For example, it may be prohibitively expensive to professionally translate all English corpora into Russian for a multilingual LLM, but by identifying which concepts will produce Russian errors, we can prioritize collecting Russian data in problem areas like *notable deaths in Russian*. Across multiple tasks, we will predict these compositional errors from the geometry of their atomic concepts.

These predictions rely on our hypothesis that models make mistakes when features stored in *superposition* cause interference. LLMs rely on superposition [4, 5] to encode many features in limited dimensions by sharing linear directions. It has been shown that, even without orthogonal feature encodings, this compression can in principle be perfectly *lossless*—that is, invertible—as long as feature activations are sparse [6]. In practice, however, we posit that models learn a *lossy* coding of their input features: when multiple non-orthogonal feature encodings are simultaneously active, they can interfere with one another and damage model performance.

This interference is especially salient in compositional settings, where multiple task-relevant features must be jointly represented. We hypothesize that LLM failures can arise from geometric interference between atomic concept representations, and that these failures can be predicted from the angles between their encoded feature directions. Our conceptual model can use representational geometry to proactively identify failure cases, paving the way for dynamic stress testing and scalable active learning. Our contributions are as follows:

1. **Description and analysis of compositional interference in lossy superposition.** We attribute compositional errors to interference between non-orthogonal atomic concept representations. Because only a few features activate at a time, prior theory argues that an ideal decoder can recover the active features even when they are encoded in superposition. We posit that when decoding is lossy, recovery error is governed by interference between the features being composed.
2. **Proof-of-concept in a controlled setting (SCAN).** We validate this hypothesis in SCAN [7], a synthetic compositional generalization benchmark. In this controlled setting, we measure pairwise interactions between concept representations and show that smaller angles correlate with compositional errors. This result holds across various data conditions and model sizes, confirming the predicted relationship between interference and failure.
3. **Predicting compositional failure in real LLMs.** We successfully predict successful compositional generalization in realistic LLM tasks from geometric interference between the constituent concepts. In multihop question answering (QA), the angle between component single-hop representations predicts whether the LLM will successfully compose them. In multilingual factual recall, the angle between fact representations and language subspaces predicts retrieval accuracy. Across both settings, we find that greater separation between

atomic representations corresponds to more reliable composition, demonstrating that representation geometry can predict compositional failures without evaluating the composed task itself. This establishes a scalable foundation for identifying challenging input scenarios and guiding active learning in real-world deployment.

2 Compositionality and Lossy Superposition

We will first explain why LLMs may fail to compose non-orthogonal feature representations and then explain how to leverage this phenomenon for error prediction. Compositionality has long been associated with orthogonal atomic feature representations [8], but modern theories of feature superposition permit lossless reconstruction from non-orthogonal encodings, even with exponentially more input features than representation dimensions [6, 9, 10]. In practice, however, we contend that superposition is *lossy* and, therefore, that compositional errors are predictable from the geometry of the active features being combined. We then describe how features can be extracted from language models and used to empirically measure angular distance between them across different settings.

2.1 Background: Lossy Superposition

Intuitively, one can store d features in d linear dimensions. So why is lossless compression of more than d features possible under current theories of superposition? The key assumption is k-**sparsity**: only $k \ll d$ features can be simultaneously active in any given input. This realistic sparsity assumption permits sufficiency theorems from compressed sensing [6], guaranteeing that the encoded representation can be inverted, recovering the original features exactly. This sparsity allows us to recover an *exponential* number of features m from a *linear* number of dimensions d. Specifically, if we know the **linear encoder** $A \in \mathbb{R}^{d \times m}$, we can exactly recover any specific **feature vector** $z \in [-1, 1]^m$ with k-sparse **support** $\text{supp}(z) \subset [m]$ from its encoded representation Az. Given Az, we can reconstruct the feature vector $\hat{z}$ with zero **recovery error**,

$$\|z - \hat{z}\|_2 = 0. \tag{1}$$

Crucially, *this guarantee does not degrade when feature representations are non-orthogonal*. Even if we prohibit nonlinear decoding by the next LLM layer, concepts encoded with very similar representations can still be recovered using a biorthogonal decoder dictionary [9].

Although this ideal decoding exists under mild assumptions,[1] it is unlikely to hold during real LLM inference; we can assume some noise in the representation. However, *lossless recovery is only guaranteed in noiseless settings.* In compressed sensing theory, recovery is more sensitive to noise when the feature encodings have high similarity [12]. Specifically, robust decoding requires an encoder matrix A with low global coherence, defined as the maximum similarity between all columns,

$$\rho = \max_{\substack{i,j \in [m] \\ j \neq i}} |\cos(a_i, a_j)|.$$

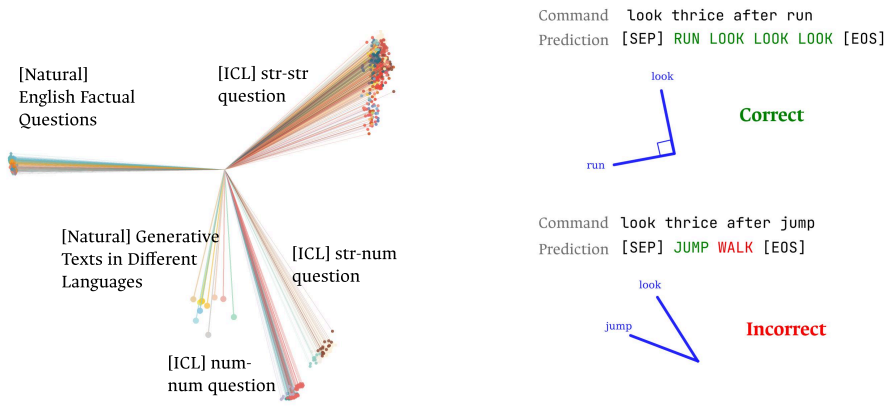
Worst-case recovery error bounds do not depend on low coherence with an ideal decoder, but do when decoding under noise. Robust decoding motivates why correlated features are known to be encoded orthogonally, while anti-correlated features exhibit negative interference [5].

Global coherence provides bounds on worst-case errors, but we are concerned about errors on specific feature combinations. In a specific scenario, not all interference is equal. We are most concerned with features *salient to that scenario*—specifically, the sparse support $\text{supp}(z)$ and any features that are relevant in the context of the sparse support.[2] Regardless of how the LLM identifies the **salient support** $\mathcal{S}$, if it restricts decoding to only salient features, then the dimension required for robust recovery is controlled with high probability by the structured bound of Adcock et al. [13]. In effect, this bound depends on each example's interaction with the salient support, which across examples of a shared scenario is bounded by the salient support's **local cumulative coherence**,

$$\alpha(\mathcal{S}) = \max_{i \in \mathcal{S}} \sum_{\substack{j \in \mathcal{S} \\ j \neq i}} |\cos(a_i, a_j)|. \tag{2}$$

[1] For classic compressed sensing, the relevant assumption is the Restricted Isometry Property [11].

[2] One simple way to limit the salient feature set is by leveraging structured sparsity and assuming features are organized into active blocks of related features. This approach is suggested by Adcock et al. [13] and inspires our concept-salient subspace construction by SVD in Section 4.



(a) Large-scale structure in representation space. (b) SCAN example.

Figure 2: **After controlling large-scale representation structure, compositional generalization can be predicted from feature interference.** (a) Representations cluster by global structures that are irrelevant to the concepts of interest. Colors indicate examples from the same topic subset (Layer 9). (b) Hypothesized illustration of compositional failure for Section 3. Two commands share the same structure but differ in a single concept (`run` vs. `jump`). The relevant atomic concepts at the top are near-orthogonal, enabling correct composition.

Further theoretical details on the relevant bound, as well as a related bound for robust *linear* compressed sensing, are provided in Appendix A.

Why might feature recovery be less robust when the salient support has high cumulative coherence? Intuitively, robust recovery is hampered by destructive interference from the most damaging feature: the salient feature that has the smallest angle with other features salient to the support. This is captured by the local cumulative coherence term, which we operationalize below as *Compositional Interference* (CI). We will leverage this interference metric to rank concept combinations by likelihood of LLM compositional failure. Our intuition is simple: when an LLM cannot robustly recover a set of features in superposition due to interference, it is more likely to make mistakes during processing.

2.2 Measuring compositional interference

The theory above predicts that compositional failures should be more likely when the features active or salient in a compositional input have high cumulative coherence. How can we compute this value *without access to an example* of the compositional scenario?

In real LLMs, the ground-truth concepts learned by the model are not directly accessible. While methods such as Sparse Autoencoders (SAEs) [14] are proposed to disentangle the representation space into discrete features, they introduce additional assumptions and implementation challenges. We instead seek a simple proxy for compositional interference using models' residual representations.

Atomic concepts and salient features. We distinguish between atomic *concepts* in the input and latent *features* in the model representation. Let $\mathcal{C}$ denote the set of atomic concepts, and let $C \subseteq \mathcal{C}$ denote a set of active atomic concepts which are instantiated by a specific input $x \in \mathcal{X}(C)$. These concepts correspond to a salient feature support $\mathcal{S}(C)$, i.e., the indices of latent features that are active or relevant for these concepts.[3] For example, the concept `Spanish` may correspond not to a single direction, but to a set of feature directions that are highly active in Spanish text and encode different aspects of the language.

Estimating salient feature encodings from examples. In real LLMs, neither the salient support $\mathcal{S}(C)$ nor the salient feature encodings $A_{\mathcal{S}(C)} = \{a_i : i \in \mathcal{S}(C)\}$ are directly observed. We therefore estimate $A_{\mathcal{S}(C)}$ using residual-stream representations from examples that instantiate each concept in C. Preferably, for an atomic concept $c \in C$, a vector representation would simply use a residual-

[3] In particular, each input x has a corresponding feature vector $z(x)$; for any $x \in \mathcal{X}(C)$, $\mathrm{supp}(z(x)) \subseteq \mathcal{S}(C)$.

stream representation from a single input $x \in \mathcal{X}(c)$ that isolates concept c. In practice, the best proxy depends on how the concept is instantiated in each empirical setting. Accordingly, we approximate concept representations using a single activation vector, an average across contexts, or a set of high-variance directions across examples of the concept; in the last case, multiple feature encodings in $A_{\mathcal{S}(C)}$ would jointly represent the atomic concept. For example, in multilingual fact recall, the input `la capital de España es` instantiates both a factual concept, `capital of Spain`, and a language concept, `Spanish`. We estimate their representations separately: the fact is represented by activations from the English version of the input (`the capital of Spain is`) while the Spanish concept is represented by high-variance feature directions across Spanish texts.

Accounting for multiscale structure. Raw representations in LLMs are often dominated by large-scale structure irrelevant to the atomic concepts of interest. In our LLM experiments (Section 4), residuals cluster by prompt type, task family, and other contextual factors. (Dominant clusters are highlighted in Figure 2a.) As a result, these cluster identities can dominate the inner products used to estimate angles between atomic representations (see Appendix C). Empirically, angles computed from raw representations are dominated by background cluster identities (Appendix Figure 7(a)).

To mitigate this effect, we apply mean-centering to the dominant clusters in raw representations before estimating salient feature encodings. Let $x \in \mathcal{X}(C)$ be an example, let $\gamma(x)$ denote its background cluster, and let $\mu_{\gamma(x)}$ be the empirical mean residual representation of examples in that cluster. We define the cluster-centered residual representation as $h_c(x) = h(x) - \mu_{\gamma(x)}$. We then use these centered residuals to estimate the salient feature encodings $A_{\mathcal{S}(C)}$ used for angle computation. This centering places angle values from different background clusters on a more comparable scale, reducing artificial binning effects caused by dominant clusters (Appendix Figure 7(b)) and improving predictions (Figure 22(a,c)).

Estimating compositional interference. We define **compositional interference** (CI) for a composition with active concept set C as a normalized[4] variant of local cumulative coherence over its salient support:

$$\mathrm{CI}(C) = \max_{i \in \mathcal{S}(C)} \frac{1}{|\mathcal{S}(C)|} \sum_{j \in \mathcal{S}(C)} |\cos(a_i, a_j)|. \tag{3}$$

Here, a_i and a_j denote empirical encodings of the salient latent feature directions indexed by i and j. CI lets us estimate interference for compositional inputs without evaluating the model on any specific composed input, since a_i and a_j are estimated from examples of the constituent atomic concepts.[5] According to Section 2.1, higher interference among salient features predicts higher recovery error.

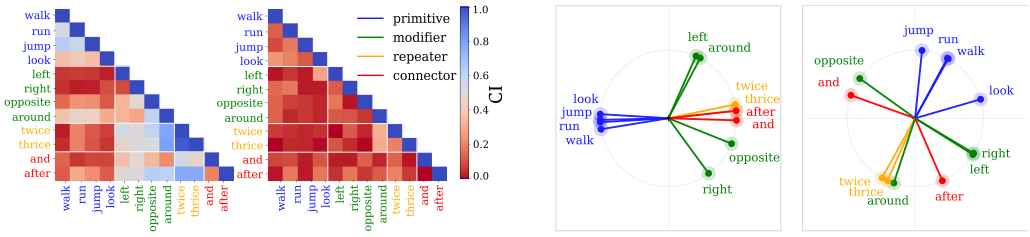
2.3 Using and evaluating predictions for ACS

We hypothesize that recovery error affects model accuracy in systematic compositional scenarios: if the model cannot reliably recover the atomic concepts in representation space, it is less likely to execute the composition correctly. This claim allows us to use CI, computed from examples of atomic concepts, for our proposed objective of Adversarial Concept Search. Generally, ACS aims to answer, "If resource constraints only allow you to test inputs for only a small subset of all possible concept combinations, which subset should you collect to maximize error coverage?" To demonstrate the efficacy of CI in choosing these adversarial scenarios, we will use it to predict compositional errors. For each experiment, we measure errors by full-prediction exact-match accuracy. We select a representation layer to provide residual-stream feature representations using a 10% validation set.

In each setting, we illustrate that CI provides a useful ranking of difficulty for the purposes of selecting a challenge set of scenarios under specified resource constraints. Assuming higher CI is associated with more compositional failure, we construct subsets of concept combinations with high CI with varying minimum cutoffs. As shown in Figure 1, the left side of the curve contains only the examples predicted to be hardest. The x-axis reports decreasing CI cutoffs by percentile, and the y-axis reports accuracy on the cumulative challenge set. If we selected the same size of test sets at random, we would expect a flat line at the model's overall accuracy. If CI is predictive, the curve should instead start low and monotonically increase toward the mean as lower-interference examples are included.

[4] By normalizing, we ensure that differences in example difficulty are not solely due to different salient feature counts.

[5] Appendix B provides alternative metrics for quantifying compositional interference.



(a) CI between atomic concepts.

(b) Atomic concept representations (PCA).

Figure 3: **Compositional models use orthogonal features.** Representations of atomic concepts in 64-dimension SCAN models trained with 8% coverage (left) and 80% coverage (right). Higher training coverage leads to representations that are more (a) pairwise orthogonal and (b) separated along principal components.

We also directly evaluate error predictions by their PR-AUC. We rank all examples by CI, assuming higher CI is associated with higher likelihood of errors. The PR-AUC baseline is the model's overall failure rate, $1 - \text{accuracy}$, which is the score of a majority-label baseline.

3 Adversarial Concept Search for a synthetic task

We first validate our hypothesis in a controlled synthetic setting by training toy models from scratch. This allows us to control the data distribution and model scale, isolate the role of feature interaction, and directly test whether representation geometry predicts compositional generalization.

Dataset The SCAN benchmark [15] is a testbed for systematic compositional generalization. SCAN specifies a set of primitive concepts, including actions (e.g., `jump`, `walk`) and operators (e.g., `left`, `twice`), which can be combined into instructions (see Figure 2b). This controlled setting allows us to examine whether the angles between atomic concepts can predict compositional failures.

3.1 Experiments

To predict model errors on SCAN, for each example, we use the primitive concepts that appear in the command as the active concept set c, and estimate $A_{\mathcal{S}(C)}$ by averaging residual activations for each primitive across contexts containing that concept, e.g., averaging activations at the `jump` token across commands containing `jump`. We compute CI over these estimated salient feature encodings and hypothesize that lower interference corresponds to near-orthogonal representations and reliable composition, while higher interference increases recovery error and leads to compositional failure.

To study compositional generalization under varying difficulty, we follow Lake and Baroni [15] in restricting the training set coverage to $\{4\%, 8\%, 16\%, 36\%, 64\%, 80\%\}$ of distinct commands while keeping the total number of training examples fixed at 100K. We provide training details in Appendix D. Models trained with lower coverage underperform on novel compositions at test time. To study how model capacity affects CI, we train autoregressive decoder-only Transformers with model sizes $d \in \{8, 12, 32, 64\}$ for each coverage ratio. Intuitively, smaller models enforce stronger compression and superposition, while larger models permit increasingly disentangled representations.

To establish SCAN as a testbed for our hypothesis, we first show that it allows us to indirectly control CI. As shown in Figure 3 and Appendix Figure 8, representation structure varies with training coverage and model capacity. Figure 3 shows that lower coverage causes representations to cluster by functional role (e.g., actions vs. compositional operators), limiting the separation between atomic concepts. As coverage increases, representations become more isotropic, creating a controlled setting in which interference varies systematically by hyperparameter.

Figure 4 confirms that example-level CI predicts compositional behavior within each SCAN model. The cumulative accuracy curves show that CI provides a useful ranking of example difficulty—stricter CI cutoffs provide harder test sets with lower model accuracy. Across all models with non-extreme accuracies,[6] this ranking is significantly better than random ordering ($p < 0.01$). Furthermore, the

[6]Defined as $0.2 < \text{acc} < 0.99$; outside of this range, compositional generalization either fails entirely or saturates.

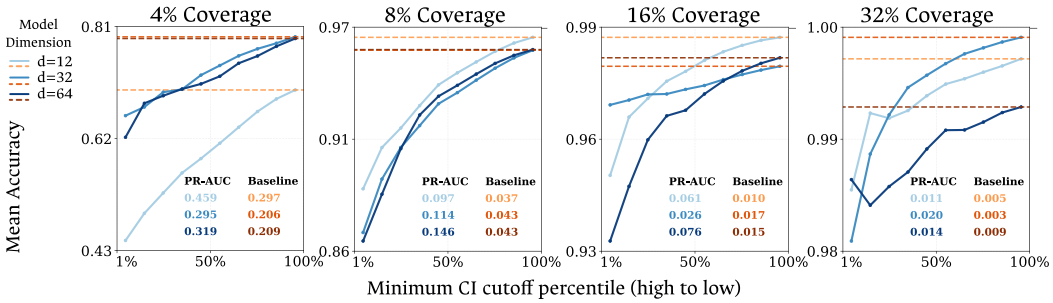


Figure 4: **SCAN models fail on examples with high interference.** Model accuracy on SCAN test sets with varying cutoffs for minimum CI. (More models in Appendix Figures 10.) Horizontal lines show each model's overall accuracy as a baseline; blue curves show test set accuracies when examples are sorted by CI. Subset accuracy responds near-monotonically to CI, significantly outperforming a random ordering. CI ranking's PR-AUC beats the error rate baseline for all models.

binned curves in Appendix Figures 11 show that the effect is not only aggregate, as accuracy is strongly negatively correlated with CI. When we control for command, length in Appendix Figures 14 and 15, the trend continues to hold. The PR-AUC metric for CI ranking consistently exceeds each model's failure-rate baseline.

Overall, CI is a reliable signal for compositional difficulty across model sizes and training regimes. Using only a single geometric property—and without access to a specific input example—we can rank compositional scenarios by difficulty.

4 Adversarial Concept Search for LLMs

We now extend our toy proof-of-concept to predict compositional errors in an LLM, `Llama-3.2-3B` [16]. We evaluate our approach on two different tasks: multihop QA and multilingual fact retrieval. This section will demonstrate the power to predict errors across diverse input distributions and surface forms.

4.1 Experiments

For both LLM tasks, we focus on examples where the LLM succeeds on the atomic components. This lets us isolate compositional errors—the model has the pieces but fails to combine them.

Multihop Reasoning. We first study two-hop factual reasoning tasks using the dataset and 10-shot prompt setup from Khandelwal and Pavlick [17] (full dataset details in Appendix E). In multihop QA, each composed query is built from two constituent atomic concepts, f and g, corresponding to the first-hop and second-hop queries. For example, a composed query $g(f(x))$ may combine an atomic first-hop query f, such as `author of 1984`, with an atomic second-hop query g, such as `birthyear of George Orwell`, into the two-hop prompt `birthyear of the author of 1984`. To focus specifically on compositional failures, we filter to examples for which the model answers the corresponding single-hop queries correctly. We extract residual activations from the last token of each atomic query and cluster-center them to obtain concept representations a_f and a_g, which estimate the salient feature encodings $A_{S(C)}$ in this setting. We then compute CI to predict failure on the composed query $g(f(x))$, using only the atomic queries for f and g.

Multilingual Fact Recall. We investigate multilingual factual recall using the KLAR dataset [18], which tests if models can answer facts across different languages consistently. When model capabilities differ by language, prior work has linked this inconsistency to a lack of alignment in internal representations [19–22]. Although this setting is not usually framed as a compositional task, prior work [18, 23–27] suggests that models implement a multi-stage process: map a non-English query onto a language-agnostic representation, retrieve the answer, then generate it in the target language. This process induces a compositionality gap analogous to multihop reasoning; even when the model can perform the atomic steps, it may still fail when they are combined. We therefore

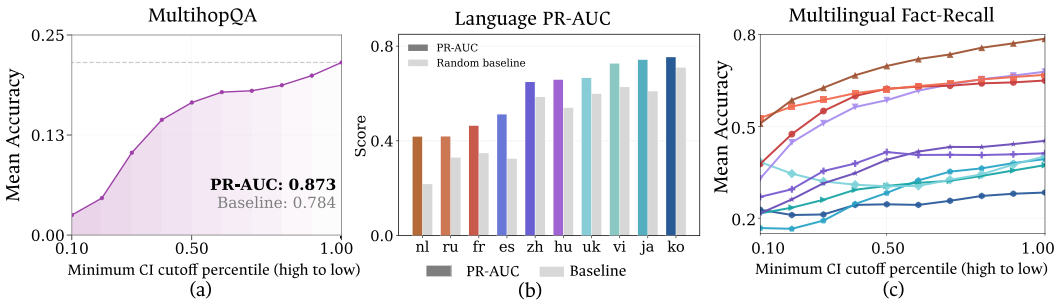


Figure 5: **LLMs fail on examples with high interference.** (a) For multihop QA, CI provides a strong ranking of example difficulty. (b) For multilingual fact recall, PR-AUC for the CI-based error prediction outperforms a majority baseline on every language. (c) For multilingual fact recall, CI ranking chooses difficult challenge sets for each language. (Each line is one language; color legend matches that in (b).) PR-AUC result details in Appendix Figure 25 and 26.

treat multilingual factual recall as requiring the composition of two atomic components: a factual representation and a language-specific representation.

For each query in target language $\ell \in \mathcal{L}$, the active concept set c consists of a factual concept q and a language concept ℓ. We extract the fact representation from the corresponding English factual query $q \in \mathcal{Q}$ and cluster-center it to obtain a_q (see Appendix C). Drawing on prior work [28], we represent the language concept ℓ as a low-rank subspace rather than a single feature vector. For each language ℓ, we collect residual representations from 8,000 samples in the multilingual OSCAR corpus [29] and apply uncentered SVD to obtain an orthonormal basis B_ℓ.[7] In this setting, we estimate $A_{\mathcal{S}(C)} = \{a_q\} \cup B_\ell$ and compute CI to predict failure on the multilingual fact query.

This setting makes the combinatorial challenge especially pronounced: each fact can be paired with any language, so the number of fact-language combinations is $|\mathcal{Q}| \times |\mathcal{L}|$. Evaluating—and, in a real-world setting, translating—all such combinations is expensive. We therefore ask whether CI, computed from an English fact activation vector and a target-language subspace, can predict cross-lingual transfer failure without access to the translated input.

4.2 Results

Ranking multihop reasoning examples. As seen in Figure 5(a), CI is highly predictive of compositional failure in multihop reasoning. The curve shows a clear monotonic trend: the LLM has lower accuracy on ACS challenge sets with stricter CI cutoffs. Appendix Figure 22(b) shows the same relationship at the bin level, confirming a strong negative correlation between CI and accuracy ($r = -0.855$). Failure PR-AUC beats the model's failure-rate baseline, indicating that CI helps identify the specific multihop examples the model gets wrong. Crucially, this prediction is obtained without evaluating the composed query itself: we can anticipate which multihop questions will be difficult solely from interference between their constituent atomic queries.

Ranking multilingual fact examples. Multilingual factual recall shows the same pattern. We explicitly evaluate trends within each language, rather than only across languages, because multilingual knowledge transfer is determined primarily by language-specific factors like resource level and orthography—factors that identify similarity to English, and therefore directly control interference with the English-language "atomic" fact. For each target language, Figure 5(c) shows that the LLM performs worse on datasets with higher interference between the English fact representation and the target-language subspace. Appendix Figure 27 confirms that accuracy also correlates with individual CI bin ranges for each language. PR-AUC metrics (Figure 5(b)) confirm the predictive accuracy of CI: for each language, CI beats the corresponding language-specific baseline.

As an illustration, Figure 6(a) visualizes example fact-language combinations, displayed with CI scores and whether the LLM answers correctly. Qualitatively, we observe that correct model predic-

[7] We retain the basis capturing 0.99 of the variance; we tune this variance threshold on the development set by sweeping 0.85, 0.90, 0.95, and 0.99.

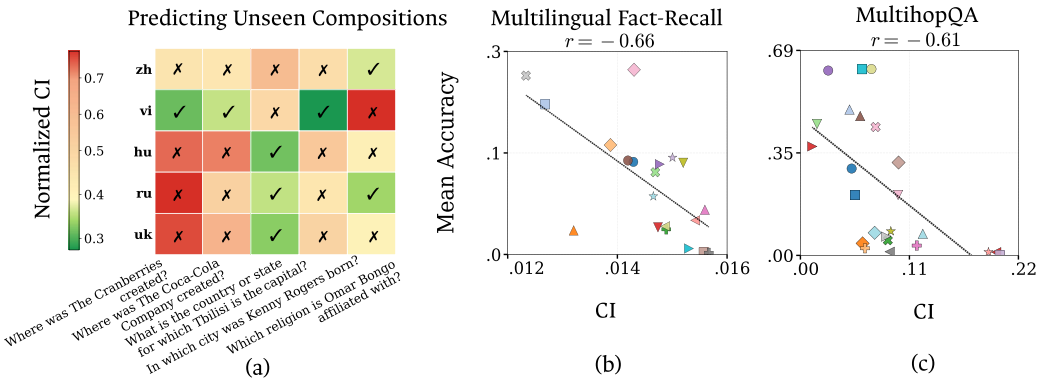


Figure 6: **LLMs fail on datasets with high interference.** (a) Fact recall across languages: columns are English facts, rows are target languages, marks indicate LLM correctness on each multilingual fact, and color shows normalized CI within each language. CI is broadly predictive of whether the model answers the compositional question correctly. (b,c) Subspace-level CI correlates strongly with accuracy for most languages in multilingual fact recall; Japanese is shown here, with all languages in Appendix Figure 30 and 31. In multihop reasoning, the correlation is weaker but follows the same trend. Both (b) and (c) show significant correlations ($p < 0.01$).

tions have a lower CI, while model errors are associated with a higher CI. Overall, the multilingual results suggest that compositional interference predicts cross-lingual transfer failures more precisely than assumptions based on the resource level of each language. It uses only the geometry between English fact representations and target-language subspaces, without access to the true translation.

Ranking datasets of coarse-grained concepts. The results thus far have shown that CI predicts the difficulty of a specific scenario. However, we can also predict the difficulty of more general scenarios—rather than evaluating the likelihood of the model answering incorrectly when prompted for `the capital of Spain` in Spanish, we can predict its overall accuracy on a dataset of national capitals in Spanish. As when we measured the angle between a fact vector against a language subspace, we will define coarse categories according to their subspaces.[8] In multilingual factual recall, CI shows a strong negative correlation with LLM accuracy for most languages (Figure 6(b), Appendix Figures 30 and 31). In multihop reasoning, the correlation is also negative and significant (Figure 6(c)). These results suggest CI can predict the general difficulty of broad compositional tasks.

5 Discussion and Future Work

Compositionality and orthogonality. Orthogonal structures are associated with adaptation to new contexts in brain scans [30, 31] and in theory. Smolensky [8] first proposed orthogonal representations as a path to variable binding in connectionist systems. Plate [32] later proposed a method for encoding systematically compositional atomic properties with quasi-orthogonal vectors in a high-dimensional space, a preview of modern overparameterized neural nets. Recent work in toy models [33] suggested that compositional models use a linearly factored representational structure. Relatedly, Olah [34] framed composition and superposition as competing strategies for allocating limited capacity among near-orthogonal feature directions. However, these previous works are limited to synthetic or small toy settings and do not attempt to explicitly predict whether a model will succeed at specific systematic combinations based on their geometry. Furthermore, compressed sensing addresses the concern of catastrophic interference between non-orthogonal feature encodings in realistic settings [35], permitting efficient compression and, in theory, decoupling orthogonality from compositionality. By assuming superposition to be lossy, we explain the importance of orthogonality in practice and operationalize our hypothesis through testable predictions in realistic and natural settings in LLMs.

[8] Similarly, we extract subspaces by applying uncentered SVD to the cluster-centered residual representations and retaining the basis vectors that explain a fixed proportion of variance. We sweep variance thresholds of 0.85, 0.90, 0.95, and 0.99 across layers on the development set, and report the best-performing configuration. When constructing category subspaces, we discard category groups with fewer than five correct examples.

Adversarial examples. Neural networks make significant classification errors after small perturbations of their inputs [36, 37]. This phenomenon has long been attributed to compression artifacts. Elhage et al. [5] found that a toy model became vulnerable to adversarial attack as superposition emerged during training, and Gorton and Lewis [38] further argued that adversarial examples arise partly from superposition-induced feature interference. In their account, adversarial attacks exploit worst-case interference: perturbations can coordinate many superposed features so that small changes accumulate into a large downstream error. Relatedly, Aden-Ali et al. [39] show that many individually small similarities between near-orthogonal dataset-example representations and a target behavior can add coherently during fine-tuning, causing the model to behave as if it had been given a hidden system prompt. Our work studies the same underlying failure mode, but focuses on the subset of salient features required by a particular composition. The underlying causes of adversarial examples are therefore closely related to the ones we exploit to identify challenging compositions, though our work explores the discrete space of input descriptions, rather than generating specific inputs.

Traditional adversarial generation requires a ground truth input and a target output. Adversarial examples are then found by locally perturbing the ground truth input to elicit the target output. In discrete modalities like language modeling this has limitations: continuous perturbations, however small, rarely map onto valid, coherent input sequences. By contrast, our method does not take a specific input or target output. Instead of searching the continuous representation space for specific errors, we search the combinatorial space of atomic concepts to identify difficult scenarios for hypothetical inputs.

Predicting model behavior In interpretability, hypotheses are often validated by predicting how a model will respond to targeted mechanistic interventions [40]. However, it rarely leverages these mechanistic insights to predict how an *unaltered* model will naturally behave when processing novel, complex *inputs*. If our goal is to understand holistic *computation* rather than merely local *implementation*, we must validate our theories by anticipating a model's edge-case failures purely from its internal representations, without requiring manual perturbations.

Recent work provides evidence that such prediction is possible. Prior studies have shown that OOD performance can be predicted from different forms of internal structure, including loss-landscape geometry [41], hidden activation patterns [42], and mechanistic accounts of character-counting circuits [43]. In contrast, Huang et al. [44] show that causal mechanisms are more predictive of OOD behavior, while Chou et al. [45] find that task-relevant geometric properties of in-distribution object manifolds forecast poor OOD generalization in image classification. More directly related to compositional generalization, An and Du [46] predict OOD compositional generalization on SCAN, and Blum et al. [27] show that the degree of representational alignment among training examples predicts cross-lingual generalization in fully trained models. However, both approaches require representations from compositional examples, whereas we only require the atomic concept representations. Overall, these analyses are constrained to simple tasks with well-understood algorithmic structure or to specific task domains. By contrast, our approach applies to any setting where models are expected to systematically compose atomic concepts.

Limitations and Future work We deliberately restrict our method to a single geometric measurement calculated on simple, easily derived atomic concept representations. This minimal setup makes the analysis transparent and suggests that representation geometry contains useful information about compositional performance. At the same time, it leaves open richer ways to identify and analyze concepts, such as SAE features, causal features, or hierarchical manifold structures [5, 14, 47–49].

For our demonstration, we explore compositional scenarios which all have existing inputs. But future work could efficiently search larger combinatorial spaces and generate *de novo* inputs from arbitrary combinations of concepts, enabling more systematic adversarial data synthesis for language models. The efficiency of the search itself could be improved; our method finds the highest interference pairs from n concept combinations with $O(n)$ vector multiplications, but a carefully-designed search could test combinations more selectively. Another gap is our exclusive focus on destructive interference, where overlap between active representations impairs independent recovery. Correlated features can also exhibit *constructive* interference, where overlap supports recovery [50]. Future work should clarify when non-orthogonality predicts failure or provides useful co-activation structure.

6 Acknowledgements

This work was enabled in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. IL and NS are supported by a grant from Coefficient Giving and the Berkeley Existential Research Institute (BERI). We thank Melanie Weber, Thomas Fel, Joshua Batson, Annabelle Michael Carrell, Aaron Mueller, Yonatan Belinkov, Jing Huang, Victoria R. Li, Ekdeep Lubana, David Klindt, Sanchit Ahuja, SueYeon Chung, Ekaterina Shutova, Sebastian Ruder, Hadas Orgad, Sweta Karlekar, Apoorv Khandelwal, Michael Lepori, Tianze Hua, Zhuonan Yang and other members of the LUNAR lab for helpful discussion and feedback on this work.

References

[1] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[2] 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.

[3] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[4] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

[5] 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.

[6] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[7] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

[8] Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

[9] Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis?, 2026. URL `https://arxiv.org/abs/2602.11246`.

[10] Brian Cheung, Alex Terekhov, Yubei Chen, Pulkit Agrawal, and Bruno Olshausen. Superposition of many models into one, 2019. URL `https://arxiv.org/abs/1902.05522`.

[11] Emmanuel Candes and Terence Tao. Decoding by linear programming, 2005. URL `https://arxiv.org/abs/math/0502327`.

[12] Zvika Ben-Haim, Yonina C. Eldar, and Michael Elad. Coherence-based performance guarantees for estimating a sparse vector under random noise. *IEEE Transactions on Signal Processing*, 58(10):5030–5043, October 2010. ISSN 1941-0476. doi: 10.1109/tsp.2010.2052460. URL `http://dx.doi.org/10.1109/TSP.2010.2052460`.

[13] Ben Adcock, Claire Boyer, and Simone Brugiapaglia. On oracle-type local recovery guarantees in compressed sensing. *Information and Inference: A Journal of the IMA*, 10(1):1–49, 2021.

[14] 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.

[15] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] Apoorv Khandelwal and Ellie Pavlick. How do language models compose functions?, 2025. URL https://arxiv.org/abs/2510.01685.

[18] Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.253. URL https://aclanthology.org/2025.acl-long.253/.

[19] Xi Ai, Mahardika Krisna Ihsani, and Min-Yen Kan. Are knowledge and reference in multilingual language models cross-lingually consistent?, 2025. URL https://arxiv.org/abs/2507.12838.

[20] Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. Language-specific latent process hinders cross-lingual performance, 2025. URL https://arxiv.org/abs/2505.13141.

[21] Yuyan Bu, Xiaohao Liu, ZhaoXing Ren, Yaodong Yang, and Juntao Dai. Align once, benefit multilingually: Enforcing multilingual consistency for llm safety alignment, 2026. URL https://arxiv.org/abs/2602.16660.

[22] Carter Blum, Katja Filippova, Ann Yuan, Asma Ghandeharioun, Julian Zimmert, Fred Zhang, Jessica Hoffmann, Tal Linzen, Martin Wattenberg, Lucas Dixon, and Mor Geva. Beyond the rosetta stone: Unification forces in generalization dynamics, 2025. URL https://arxiv.org/abs/2508.11017.

[23] Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. Paths Not Taken: Understanding and Mending the Multilingual Factual Recall Pipeline, May 2025. URL http://arxiv.org/abs/2505.20546. arXiv:2505.20546 [cs].

[24] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

[25] Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319, 2024.

[26] Nadav Shani and Ali Basirat. Language dominance in multilingual large language models. In Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hosein Mohebbi, Hanjie Chen, Dana Arad, and Gabriele Sarti, editors, *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 137–148, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-346-3. doi: 10.18653/v1/2025.blackboxnlp-1.7. URL https://aclanthology.org/2025.blackboxnlp-1.7/.

[27] Carter Blum, Katja Filippova, Ann Yuan, Asma Ghandeharioun, Julian Zimmert, Fred Zhang, Jessica Hoffmann, Tal Linzen, Martin Wattenberg, Lucas Dixon, and Mor Geva. Beyond the Rosetta Stone: Unification Forces in Generalization Dynamics, August 2025. URL http://arxiv.org/abs/2508.11017. arXiv:2508.11017 [cs].

[28] Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations, 2022. URL https://arxiv.org/abs/2205.10964.

[29] Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *CMLC 2021-9th Workshop on Challenges in the Management of Large Corpora*, 2021.

[30] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270, 2022.

[31] Lennart Luettgau, Nan Chen, Tore Erdmann, Sebastijan Veselic, Rani Moran, Zeb Kurth-Nelson, and Raymond J Dolan. A neural mechanism for compositional generalization of structure in humans. *bioRxiv*, pages 2024–09, 2024.

[32] Tony A Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, 6 (3):623–641, 1995.

[33] Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Does data scaling lead to visual compositional generalization?, 2025. URL https://arxiv.org/abs/2507.07102.

[34] Chris Olah. Distributed representations: Composition & superposition. https://transformer-circuits.pub/2023/superposition-composition/index.html, May 2023. Transformer Circuits Thread. Published May 4, 2023.

[35] Vitória Barin Pacela, Shruti Joshi, Isabela Camacho, Simon Lacoste-Julien, and David Klindt. Stop probing, start coding: Why linear probes and sparse autoencoders fail at compositional generalisation. *arXiv e-prints*, pages arXiv–2603, 2026.

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

[37] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

[38] Liv Gorton and Owen Lewis. Adversarial examples are not bugs, they are superposition. *arXiv preprint arXiv:2508.17456*, 2025.

[39] Ishaq Aden-Ali, Noah Golowich, Allen Liu, Abhishek Shetty, Ankur Moitra, and Nika Haghtalab. Subliminal effects in your data: A general mechanism via log-linearity, 2026. URL https://arxiv.org/abs/2602.04863.

[40] Naomi Saphra and Sarah Wiegreffe. Mechanistic?, 2024. URL https://arxiv.org/abs/2410.09087.

[41] Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. Linear connectivity reveals generalization strategies, 2023. URL https://arxiv.org/abs/2205.12411.

[42] Victoria R. Li, Jenny Kaufmann, Martin Wattenberg, David Alvarez-Melis, and Naomi Saphra. Can interpretation predict behavior on unseen data?, 2025. URL https://arxiv.org/abs/2507.06445.

[43] Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task, 2026. URL https://arxiv.org/abs/2601.04480.

[44] Jing Huang, Junyi Tao, Thomas Icard, Diyi Yang, and Christopher Potts. Internal causal mechanisms robustly predict language model out-of-distribution behaviors. *arXiv preprint arXiv:2505.11770*, 2025.

[45] Chi-Ning Chou, Artem Kirsanov, Yao-Yuan Yang, and SueYeon Chung. Diagnosing generalization failures from representational geometry markers, 2026. URL https://arxiv.org/abs/2603.01879.

[46] Zhiyu An and Wan Du. Representational homomorphism predicts and improves compositional generalization in transformer language model, 2026. URL https://arxiv.org/abs/2601.18858.

[47] Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models, 2025. URL https://arxiv.org/abs/2502.12892.

[48] Chi-Ning Chou, Hang Le, Yichen Wang, and SueYeon Chung. Feature learning beyond the lazy-rich dichotomy: Insights from representational geometry, 2025. URL https://arxiv.org/abs/2503.18114.

[49] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[50] Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro AM Mediano. From data statistics to feature geometry: How correlations shape superposition. *arXiv preprint arXiv:2603.09972*, 2026.

[51] Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, 2023.

A Recovery bounds related to local cumulative coherence

This section expands on our claim that robust recovery bounds can be controlled by local cumulative coherence. Specifically, we want to recover an estimate $\hat{z}$ for an m-dimensional k-sparse feature vector z from noisy conditions. Robust recovery guarantees provide high probability bounds on the recovery error $\|\hat{z} - z\|_2$.

A.1 Robust compressed sensing

The specific robust compressed sensing bound provided by Adcock et al. [13] limits the number of measurements required to guarantee a high probability of exact recovery. As is common in the superposition literature, we treat the hidden dimension d as the number of sampled measurements. In robust compressed sensing, we assume that rather than access to the encoded Az, we instead observe the noisy representation $Az + \epsilon$, where the noise is bounded by some constant $\|\epsilon\| < \eta$. Therefore, the relevant bound can be paraphrased as claiming that we can guarantee the recovery error on our estimated $\hat{z}$ from a noisy encoding satisfies,

$$\|z - \hat{z}\|_2 \leq (c_1 + c_2\sqrt{k})\eta, \tag{4}$$

for constants c_1, c_2.

Adcock et al. [13] guarantee this robust recovery with probability $1 - \varepsilon$ if the hidden dimension d is at least,

$$d \gtrsim \Theta(\text{supp}(z), A) \log^2(m/\varepsilon), \tag{5}$$

where we define the interaction term as a positive real number,

$$\Theta(\text{supp}(z), A) = \|A_{\mathcal{S}}^\top A_{\text{supp}(z)}\|_{\infty\to\infty} \tag{6}$$

In Equation 2, we refer to the bound in Equation 5 as effectively controlled by mutual coherence in the salient support,

$$\alpha(\mathcal{S}) = \max_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}} |\cos(a_i, a_j)| \tag{7}$$

$$\geq \max_{i\in\mathcal{S}} \sum_{j\in\text{supp}(z)} |\cos(a_i, a_j)| \tag{8}$$

$$= \Theta(\text{supp}(z), A). \tag{9}$$

Therefore we can satisfy Equation 5 if hidden dimension d is at least,

$$d \gtrsim \alpha(\mathcal{S}) \log^2(m/\varepsilon). \tag{10}$$

A.2 Robust linear compressed sensing

We now assume the *linear* compressed sensing setting of Garg et al. [9], which weakens the lossless recovery guarantee from compressed sensing to *near-lossless* recovery by accepting the Linear Representation Hypothesis (LRH). Under the LRH, Garg et al. [9] showed that a linear decoder of dimension d can still recover a number of features exponential in d up to a small constant error ϵ.

In theory, the number of k-sparse features which can be linearly accessible is exponential in d [9]—even if the encoded features are linearly correlated. However, the theory only guarantees that this near-lossless decoding *exists*, not that it is applied by the neural network. We argue that if the neural network's next layer does linearly decode the representation, it may not reflect the theoretical ideal probe dictionary. We will treat the LLM's linear encoding as explicit in its activations, but treat its linear decoding as implicit. We assume this implicit linear decoding to be a perturbation of the ideal decoding, leading to the following argument.

Let $A \in \mathbb{R}^{d\times m}$ be the representation matrix (columns $a_1, \ldots, a_m$) and $B \in \mathbb{R}^{d\times m}$ be the probe matrix (columns $b_1, \ldots, b_m$). Given a feature vector $z \in [-1, 1]^m$ with k nonzero features, the decoded estimate is

$$\hat{z} = B^\top Az.$$

Due to the results of Garg et al. [9], we can assume that for all k-sparse z,

$$\|\hat{z} - z\|_\infty \leq \varepsilon.$$

Now assume that the implicit linear decoding during downstream processing, $\tilde{B}$, is a perturbation of the ideal probe B, specifically $\tilde{B} = B + \Delta B$ where

$$\|\Delta B\|_{2\to\infty} \;\leq\; \eta.$$

We will show that this perturbation has its largest impact on recovery error when the sparse active features have highly correlated encodings.

Theorem 1 (Perturbation sensitivity bounded by correlations in $A_{\mathcal{S}}$)**.** *Let $\tilde{B} = B + \Delta B$, and suppose each probe vector changes by at most η:*

$$\|\Delta b_i\|_2 \leq \eta \quad \textit{for all } i.$$

Let z be a k-sparse feature vector with support $\mathrm{supp}(z)$*. Then*

$$\|\tilde{B}^\top A z - z\|_\infty \;\leq\; \varepsilon \;+\; \eta\, k\, \|A_{\mathrm{supp}(z)}\|_2. \tag{11}$$

In particular, if the product $\langle a_i, a_j\rangle$ for $i, j \in \mathcal{S}$ is large and positive, that feature pair increases the perturbation sensitivity on this input.

Proof. We decompose the new error into the original decoding error plus the effect of changing the probe:

$$\tilde{B}^\top A z - z = (B^\top A z - z) + \Delta B^\top A z.$$

By assumption, the first term has $\|B^\top A z - z\|_\infty \leq \varepsilon$, so we only need to bound $\|\Delta B^\top A z\|_\infty$. The i-th coordinate of this perturbation is bounded using Cauchy–Schwarz and the per-column bound $\|\Delta b_i\|_2 \leq \eta$,

$$|(\Delta B^\top A z)_i| \;\leq\; \|\Delta b_i\|_2\, \|A z\|_2 \;\leq\; \eta\, \|A z\|_2.$$

Since z is supported on $\mathrm{supp}(z)$, we can express the above using only the nonzero features and write $Az = A_{\mathrm{supp}(z)} z$, where $A_{\mathrm{supp}(z)}$ is the restriction of A to indices in $\mathrm{supp}(z)$. Using Cauchy–Schwarz and the k-sparsity constraint on z,

$$|(\Delta B^\top A z)_i| \;\leq\; \eta\, \|A_{\mathrm{supp}(z)} z\|_2 \;\leq\; \eta\, \|A_{\mathrm{supp}(z)}\|_2\, \|z\|_2 \;\leq\; \eta\, k\, \|A_{\mathrm{supp}(z)}\|_2.$$

Taking the maximum over all coordinates i,

$$\|\tilde{B}^\top A z - z\|_\infty \;\leq\; \varepsilon \;+\; \eta\, k\, \|A_{\mathrm{supp}(z)}\|_2. \tag{12}$$

□

Under Theorem 1, deviations from an ideal decoder can damage feature recovery more on inputs when encodings of the active salient support are correlated. If an LLM implicitly processes feature encodings with a flawed decoder, it will make therefore more mistakes when composing non-orthogonal encodings.

Note that the quantity of interest in our robust recovery bound differs depending on whether the model's intrinsic decoding is linear or nonlinear. If we assume the model decodes nonlinearly, as in compressed sensing, then the quantity of interest is cumulative coherence in the salient support. If we assume the model uses linear decoding, then the quantity of interest is a simple norm on the active support—a subset of the salient support, and one that is not identifiable without analyzing specific input instantiating the scenario. Without studying specific inputs, we remain limited to measurements on the salient support, not the active subset.

Either way, cosine similarity within the salient support causes destructive interference.

B Alternative Metrics to Capture Compositional Interference

In the paper, we use cumulative coherence derived from Adcock et al. [13] and measure this quantity. We consider several alternative geometric metrics for estimating compositional interference among the salient features of a composition. Let $\mathcal{S}(C)$ denote the salient feature support, and let a_i be the representation vector associated with feature i. All metrics below are computed over pairwise similarities among active feature directions.

Minimum similarity. The minimum similarity captures the least aligned pair of salient features:

$$\mathrm{CI}_{\min}(C) = \min_{\substack{i,j \in \mathcal{S}(C) \\ i \neq j}} |\cos(a_i, a_j)|. \tag{13}$$

Maximum similarity. The maximum similarity captures the most aligned pair of salient features:

$$\mathrm{CI}_{\max}(C) = \max_{\substack{i,j \in \mathcal{S}(C) \\ i \neq j}} |\cos(a_i, a_j)|. \tag{14}$$

Mean similarity. The mean similarity averages pairwise alignment across all salient feature pairs:

$$\mathrm{CI}_{\mathrm{mean}}(C) = \frac{1}{|\mathcal{S}(C)|(|\mathcal{S}(C)| - 1)} \sum_{\substack{i,j \in \mathcal{S}(c) \\ i \neq j}} |\cos(a_i, a_j)|. \tag{15}$$

When $|\mathcal{S}(C)| = 2$, different CI aggregation metrics reduce to the same pairwise comparison and therefore yield the same correlational result. When $|\mathcal{S}(C)| > 2$, however, these metrics can diverge because they aggregate multiple pairwise interactions differently. Therefore, whenever examples contain more than two salient features, we additionally report results for alternative aggregation metrics to enable comparison across definitions of interference (Figures 16, 17, 20, 21, 19, 18, 28, 29, and Table 5).

Across these comparisons, most similarity-based metrics remain predictive in most settings. Since the alternative metrics are partly intercorrelated, this suggests that the central signal comes from angular similarity among concept representations. Nevertheless, empirical results in the paper show that the cumulative coherence metric derived from the bounds in Section 2.2 is the most stable metric to measure compositional interference across all scenarios.

C Cluster Mean Centering

We expand on Section 2.2, "Accounting for multiscale structure," to provide more mathematical intuition and details.

For an input $x \in \mathcal{X}(c)$, let C denote the active concept set, let $\mathcal{S}(C)$ denote its salient support, and let $A_{\mathcal{S}(C)} = \{a_i : i \in \mathcal{S}(C)\}$ denote the corresponding salient feature encodings. To provide a simple intuition, we consider the setting in which each salient feature is represented by a single linear direction, and we use the linear representation hypothesis to model residual stream activations. The empirical estimate $\tilde{a}_i$ of a salient feature encoding may contain not only the direction of interest, but also background components induced by prompt format, task family, language, or other contextual structure. We write

$$\tilde{a}_i = z_i a_i + \sum_{u \in \mathcal{B}_i} z_u^{(i)} a_u,$$

where $z_i a_i$ is the salient feature contribution we aim to isolate, z_i is its activation coefficient, and $\mathcal{B}_i$ indexes other background directions mixed into the empirical estimate of feature i. Then the raw inner product between two empirical estimates $\tilde{a}_i$ and $\tilde{a}_j$, for $i, j \in \mathcal{S}(c)$, expands as

$$\begin{aligned} \langle \tilde{a}_i, \tilde{a}_j \rangle = z_i z_j \langle a_i, a_j \rangle &+ z_i \sum_{\ell \in \mathcal{B}_j} z_\ell^{(j)} \langle a_i, a_\ell \rangle \\ &+ z_j \sum_{u \in \mathcal{B}_i} z_u^{(i)} \langle a_u, a_j \rangle \\ &+ \sum_{u \in \mathcal{B}_i} \sum_{\ell \in \mathcal{B}_j} z_u^{(i)} z_\ell^{(j)} \langle a_u, a_\ell \rangle. \end{aligned} \tag{16}$$

The first term is the salient feature interaction we aim to measure, while the remaining terms arise from background structure. Thus, raw inner products between empirical feature estimates can be dominated by shared background structure rather than by the local interaction of interest.

To mitigate this effect, we apply cluster mean-centering at the level of residual representations before estimating salient feature encodings. Let $\gamma(x)$ denote the background cluster associated with example x, and let $\mu_{\gamma(x)}$ be the empirical mean residual representation of examples in that cluster. We define the centered residual representation as

$$h_c(x) = h(x) - \mu_{\gamma(x)}.$$

We then use these centered residuals to estimate the salient feature encodings $A_{\mathcal{S}(C)}$ used for angle computation. Mean-centering reduces dominant cluster-level structure and isolates residual variation that more directly reflects interactions between salient feature directions.

In this work, we identify background clusters by visualizing empirical representations with SVD and observing that they group by prompt type, task family, language, or other contextual factors. We then center each representation by the empirical mean of its corresponding cluster. Automatically discovering these background clusters is an important direction for future work.

Beyond the mathematical intuition, empirical observations in Appendix Figures 7, 22, 23, and 24 further motivate the need for cluster mean-centering.

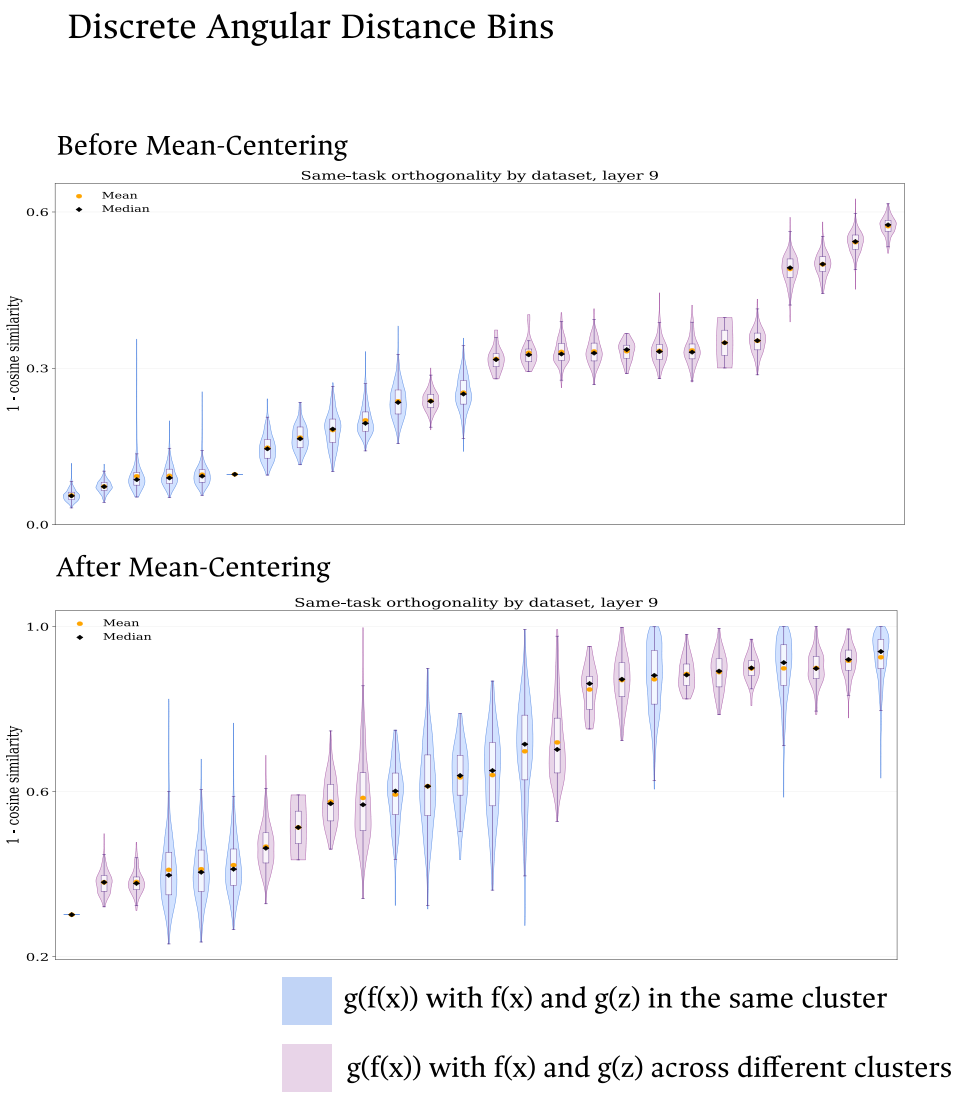


Figure 7: Angle distributions of concept interactions in the multi-hop dataset before and after mean-centering. The top panel shows that, without mean-centering, angles between interacting concepts from different clusters are dominated by global clustering structure. After mean-centering, shown in the bottom panel, this global effect is reduced, allowing us to better capture local interactions between concepts.

Cov.	$d = 8$	$d = 12$	$d = 32$	$d = 64$
4%	0.169	0.703	0.794	0.792
8%	0.468	0.963	0.957	0.957
16%	0.767	0.990	0.983	0.985
32%	0.950	0.995	0.997	0.992
64%	0.947	0.995	0.999	0.998
80%	0.960	0.996	0.998	0.999

(a) Average accuracy.

Cov.	$d = 8$	$d = 12$	$d = 32$	$d = 64$
4%	-0.067^{***}	-0.283^{***}	-0.183^{***}	-0.219^{***}
8%	-0.121^{***}	-0.160^{***}	-0.166^{***}	-0.176^{***}
16%	-0.075^{***}	-0.112^{***}	-0.045^{***}	-0.139^{***}
32%	-0.140^{***}	-0.049^{***}	-0.073^{***}	-0.038^{***}
64%	-0.181^{***}	0.030^{*}	-0.015	-0.046^{***}
80%	-0.078^{***}	0.051^{**}	0.033^{*}	-0.042^{*}

(b) Point-biserial correlation with CI.

Table 1: SCAN results across training coverage and model dimension. Left: average exact-match accuracy. Right: point-biserial correlation r_{pb} between CI score and binary correctness. The point-biserial correlation is the Pearson correlation between a continuous variable and a binary variable; here, it measures whether examples with higher CI are less likely to be answered correctly. Since the original correlations were computed with $1 - \mathrm{CI}$, we flip their signs when reporting correlations with CI. Except in near-saturated regimes with extreme accuracies, correlations are consistently negative, indicating that higher CI predicts lower item-level correctness. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 10^{-3}$.

D SCAN Additional Details

We expand on the experimental details from Section 3.

Training Details We train decoder-only Transformers on the SCAN benchmark using a size-variation data regime. Specifically, we vary the fraction of training commands seen by the model across coverage ratios of $4\%, 8\%, 16\%, 32\%, 64\%$, and 80% of the full SCAN training split. We evaluate four model sizes with hidden dimension $d \in \{8, 12, 32, 64\}$. Each model uses a feed-forward dimension of $4d$, 4 attention heads, and 10 Transformer layers. All models are trained with the Adam optimizer using a learning rate of 10^{-3} and batch size 256. Models are trained for up to 3,000–4,000 epochs with early stopping: training halts if development-set exact-match accuracy does not improve for 200 epochs for $d = 8$, or for 300 epochs for $d \in \{12, 32, 64\}$. The development set is drawn as 10% of the held-out test split.

Input sequences are formatted as `[BOS] input [SEP] output [EOS]`, and the training loss is cross-entropy computed only over the output tokens. During evaluation, we use greedy autoregressive decoding and report full-sequence exact-match accuracy. All runs use a single random seed (42).

Additional Results Figure 8 and Figure 9 visualize how atomic concept representations vary with training coverage and model capacity. Figure 10 and Figure 11 show that examples with higher CI tend to have lower accuracy, supporting CI as a useful ranking signal for challenge-set construction. Figure 12 and Figure 13 further report the corresponding PR-AUC results, showing that CI provides a predictive signal across coverage and model-size settings. Table 1 shows that CI is negatively correlated with correctness under point-biserial correlation in most settings. Finally, to verify that CI is not merely tracking dataset distribution, specifically sequence length, Figure 14 and Figure 15 break the results down by command length and show that the trend largely persists within length groups.

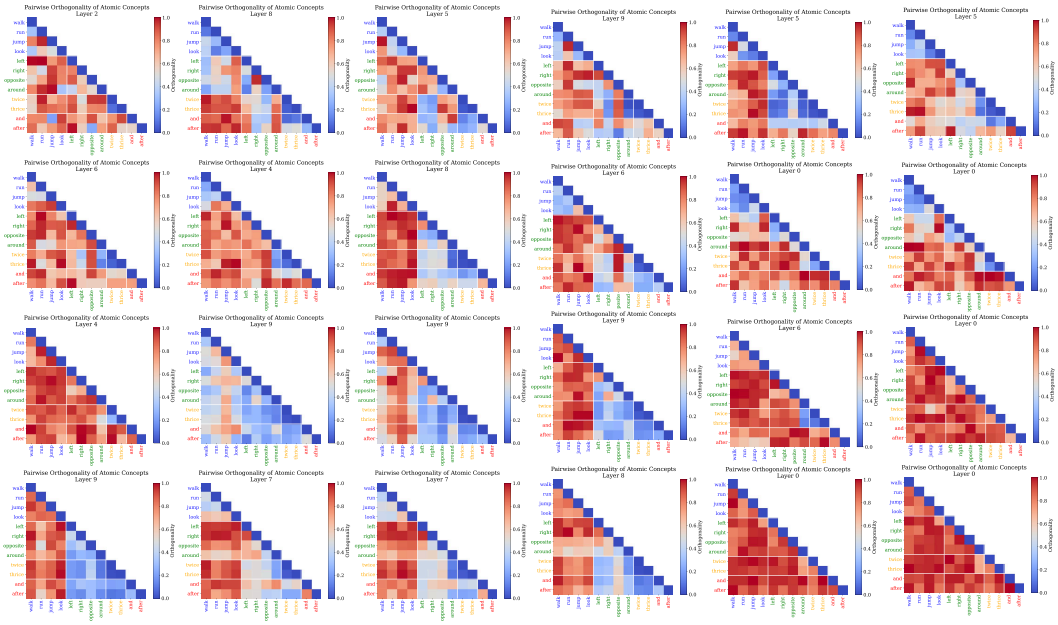


Figure 8: Heatmaps of pairwise angular distances between atomic concepts in the SCAN dataset, computed from models with different hidden dimensions and training coverage levels. Columns correspond to training coverage ratios of $4\%, 8\%, 16\%, 36\%$, and 64%, while rows correspond to model dimensions $d \in \{8, 12, 32, 64\}$. For each setting, we show the layer selected on the development set based on the best AUC.

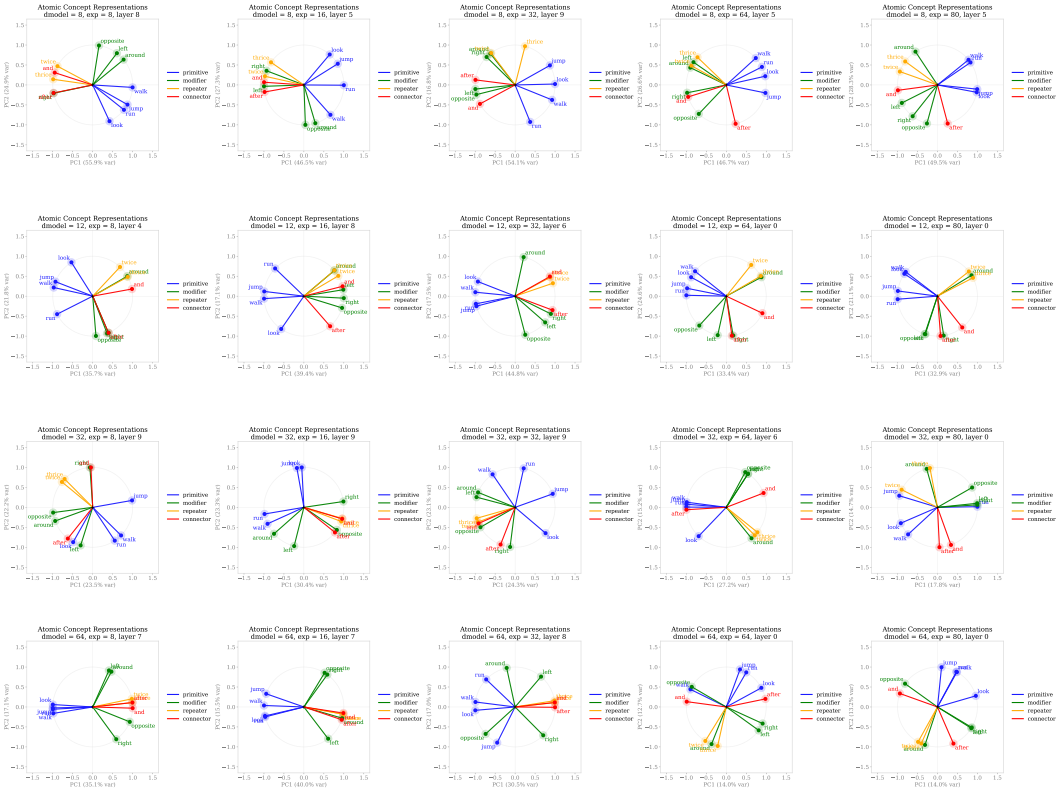


Figure 9: PCA visualization of atomic concept representations in the SCAN dataset computed from models with different hidden dimensions and training coverage levels. Columns correspond to training coverage ratios of $4\%, 8\%, 16\%, 36\%$, and 64%, while rows correspond to model dimensions $d \in \{8, 12, 32, 64\}$. For each setting, we show the layer selected on the development set based on the best AUC.

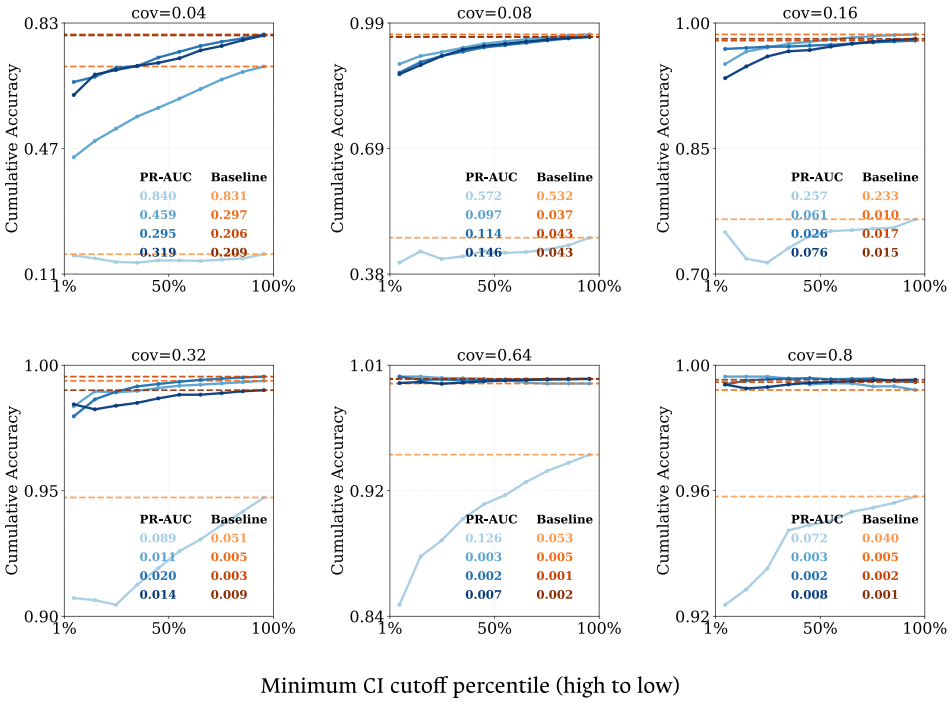


Figure 10: Cumulative plots for models of all sizes on the SCAN dataset across training coverage ratios.

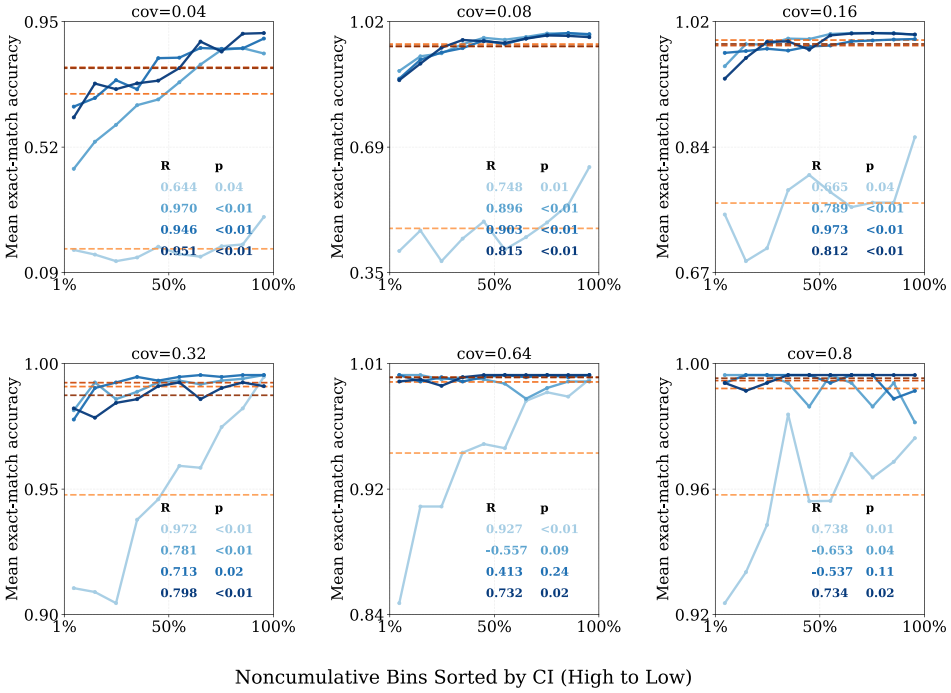


Figure 11: Noncumulative plots for models of all sizes on the SCAN dataset across training coverage ratios.

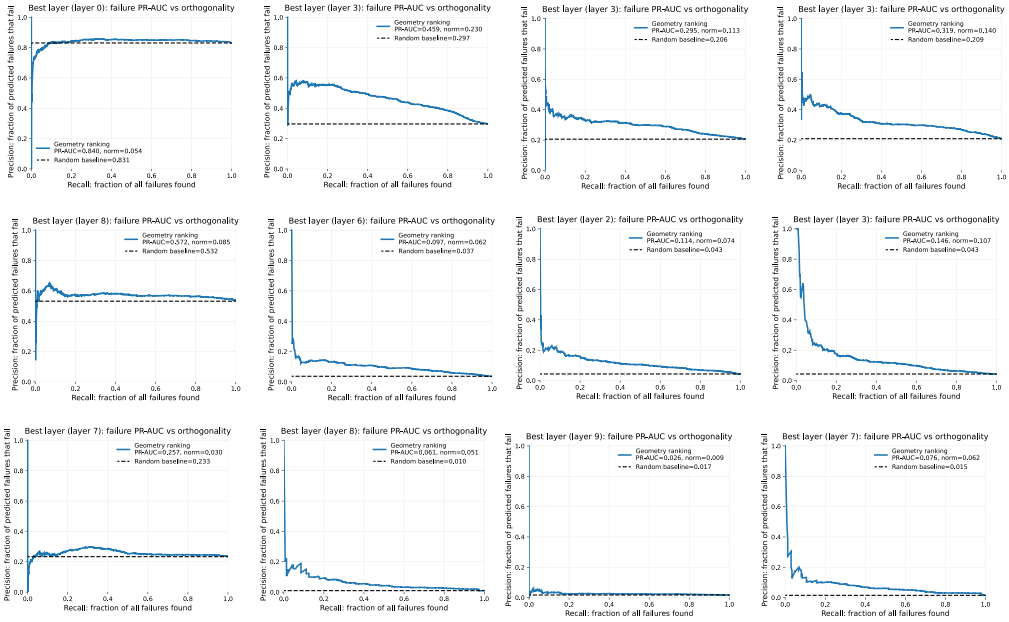


Figure 12: PR-AUC for different model variants on the SCAN dataset. Rows correspond to coverage ratios $cov \in \{4\%, 8\%, 16\%\}$, and columns correspond to model dimensions $d \in \{8, 12, 32, 64\}$. PR-AUC for different model variants on the SCAN dataset, where CI is computed as the mean cosine similarity between interacting concepts.

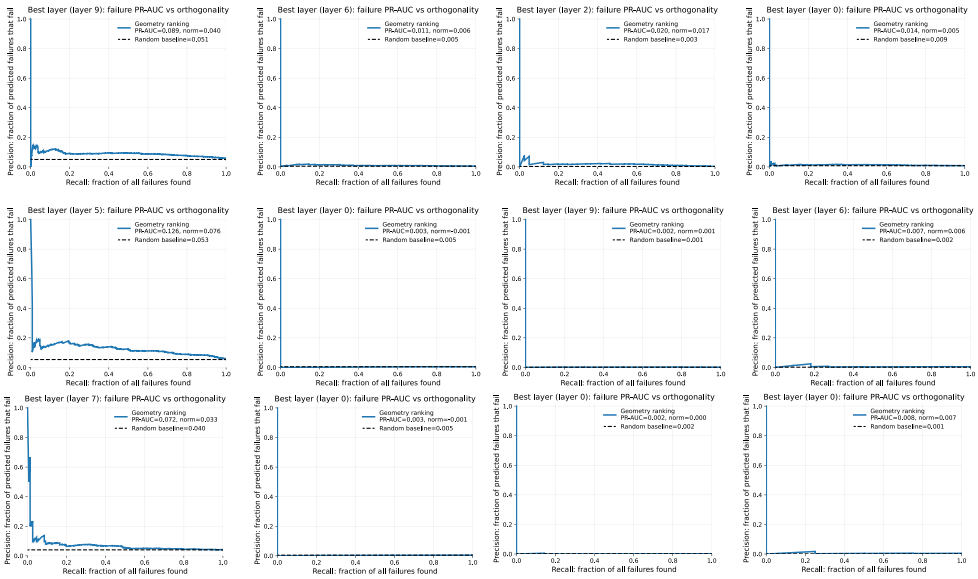


Figure 13: PR-AUC for different model variants on the SCAN dataset. Rows correspond to coverage ratios $cov \in \{32\%, 64\%, 80\%\}$, and columns correspond to model dimensions $d \in \{8, 12, 32, 64\}$. PR-AUC for different model variants on the SCAN dataset, where CI is computed as the mean cosine similarity between interacting concepts.

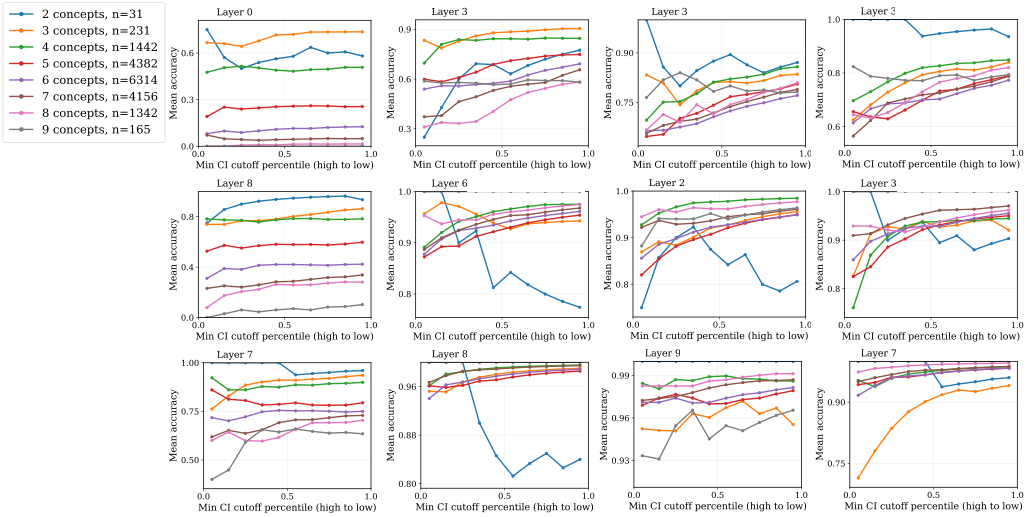


Figure 14: SCAN predictive trends broken down by command length for $\mathrm{CI}_{\mathrm{mean}}$. Rows correspond to coverage ratios $cov \in \{4\%, 8\%, 16\%\}$, and columns correspond to model dimensions $d \in \{8, 12, 32, 64\}$. To check whether the main trend is driven only by command length, we plot cumulative accuracy curves separately for each length group at the selected layers. The trend is generally consistent across length groups, but is weaker for the shortest commands, especially for 2-concept examples. Manual inspection of this group suggests that many failures reflect operator confusion rather than interference between the action and operator representations: for example, the model may map `jump twice` to three jumps, or produce four `walk` actions for `walk thrice`. Higher-length groups contain more active components and more directly reflect the multi-compositional interference captured by our metric.

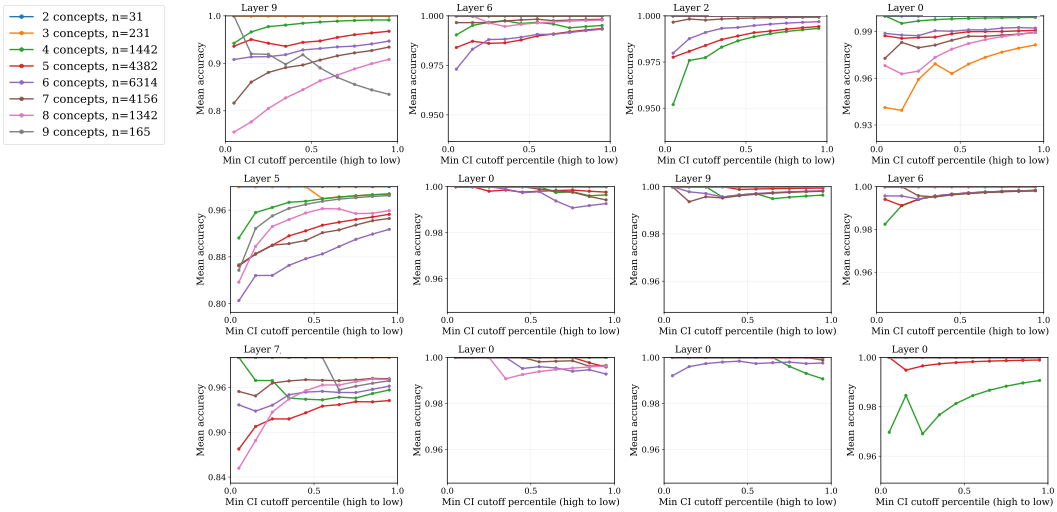


Figure 15: SCAN predictive trends broken down by command length for $\mathrm{CI}_{\mathrm{mean}}$. Rows correspond to coverage ratios $cov \in \{32\%, 64\%, 80\%\}$, and columns correspond to model dimensions $d \in \{8, 12, 32, 64\}$.

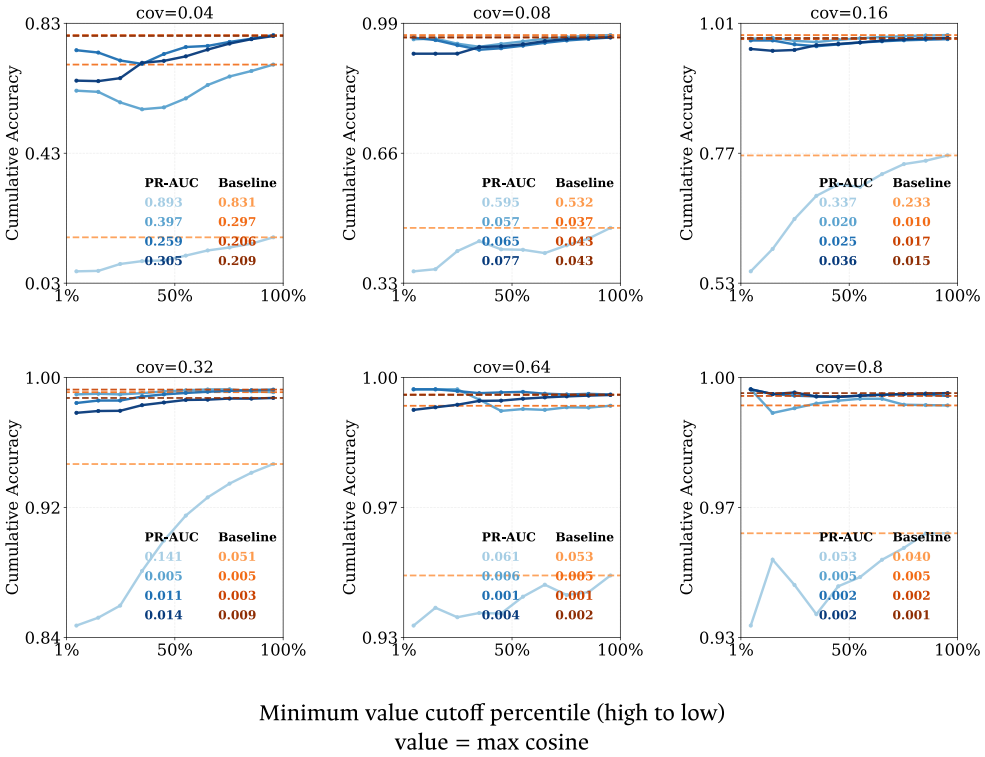


Figure 16: Cumulative plots for models of all sizes on the SCAN dataset across training coverage ratios, where the X axis is the value when computing the interference using maximum cosine.

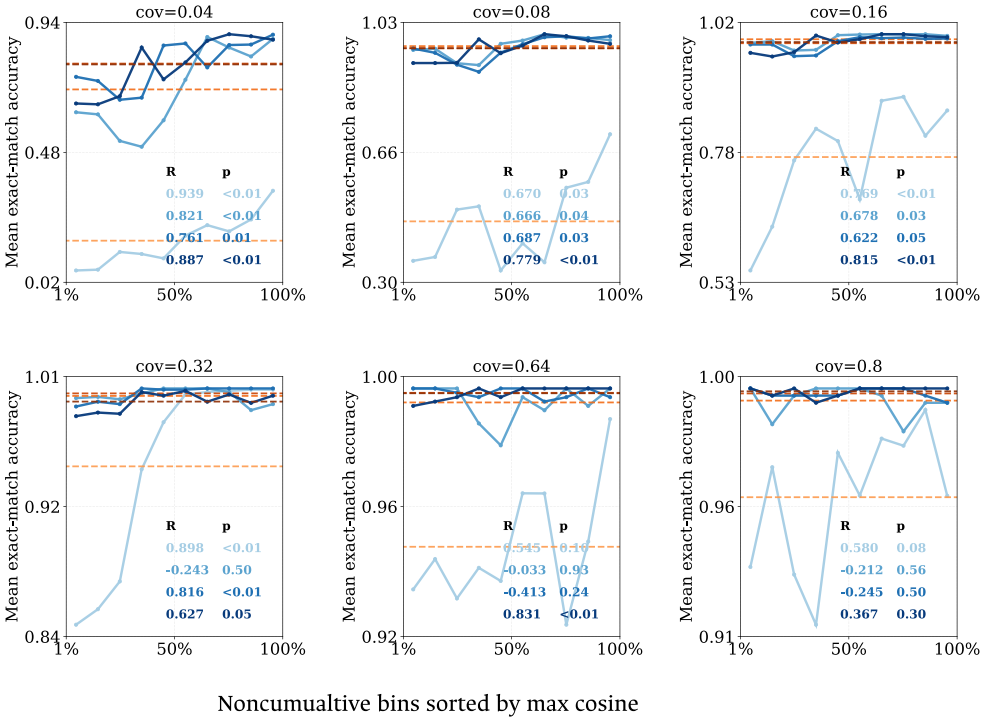


Figure 17: Noncumulative plots for models of all sizes on the SCAN dataset across training coverage ratios, where the X axis is the value when computing the interference using maximum cosine.

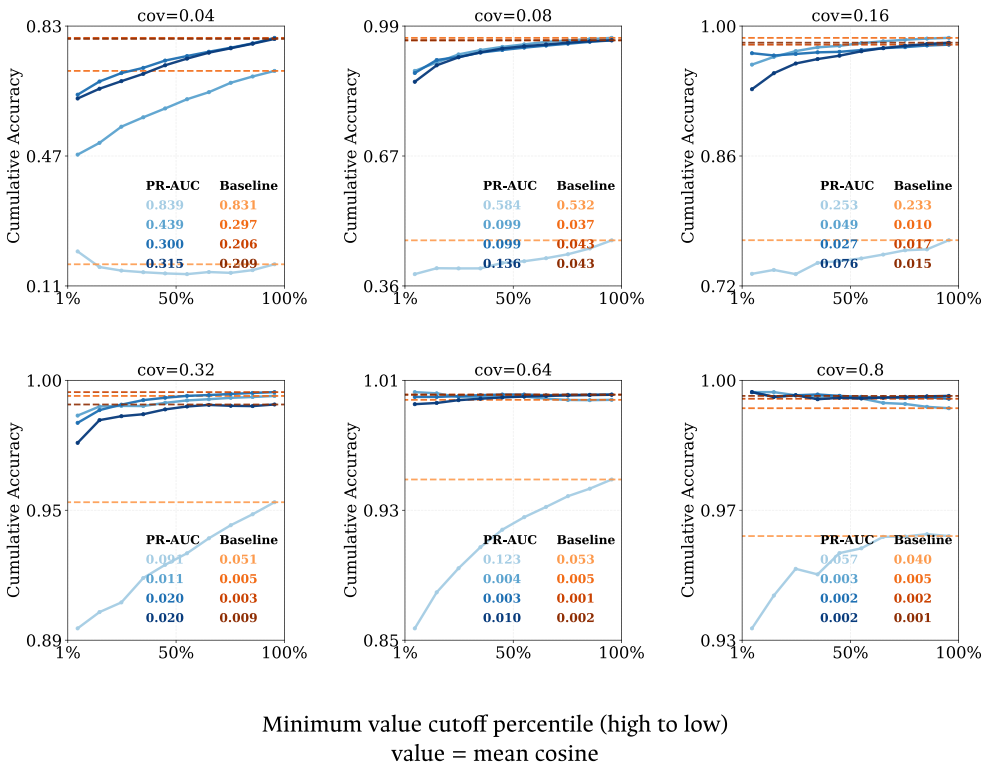


Figure 18: Cumulative plots for models of all sizes on the SCAN dataset across training coverage ratios, where the X axis is the value when computing the interference using mean cosine.

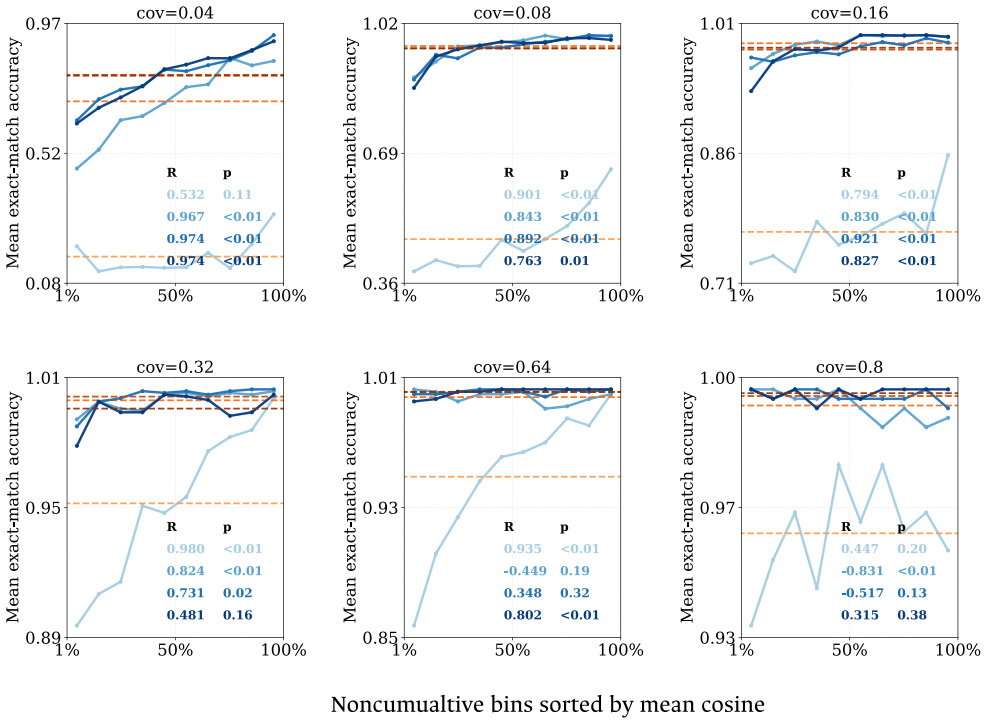


Figure 19: Noncumulative plots for models of all sizes on the SCAN dataset across training coverage ratios, where the X axis is the value when computing the interference using mean cosine.

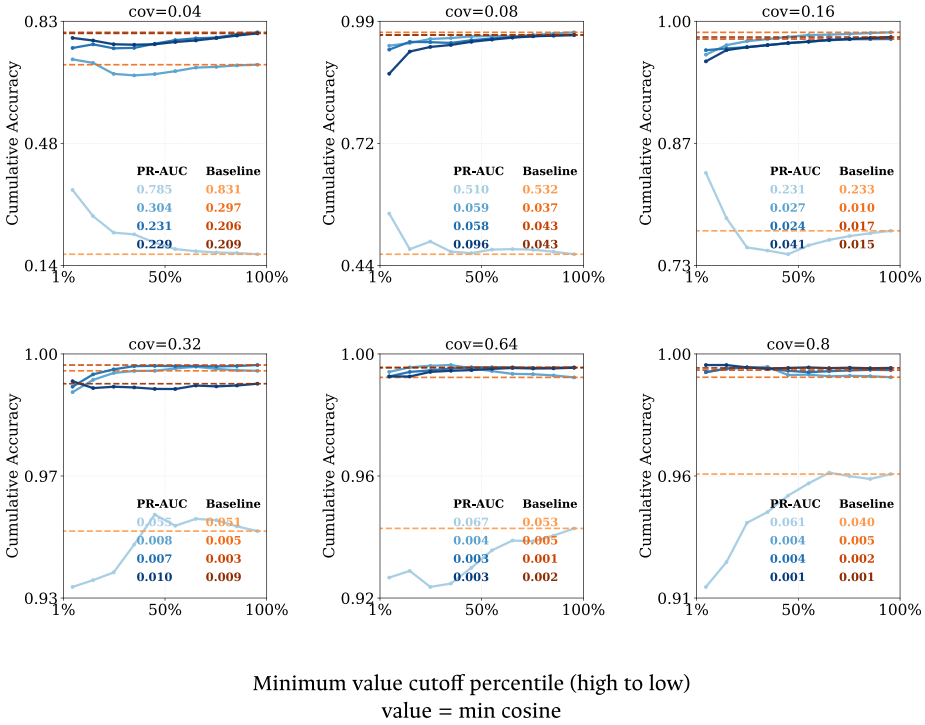


Figure 20: Cumulative plots for models of all sizes on the SCAN dataset across training coverage ratios, where the X axis is the value when computing the interference using minimum cosine.

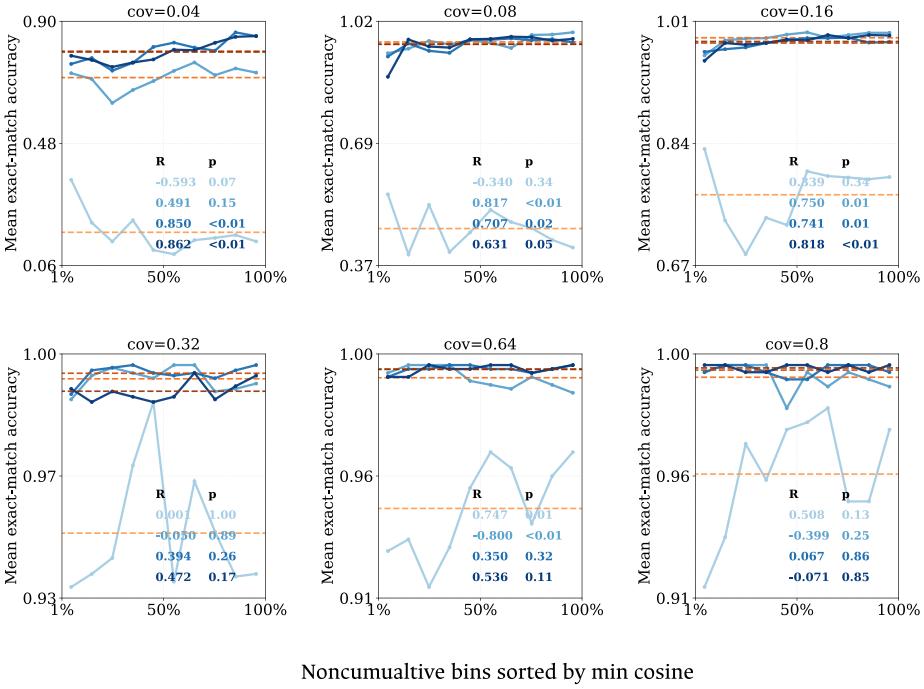


Figure 21: Noncumulative plots for models of all sizes on the SCAN dataset across training coverage ratios, where the X axis is the value when computing the interference using minimum cosine.

E Multihop Dataset Details

We expand on the dataset details provided in Section 4 for multihop QA task. We study two-hop factual reasoning tasks using the datasets and 10-shot prompting setup from Khandelwal and Pavlick [17]. To test whether our method generalizes beyond common compositions, we additionally construct several datasets targeting less common multihop compositions that span different representation clusters. These datasets include `int-plus8-parity`, `int-plus2-parity`, `int-plus5-parity`, `int-plus2-str`, `int-plus5-str`, `int-plus8-str`, and `artist-birthyear-times-two`. Since these are structured tasks, we construct the integer-based datasets by sampling 1000 valid input–output combinations from the space of possible cases. The exception is `artist-birthyear-times-two`, which we construct by adapting the original artist birth-year data from Khandelwal and Pavlick [17].

We filter to examples for which the model answers both constituent single-hop queries correctly, ensuring that errors reflect compositional failures rather than missing single-hop knowledge. In total, our multihop evaluation contains 26 datasets.

F Multihop Additional Results

We expand on the experimental details for the multihop QA task from Section 4.

Figure 22 shows that CI predicts multihop failures, with both cumulative and noncumulative trends holding after mean-centering. Table 2 confirms this quantitatively using point-biserial correlation: CI is negatively correlated with correctness after mean-centering. Figure 22 and Table 2 also show that, without mean-centering, the trend does not hold, demonstrating the importance of the cluster-based mean-centering procedure from Section 2.2, "Accounting for multiscale structure." Figure 23 further motivates cluster centering by showing that the raw aggregate trend is obscured by group-level structure, even though strong within-group trends remain. Figure 24 shows that CI is predictive within each cluster-pair setting, motivating cluster mean-centering for global comparison.

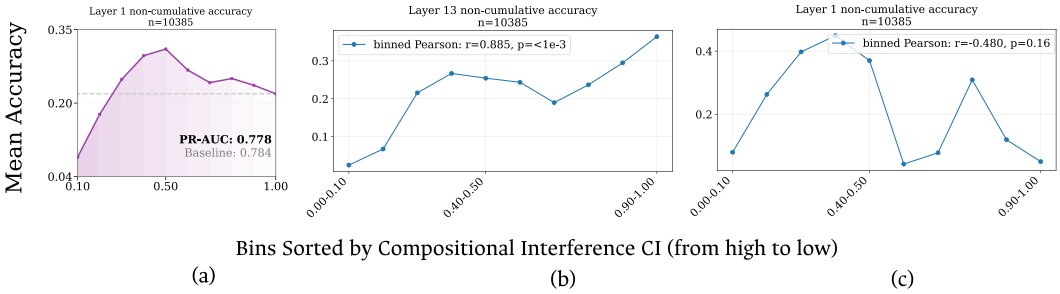


Figure 22: Multihop Dataset. (a) Cumulative plots when we extract the atomic representations for concepts without mean-centering. (b) Non-cumulative plots when we extract the atomic representations for concepts with mean-centering. (c) Non-cumulative plots when we extract the atomic representations for concepts without mean-centering.

Setting	r_{pb}
With mean-centering	-0.210 $(p < 0.01)$
Without mean-centering	0.178 $(p < 0.01)$

Table 2: Point-biserial correlations between CI and binary correctness (before and after mean-centering). With mean-centering, CI is negatively correlated with correctness, while without mean-centering, the sign reverses.

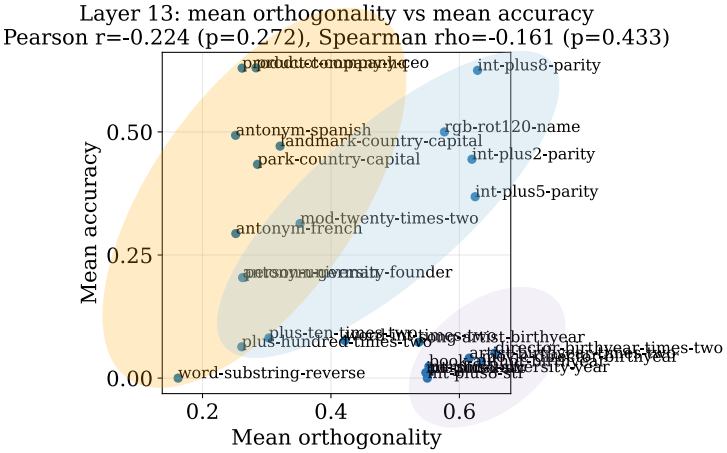


Figure 23: When we average $1 - \mathrm{CI}$ over all examples in each dataset and compare this dataset-level score against mean dataset accuracy, the overall correlation is weak and not statistically significant. This is consistent with the cumulative curve without mean-centering, where the PR-AUC signal is also weak (Figure 22). However, the same plot reveals strong within-group correlations: within each group, $1 - \mathrm{CI}$ is predictive of model accuracy. This suggests that CI does capture interference and failure likelihood, but the signal is obscured when aggregating across groups. This observation motivates our mean-centering procedure. The weak aggregate trend is driven by discrete orthogonality bands induced by group-level background structure, rather than by the absence of a CI signal. Mean-centering compensates for these group-level offsets, allowing the within-group relationship between CI and model failure to become visible.

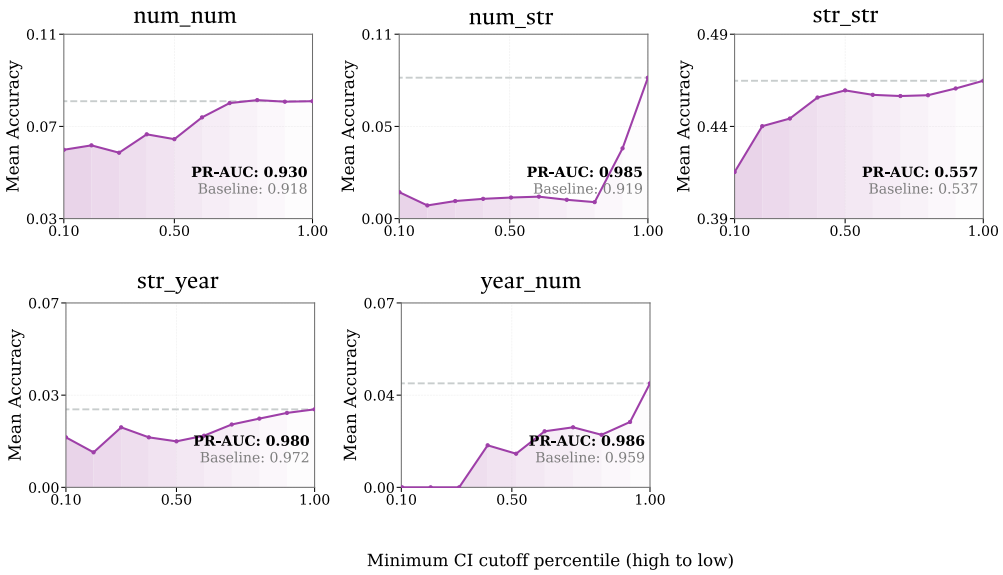


Figure 24: As discussed in Section 2.2, model representations can cluster by domain-level structure that is not directly relevant to compositional interference. When we break examples into within-cluster-pair settings, the trends remain strong within each setting, indicating that CI is still predictive of interference and model failure. However, the absolute CI value ranges are shifted across cluster pairs, making raw CI values difficult to compare globally. This motivates our mean-centering procedure: by adjusting for these cluster-level offsets, we can recover a comparable CI signal across different representation clusters.

G Multilingual Fact-Recall Additional Results

We expand on the experimental details for the multilingual fact-recall setting from Section 4. Table 3 provides the mapping from language abbreviations to full language names. Figure 26 shows PR-AUC curves across languages, showing that CI provides a predictive signal for multilingual fact-recall errors. Figure 25 reports the corresponding PR-AUC values and baselines for each language. Figure 27 shows the noncumulative trend, indicating that examples with higher CI are more error-prone across languages.

Abbreviation	Language
`nl`	Dutch
`ru`	Russian
`fr`	French
`es`	Spanish
`zh`	Chinese
`hu`	Hungarian
`uk`	Ukrainian
`vi`	Vietnamese
`ja`	Japanese
`ko`	Korean

Table 3: Language abbreviations and their corresponding full names.

Lang	PR-AUC	Baseline
nl	0.42	0.22
ru	0.50	0.33
fr	0.49	0.35
es	0.51	0.33
zh	0.65	0.59
hu	0.66	0.54
uk	0.67	0.60
vi	0.72	0.63
ja	0.70	0.61
ko	0.77	0.71

Figure 25: PR-AUC values for different languages.

Lang.	r_{pb}
es	-0.261 ($p < 0.01$)
fr	-0.126 ($p < 0.01$)
hu	-0.186 ($p < 0.01$)
ja	-0.267 ($p < 0.01$)
ko	-0.088 ($p < 0.01$)
nl	-0.288 ($p < 0.01$)
ru	-0.125 ($p < 0.01$)
uk	-0.181 ($p < 0.01$)
vi	-0.191 ($p < 0.01$)
zh	-0.081 ($p < 0.01$)

Table 4: Point-biserial correlations for multilingual fact-recall. Each row reports r_{pb}, the point-biserial correlation between CI and binary correctness. All reported correlations are significant at $p < 0.01$.

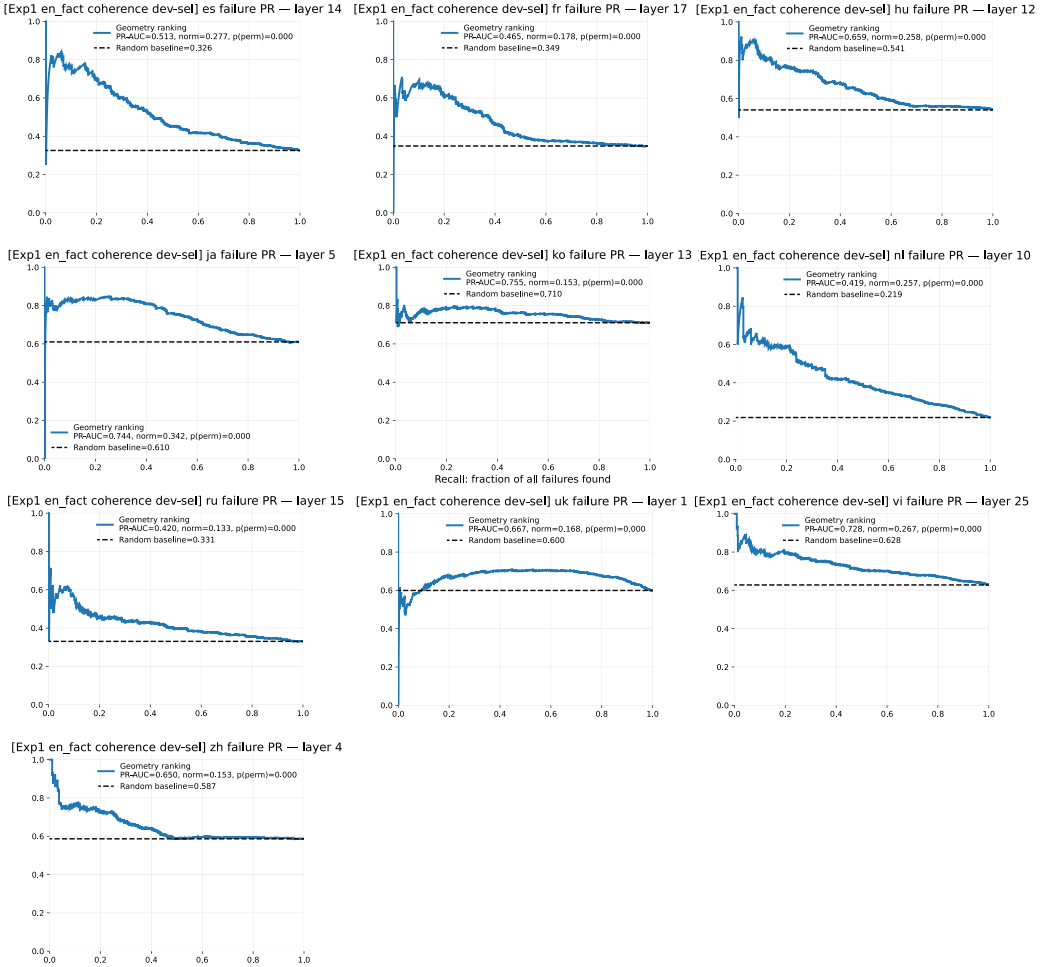


Figure 26: PR-AUC plots for the multilingual factual recall dataset for different languages.

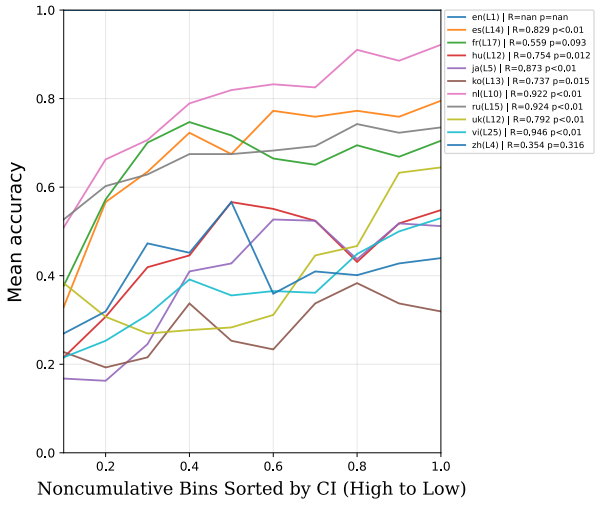


Figure 27: Noncumulative plot with x-axis sorted from high to low CI on the multilingual factual recall dataset. The trend is monotonically increasing across languages.

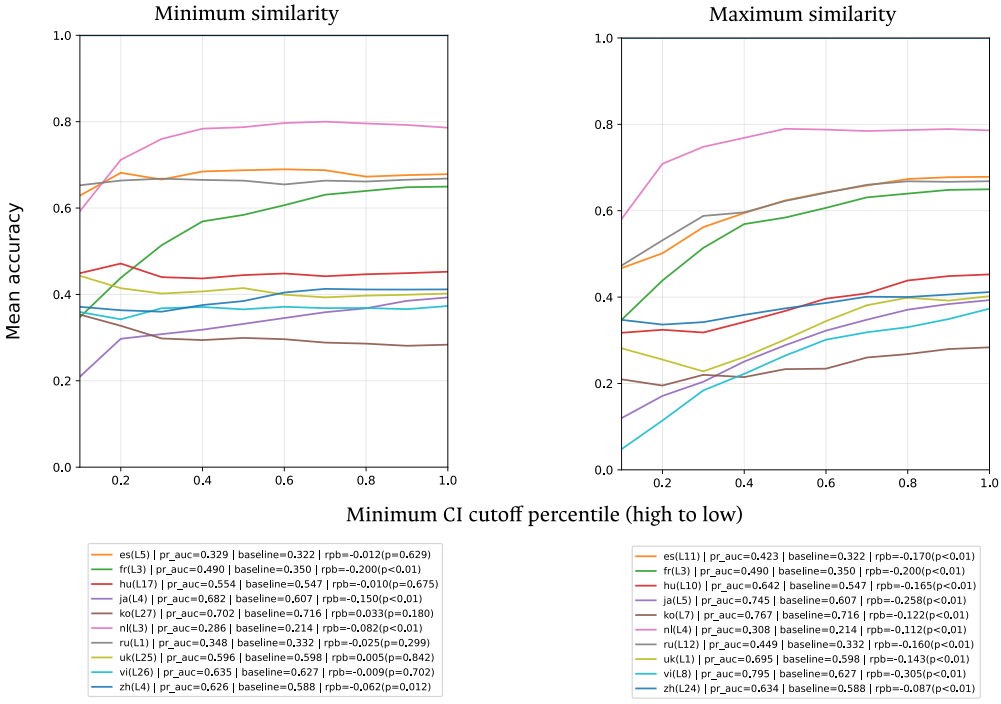


Figure 28: We also evaluate alternative interference metrics, as described in Appendix B. We omit the mean-similarity metric because, in this setting, it is mathematically equivalent to CI. Since the subspace bases are orthonormal, averaging signed basis-direction similarities equals zero, so it is the same quantity as captured by CI.

H Coarse-grained concepts

We expand on the additional experimental details for the coarse-grained concept analysis from Section 4. For multihop QA, Figure 29 shows that subspace-level CI also predicts the overall difficulty of broader task categories; it further reports results with alternative interference metrics from Section B, showing that the predictive trend is generally robust across metric choices. For multilingual fact-recall, Figure 30 and Figure 31 show that subspace-level CI also predicts coarse-grained task difficulty across languages. Table 5 reports per-language correlations for alternative subspace-level interference metrics. Figure 32 shows that when all language-topic pairs are pooled together, the overall CI–accuracy correlation is only weakly negative, likely because language-specific factors shift baseline difficulty and partially obscure the stronger within-language trends.

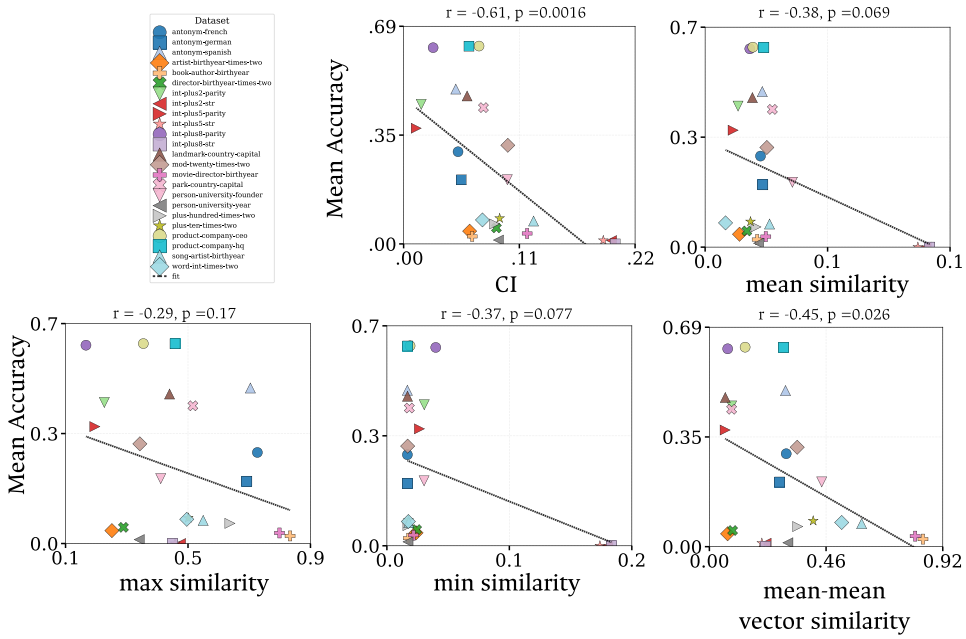


Figure 29: We define coarse task-level subspaces and use compositional interference between subspaces to predict the general difficulty of broadly described datasets. In multilingual fact-recall, CI is negatively correlated with mean accuracy across languages. We also evaluate alternative interference metrics, as described in Appendix B. Other metrics all show moderate negative correlation, which strengthens the point that the central signal comes from angular similarity among concept representations, while the derived cumulative coherence bound provides the highest predictive power. We further test a task-vector variant, where examples within the same coarse category are averaged to form a mean vector representation, following the intuition of task vectors [51]. Similarity between these task vectors also acts as a moderate negative predictor of accuracy.

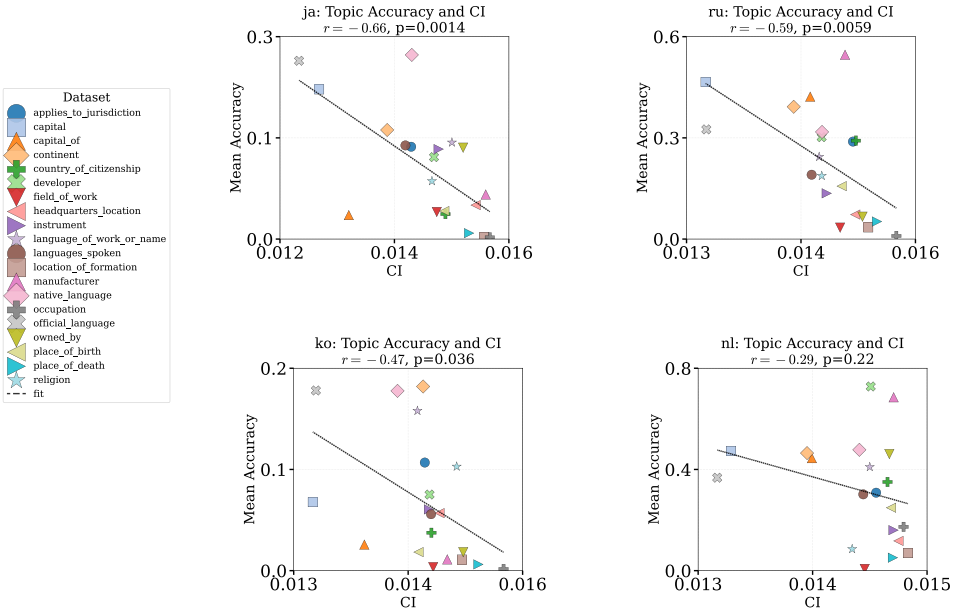


Figure 30: We define coarse task-level subspaces and use compositional interference between subspaces to predict the general difficulty of a broadly-described dataset. Correlation between mean accuracy across languages in the multilingual factual recall dataset and CI; this figure includes languages Japanese, Russian, Korean, and Dutch.

Lang.	CI	**Min. Sim.**	**Mean Sim.**	**Max Sim.**
nl	-0.29 ($p = 0.215$)	-0.54 ($p = 0.013$)	-0.43 ($p = 0.056$)	-0.40 ($p = 0.078$)
ru	-0.59 ($p = 0.006$)	-0.61 ($p = 0.005$)	-0.54 ($p = 0.014$)	-0.22 ($p = 0.353$)
fr	-0.34 ($p = 0.139$)	-0.46 ($p = 0.044$)	-0.35 ($p = 0.134$)	-0.37 ($p = 0.107$)
es	-0.15 ($p = 0.536$)	-0.42 ($p = 0.066$)	-0.37 ($p = 0.111$)	-0.44 ($p = 0.053$)
zh	-0.30 ($p = 0.192$)	-0.50 ($p = 0.026$)	-0.46 ($p = 0.042$)	-0.20 ($p = 0.395$)
hu	-0.18 ($p = 0.458$)	-0.48 ($p = 0.033$)	-0.38 ($p = 0.101$)	-0.32 ($p = 0.163$)
uk	-0.44 ($p = 0.052$)	-0.52 ($p = 0.020$)	-0.39 ($p = 0.086$)	-0.26 ($p = 0.263$)
vi	-0.06 ($p = 0.812$)	-0.33 ($p = 0.154$)	-0.26 ($p = 0.273$)	-0.42 ($p = 0.065$)
ja	-0.66 ($p = 0.001$)	-0.56 ($p = 0.011$)	-0.70 ($p < 10^{-3}$)	-0.26 ($p = 0.272$)
ko	-0.47 ($p = 0.036$)	-0.61 ($p = 0.005$)	-0.69 ($p < 10^{-3}$)	-0.24 ($p = 0.306$)
en	-0.46 ($p = 0.042$)	-0.67 ($p = 0.001$)	-0.57 ($p = 0.008$)	-0.45 ($p = 0.046$)

Table 5: Per-language Pearson correlations between task accuracy and subspace-level interference values computed with different metrics. Each cell reports r, with the corresponding p-value in parentheses. Correlations are computed over $n = 20$ task groups per language.

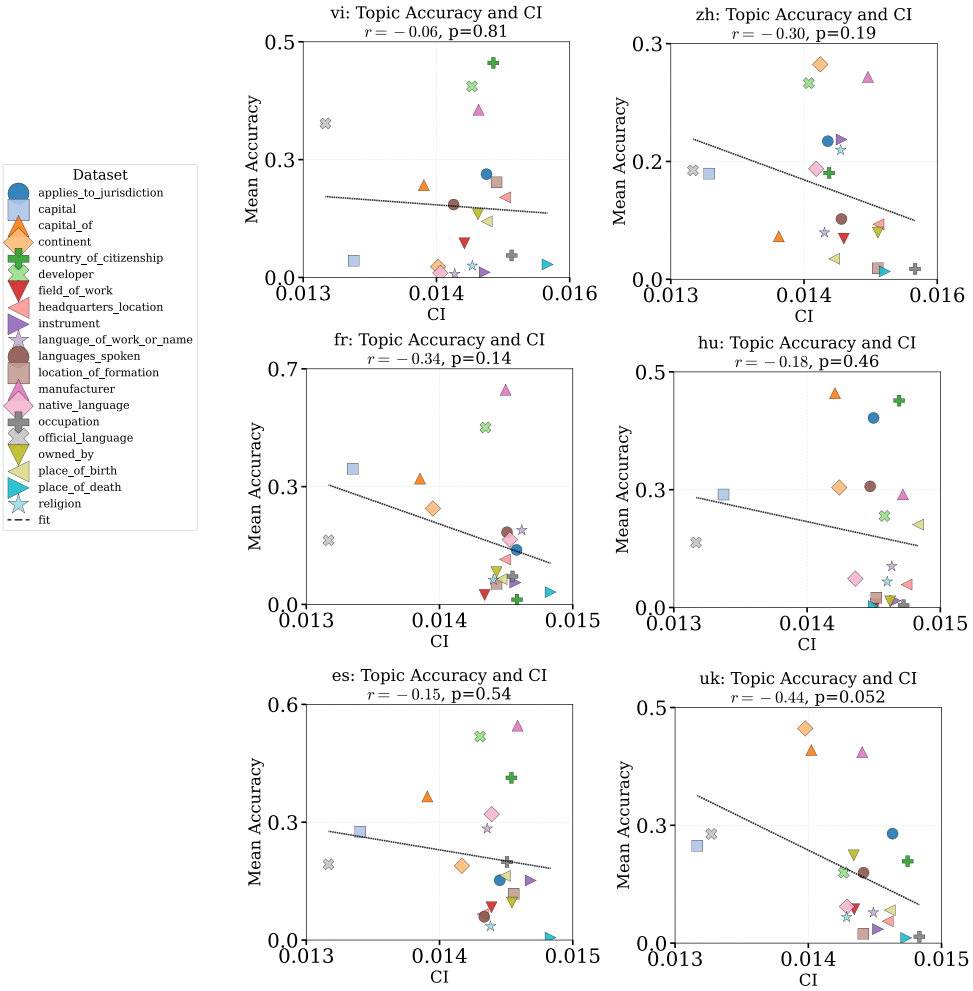


Figure 31: We define coarse task-level subspaces and use compositional interference between subspaces to predict the general difficulty of a broadly-described dataset. Correlation between mean accuracy across languages in the multilingual factual recall dataset and CI; this figure includes languages Vietnamese, Chinese, French, Hungarian, Spanish, and Ukrainian.

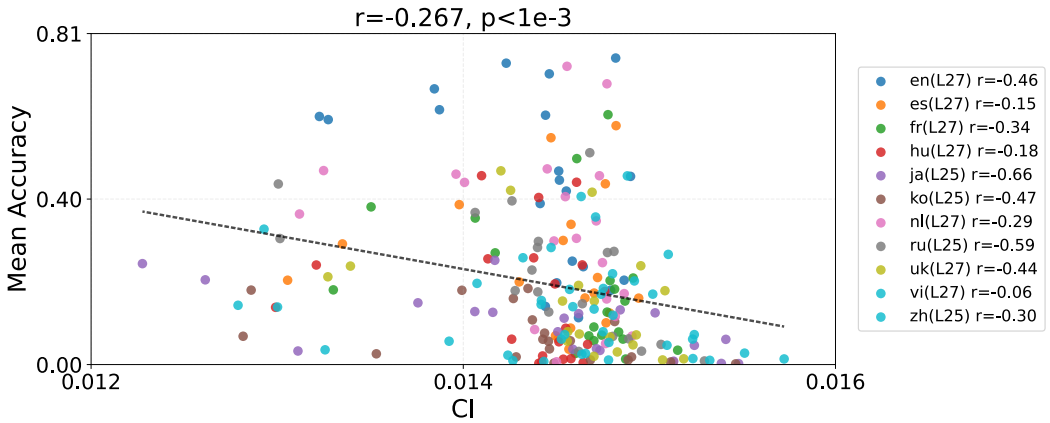


Figure 32: Coarse task-level correlation between Accuracy and CI for multilingual fact recall when all language-topic pairs are pooled together. Overall, the pooled correlation between CI and accuracy is weakly negative, even though the within-language trends remain moderate to strong. This suggests that cross-language comparisons are affected by language-specific factors, such as resource level, orthographic similarity to English, and transfer from English-language representations. These factors can shift the baseline difficulty of each language and partially obscure the within-language relationship between CI and model accuracy.

I Code and Compute

We will release the code upon the paper decision. For the SCAN experiments in Section 3, each toy model was trained on a single NVIDIA GeForce RTX 3090 GPU. For the real-LLM experiments in Section 4, all runs on Llama are conducted on a single 3090 GPU as well.