Metric Choice Determines Semantic Geometry in LLM Hidden States: A measurement Study of Hidden-State Geometry

Rui Gao [1] **Velibor Bojković** [1 *] **Hilal Mohammad Hilal AlQuabeh** [1 *]

Abstract

Large language models (LLMs) can be viewed as functions mapping discrete prompts to continuous hidden representations. Recent work shows that decoder-only transformers are almost surely injective, implying that distinct prompts are not collapsed into identical hidden states. Building on this functional perspective, we study whether LLM representations also exhibit expansion: the layerwise growth of distances between prompt representations. We empirically measure pairwise hidden-state distances across transformer depth and find that raw Euclidean (L2) distances generally increase with depth, indicating broad expansion in representation space. Although expansion appears across meaningful, unrelated, and random-token prompt families, these families are not geometrically identical. Raw L2 distance primarily reflects shared norm growth across layers, making expansion appear largely independent of semantic meaning or lexical coherence. However, when representations are compared using centered cosine distance, semantic and topic-level distinctions become visible. These results separate two geometric phenomena in LLMs: intrinsic expansion, which appears to be a general property of hidden-state evolution, and semantic geometry, which becomes visible only under appropriate measurement protocols.

1. Introduction

Large language models (LLMs) can be viewed as functions from discrete prompts to continuous hidden representations (Vaswani et al., 2017; Radford et al., 2019). Recent work strengthens this view by showing that decoder-only transformers are almost surely injective under standard assumptions: distinct prompts map to distinct hidden states (Nikolaou et al., 2025). Injectivity rules out exact prompt collisions, but opens an important metric question: **how does prompt geometry evolve across layers?**

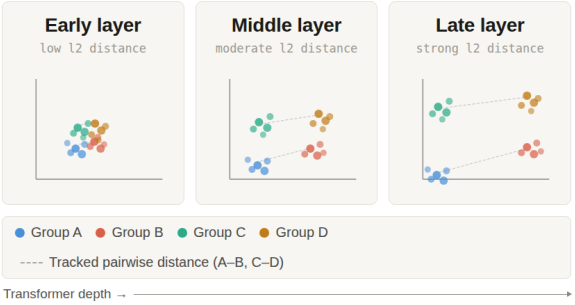

Figure 1. Schematic view of hidden-state geometry and raw L2 distance growth across layers.

In this work, we study one aspect of this question: empirical, layerwise *expansion*. Given a prompt set, we represent each prompt by its hidden state at each layer and track raw norms and pairwise distances across depth. Expansion here means observed growth of prompt-pair distances in hidden-state space, not a global mathematical proof of the model.

This matters for interpretability. If expansion reflects semantic processing, then paraphrases, same-topic prompts, unrelated prompts, and nonsensical prompts should have different distance-growth profiles. If they expand similarly, raw expansion is better understood as an intrinsic geometric property rather than direct evidence of semantic separation.

Our experiments support the latter view for raw Euclidean geometry. Raw hidden-state norms and pairwise L2 distances generally increase with depth, giving empirical evidence of broad expansion. Yet this growth is broad: meaningful prompts, unrelated prompts, and unigram-random prompts constructed from the same lexical distribution show similar patterns. Thus, raw L2 expansion is real, but largely insensitive to semantic meaning and lexical coherence.

This motivates a second question: whether prompt-space geometry contains semantic structure that raw distance fails to expose. We find that it does. Raw Euclidean distance is dominated by norm growth and shared components, and

[*]Equal contribution [1]MBZUAI. Correspondence to: < >.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

therefore does not reliably separate related from unrelated prompt pairs under controlled comparisons. When groups are centered with a shared centroid and then compared using cosine distance, sentence-, topic-, and category-level distinctions become visible. Lexical word-level evidence is weaker and treated as suggestive. Centering removes common directions, while cosine normalization reduces the effect of layerwise norm growth.

Overall, our results distinguish two geometric phenomena in LLMs. The first is *broad empirical expansion*: pairwise distances tend to grow across layers largely independently of prompt meaning. The second is *semantic geometry*: structure that reflects topic or meaning becomes recoverable only under appropriate measurement choices. This suggests that prompt-space geometry can support interpretability, but only when treated as a controlled measurement problem rather than a direct proxy for meaning.

We test three falsifiable hypotheses about prompt-level hidden-state geometry: **H1**, raw hidden-state norms and pairwise L2 distances grow with depth across prompt families; **H2**, this growth is not sufficient evidence of semantic separation because it also appears for incoherent or unrelated controls; and **H3**, sentence-, topic-, and category-level structure becomes more visible after shared-centroid centered-cosine comparison.

Our goal is not to identify a causal circuit or mechanism for generation, but to determine which geometric measurements can and cannot support semantic claims about hidden states. The results support these hypotheses, but they do not show that the centered-cosine geometry is causally used by the model during generation.

2. Related Work

LLMs as functions and representation geometry. Recent work studies transformer language models as functions from discrete sequences to hidden representations, including the injectivity result of Nikolaou et al. (2025). Complementary work analyzes how transformer geometry evolves across layers, including syntactic and semantic subspaces, intrinsic dimension, neighborhood structure, and representational separation (Coenen et al., 2019; Valeriani et al., 2023). We build on these perspectives by asking whether layerwise expansion is merely a generic geometric effect or whether it reflects semantic structure.

Metrics, anisotropy, and interpretability. Transformer representations are known to be anisotropic, and raw similarity or distance measures can be distorted by norm growth, directional concentration, rogue dimensions, and common components (Ethayarajh, 2019; Gao et al., 2019; Li et al., 2020; Godey et al., 2024; Timkey and van Schijndel, 2021). Post-processing methods such as mean removal, common-component removal, standardization, or whitening have therefore been used to improve geometric comparisons in embedding spaces (Mu and Viswanath, 2018; Li et al., 2020; Timkey and van Schijndel, 2021). Probing studies further show that linguistic and semantic information is encoded in hidden states (Hewitt and Manning, 2019; Tenney et al., 2019; Coenen et al., 2019), but often assume that geometric similarity is meaningful. In contrast, we explicitly test when geometric comparisons recover semantic relationships and when they are dominated by representation dynamics.

3. Methodology

We compare several prompt families across decoder-only models, extraction modes, and distance metrics. Full construction rules, metric definitions, aggregation details, and implementation settings are given in Appendix A, Appendix A.4, and Appendix A.5. The corresponding code, prompts, and results are available in the anonymous repository https://anonymous.4open.science/r/LLM_Expansion-A26E/.

Prompt Suite Design We use two prompt suites. The first is a small MIND diagnostic suite that compares meaningful news prompts with length-matched unigram-random controls. The random controls are sampled from the same empirical word pool as the meaningful prompts, but their word order and coherence are destroyed. This suite tests whether raw expansion depends on semantic structure.

The second is a broader pairwise suite drawn from STS-B, AG News, MIND, and SimLex-999 (Cer et al., 2017; Zhang et al., 2015; Wu et al., 2020; Hill et al., 2015). It contains prompt pairs designed to compare related and unrelated examples. This suite tests whether geometric metrics distinguish semantic relatedness once all prompts are meaningful. The full prompt inventory, dataset-specific construction rules, random-control definitions, and pair surfaces are reported in Appendix A.

Hidden-State Extraction For each prompt and layer, we extract one vector representation. Our primary representation is `mean_pool`, the average hidden state over all non-padding tokens. We also report `last_token`, the last token hidden state as a baseline. Comparing these two modes checks whether the observed geometry is a prompt-level effect or is specific to the final token.

Metrics and Centering We separate metrics according to their role. Raw norm and raw L2 distance measure representation-scale growth and are used to diagnose empirical expansion. Family-centered norm measures how far prompts spread around their family centroid. For pair-

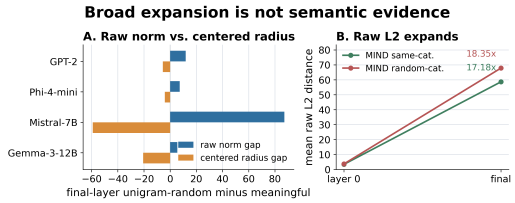

Figure 2. Raw expansion increases distances uniformly across inputs. **Left** Difference between distances between "non-sensical" and "meaningful" prompts. **Right** The total expansion factor at the penultimate layer.

wise centered-cosine results, the primary metric subtracts a centroid computed over both groups being compared, then computes cosine distance between the centered vectors.

This separation is important because raw Euclidean geometry can be dominated by norm growth and common representation components (Mu and Viswanath, 2018; Timkey and van Schijndel, 2021; Li et al., 2020). Centered cosine is therefore used to test whether meaningful relational structure remains after these global effects are reduced. Full metric definitions and centroid conventions are given in Appendix A.4.

Experimental Setup We evaluate four decoder-only checkpoints: GPT-2, Phi-4-mini, Mistral-7B, and Gemma-3-12B (Radford et al., 2019; Jiang et al., 2023; Microsoft, 2025; Gemma Team, 2025). For each model, we extract hidden states across layers. Designed prompt families are summarized over their predefined pairs, while random families are summarized over all eligible unordered pairs.

The main results report aggregate layerwise behavior across models, prompt families, extraction modes, and metrics. Implementation details, including model checkpoints, truncation, aggregation rules, and visualization conventions, are provided in Appendix A.5 and Appendix A.7.

4. Results: From Expansion to Sentence, Topic, and Category Structure

We first show that decoder-only LLMs exhibit empirical expansion: hidden-state norms and pairwise prompt distances generally grow with depth. This expansion appears broadly across all designed prompts. Then, we show that raw L2 distance does not reliably separate sentence-, topic-, or category-level structure under controlled comparisons. Finally, we show that centered cosine distance recovers these structures that raw expansion obscures. Additional experimental summaries and plots supporting these results are provided in Appendix A.7, with the main final-layer values summarized in Table 6.

4.1. Expansion Is Broad Across Prompt Types

Raw expansion is broad. We use expansion operationally: prompt representations are expanding when their hidden-state norms or pairwise distances increase across layers. Under this definition, expansion appears consistently across all four model choices.

Table 1. Raw L2 expands for both related and control MIND pairs.

Contrast	Layer 0 raw L2	Final raw L2	Final / Layer 0
Same-subcategory	3.4552	59.3685	17.18×
Subcategory-disjoint	3.7333	68.5107	18.35×

We tested whether expansion depends on sentence coherence. The diagnostic compares 32 meaningful MIND news prompts with 32 length-matched unigram-random controls sampled from the same empirical word pool. The controls preserve the lexical source in expectation while destroying word order and sentence coherence. If expansion were itself a semantic signal, meaningful and unigram-random prompts should exhibit clearly different expansion profiles.

This is not what we observe. Both families show substantial raw-norm growth from the embedding layer to the final layer. Median raw-norm growth is 31.49× for meaningful prompts and 33.32× for unigram-random prompts, indicating no consistent separation between the two conditions at the level of raw hidden-state magnitude.

This prompt-level result establishes the main control: raw expansion does not differentiate coherent prompts from unigram-random controls, even when both are drawn from the same empirical word distribution. We therefore treat expansion as a background geometric effect, reflecting a general tendency of decoder-only language models to increase hidden-state norms regardless of linguistic coherence.

4.2. Raw L2 Does Not Reliably Separate Related and Control Pairs

Raw L2 is not a reliable semantic metric. In addition to failing to distinguish coherent from randomized prompts, raw L2 distance also does not reliably separate semantic relationships. Layerwise pairwise distances show that raw Euclidean distance is dominated by overall expansion rather than meaningful structure.

In the AG News mean-pool control, same-topic and cross-topic article pairs both exhibit large raw L2 distances by the final layer. At layer 0, their distances are nearly identical: 4.1259 for same-topic pairs and 4.1045 for cross-topic pairs. By the final layer, both have expanded substantially, reaching 70.4711 and 77.7148, respectively. Although cross-topic pairs are somewhat farther apart, same-topic pairs undergo a comparable 17.08× expansion, indicating that distance growth is largely driven by global hidden-state displacement rather than a clear topic-separation signal.

These distance results show that Euclidean distance is useful as a diagnostic of representation-scale growth, but not as standalone evidence for sentence-, topic-, or category-level

separation. The semantic comparison requires a metric that removes shared global structure and compares normalized directions, which motivates the centered cosine results in the next section.

4.3. Centered Cosine Shows Consistent Sentence, Topic, and Category Structure

Centered cosine provides the main controlled separation result. For each controlled comparison, we subtract a centroid computed over the union of the related and controlled prompt groups, then compare angular relationships.

In our full four-checkpoint mean-pooled pass, controlled alternatives are farther than related conditions for the strongest sentence, topic, and category contrasts. The average final-layer control-minus-related gap is 0.2730 for STS-B low- versus high-similarity sentence pairs, 0.1818 for AG News cross-topic versus same-topic article pairs, and 0.2254 for MIND random-category versus same-category news pairs. These effects are not driven by a single checkpoint: all four models have positive final-layer gaps for STS-B (0.2301–0.3081), AG News (0.1325–0.2190), and MIND (0.1630–0.2609). These positive gaps show that centered cosine reveals structure that raw L2 expansion fails to distinguish within this benchmark. The same STS-B, AG News, and MIND gaps remain positive under both family and suite centering, indicating that the result is not driven by a family-specific centering choice (Table 8; Figure 13).

We quantify this result as a consistency audit rather than as a formal population-level significance claim. Treating the four retained checkpoints as the replication unit, the three main contrasts are positive in 4/4 models under family, contrast, and suite centering. A one-sided sign test over four checkpoint signs gives $p = 0.0625$ for each 4/4 row, which is above the conventional $p < 0.05$ threshold. This should therefore be read as directional model-consistency evidence, not as a high-powered test over language models. As descriptive pair-level support, we also bootstrap final-layer family-centered `mean_pool` distances and run deterministic Monte Carlo label permutations over pair distances. The pair-level gap is positive in 4/4 models for STS-B, AG News, and MIND; bootstrap intervals are positive in 4/4 models for STS-B and MIND and in 3/4 models for AG News. Because random-cloud pair distances share prompt endpoints, these pair-level summaries are descriptive audit trails rather than independent-sample significance tests.

We exclude the lexical SimLex contrast from the main claim because it is weaker and more model-variable than the sentence-, topic-, and category results. In the full four-checkpoint `mean_pool` pass, the SimLex final-layer gap is positive on average but one checkpoint shows the opposite sign. We therefore report SimLex only as an exploratory lexical control in the appendix. The strongest centered-cosine pattern in the main text is for sentence-level similarity, topic structure, and category separation, with the scope limited to the retained checkpoints and prompt suite.

Table 2. Final-layer centered-cosine consistency audit. Contrast-gap values use contrast centering in the full four-checkpoint `mean_pool` pass. Sign-test values are one-sided over checkpoint signs and should not be read as formal population-level significance tests. Pair-bootstrap counts report how many model-specific descriptive bootstrap intervals are strictly positive.

Contrast	Contrast gap	Family signs	Suite signs	Pair bootstrap
STS-B low-high	0.2730	4/4 ($p = 0.0625$)	4/4 ($p = 0.0625$)	4/4
AG News cross-same	0.1818	4/4 ($p = 0.0625$)	4/4 ($p = 0.0625$)	3/4
MIND random-same	0.2254	4/4 ($p = 0.0625$)	4/4 ($p = 0.0625$)	4/4
SimLex random-correlated	0.1127	3/4 ($p = 0.3125$)	2/4 ($p = 0.6875$)	2/4

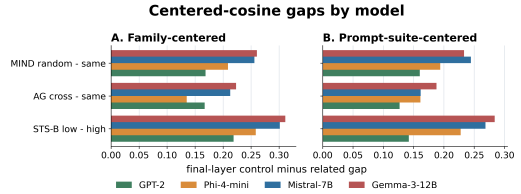

Figure 3. Centered cosine results for the main sentence, topic and category comparisons. Positive gaps mean the controlled alternative is farther than the related condition after centering both groups together. The lexical SimLex control is reported only in the appendix.

Together, these results separate broad empirical expansion from semantic geometry. Raw norms and raw pairwise L2 distances show that prompt representations expand broadly across layers, including for nonsensical and unrelated controls. Centered cosine shows that sentence-, topic-, and category-level relationships are more consistently expressed after controlling for shared representation components and norm growth.

5. Conclusion

This work treats prompt-space geometry as a measurement problem rather than as a direct readout of semantic meaning. Across multiple decoder-only language models, we find that hidden-state representations undergo broad expansion with depth: raw norms and pairwise distances tend to increase for coherent prompts as well as controlled alternatives. This suggests that expansion is a consistent empirical feature of the hidden-state measurements we study, but not sufficient evidence of semantic organization on its own.

At the same time, semantic structure becomes visible when geometry is measured under a controlled and consistent protocol. When choices about representation, centering, pair construction, and metric are fixed, related sentence, topic, and category prompts remain consistently closer than appropriate alternatives in the retained comparisons. The key result is therefore not that hidden-state geometry directly encodes meaning, but that meaningful structure can be exposed by careful measurement.

Overall, our findings separate *broad empirical expansion* from *semantic geometry*. They suggest that analyses of language-model representations should be interpreted as protocol-dependent measurements of representational structure, rather than as direct maps of meaning.

Limitations and future work This study is only measurement-focused. We do not prove that decoder-only transformers are globally expanding functions, nor do we show that the measured geometry is causally used during generation. Our results show empirical layerwise expansion and recoverable sentence-, topic-, and category-level structure under a specific centered-cosine protocol.

Another key limitation is that centering is itself a measurement choice. The centering-scope analysis reduces concern about family-specific artifacts because the prompt-suite gaps remain positive under family and suite centering. However, centering may still bias comparisons by determining which shared directions are removed. The SimLex lexical result is therefore treated as suggestive, and further investigation is needed.

Future work should test whether these geometric signals are functionally used by the model. This can be done through hidden-state interventions such as norm scaling, directional perturbations, noise injection, or rotations (Meng et al., 2022; Wang et al., 2023; Turner et al., 2023). If such interventions affect predictions, the centered geometry may reflect functionally relevant structure; if not, it may mainly reflect architectural or measurement effects.

References

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14, 2017. doi: 10.18653/v1/S17-2001.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65, 2019. doi: 10.18653/v1/D19-1006.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, 2024. doi: 10.18653/v1/2024.eacl-long.3.

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 (Long and Short Papers)*, pages 4129–4138, 2019. doi: 10.18653/v1/N19-1419.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015. doi: 10.1162/COLI_a_00237.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9130, 2020. doi: 10.18653/v1/2020.emnlp-main.733.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

Microsoft. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

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

Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. Language models are injective and hence invertible. *arXiv preprint arXiv:2510.15511*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

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*, pages 4593–4601, 2019. doi: 10.18653/v1/P19-1452.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, 2021. doi: 10.18653/v1/2021.emnlp-main.372.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, volume 36, pages 51234–51252, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 Small. In *International Conference on Learning Representations*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, 2020. doi: 10.18653/v1/2020.acl-main.331.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657, 2015.

A. Prompt Suite and Measurement Details

This appendix records the implementation-level details needed to reproduce the prompt construction and metric computations used in the paper. The goal is to make the methodology precise: raw geometric expansion is treated as a control phenomenon, while centered cosine distance is the primary metric for semantic and topic-structure claims.

Throughout the appendix, we distinguish between paper-facing descriptive names and implementation identifiers. The descriptive names are used in prose, figures, and tables; the implementation identifiers refer to the corresponding `pair_group` values stored in the prompt-suite files.

A.1. Prompt Suites

The main pairwise suite contains 512 prompts drawn from STS-B, AG News, MIND, and SimLex-999. Each prompt is stored as one JSONL row in `prompt_suites/external_prompt_suite_stsb_agnews_words_512.jsonl`. Designed families define explicit two-item pairs. Random-control families define prompt clouds and are summarized over all eligible unordered pairs.

Paper-facing name	Implementation identifier	Source	Prompts	Pair surface
STS-B high-similarity sentence pairs	`stsb_high_similarity`	STS-B	64	32 designed sentence pairs
STS-B low-similarity sentence pairs	`stsb_low_similarity`	STS-B	64	32 designed sentence pairs
AG News same-topic article pairs	`ag_news_same_topic`	AG News	64	32 designed article pairs
AG News cross-topic article pairs	`ag_news_cross_topic`	AG News	64	32 designed article pairs
AG News stratified random cloud	`ag_news_random_64_stratified`	AG News	64	2016 all unordered pairs
MIND same-subcategory news pairs	`mind_news_same_category`	MIND	64	32 designed news pairs
MIND subcategory-disjoint news cloud	`mind_news_random_64_category`	MIND	64	2016 subcategory-disjoint unordered pairs
SimLex high-similarity word pairs	`word_correlated`	SimLex-999	32	16 designed word pairs
SimLex low-similarity word cloud	`word_random`	SimLex-999	32	496 source-pair-disjoint unordered pairs

Table 3. Prompt families in the main 512-prompt pairwise suite. Descriptive names are used in the paper, while implementation identifiers correspond to stored `pair_group` values.

We also use a separate 64-prompt MIND diagnostic suite stored at `prompt_suites/mind_meaningful_vs_unigram_random_64.jsonl`. It contains two families: MIND meaningful news prompts (`mind_meaningful_32`) and MIND unigram-random controls (`mind_unigram_random_32`). This suite tests whether raw expansion depends on linguistic coherence and word order.

A.2. Meaningful-Versus-Unigram-Random MIND Control

The MIND meaningful-versus-random diagnostic uses 32 meaningful MIND prompts and 32 length-matched unigram-random controls.

The meaningful prompts are selected from the `mind_news_same_category` family in the 512-prompt suite. The selection rule takes one prompt per MIND subcategory after filtering for sentence-like prompts. A prompt must contain 10–90 extracted word tokens, have a digit-character fraction no greater than 0.04, and must not contain the filtered sports-score patterns used by the builder.

The unigram-random prompts are constructed as follows.

- **Length matching:** each random prompt has the same extracted word-token count as its matched meaningful prompt.
- **Sampling source:** words are sampled from the empirical unigram pool formed by the selected 32 meaningful prompts, not from each individual prompt and not from the full MIND source.
- **Replacement:** sampling is with replacement.
- **Random seed:** the suite builder uses seed `0`.

- **Tokenization for construction:** words are extracted with the regex `[A-Za-z][A-Za-z'-]*` and lowercased.
- **Casing and punctuation:** random prompts are lowercased word sequences followed by a final period. Original punctuation is otherwise removed, except apostrophes and hyphens inside extracted word tokens.
- **Numbers:** numeric tokens are not included in the random unigram pool under the word-regex rule.
- **Stopwords:** stopwords are not removed. If a stopword occurs in the selected meaningful-prompt pool, it can be sampled.
- **Construction matching:** every random prompt stores `matched_to` and `match_id` metadata linking it to one meaningful prompt for length matching. The norm diagnostic compares the two prompt families as families rather than treating these construction matches as pairwise semantic examples.
- **Disruption constraints:** a sampled random prompt is rejected if it exactly matches the original word list, shares any original trigram for prompts of length at least three, has maximum repeated-token fraction above 0.25, has multiset overlap with the original prompt above 0.35, or duplicates a previously generated random prompt.

Thus, the random controls preserve the selected suite's empirical unigram pool in expectation while destroying sentence order and coherence. They should not be described as preserving the exact corpus frequency distribution in every generated prompt.

A.3. Main Pair-Construction Rules

All main-suite construction uses seed `0`. Text is normalized by unescaping HTML, removing HTML tags and URLs, collapsing whitespace, and enforcing unique normalized prompt texts. Candidate prompts are additionally filtered to be no longer than 192 GPT-2 tokens during suite construction.

STS-B. The STS-B high-similarity family (`stsb_high_similarity`) uses sentence pairs with human similarity scores in $[4.0, 5.0]$. The STS-B low-similarity family (`stsb_low_similarity`) uses sentence pairs with scores in $[0.0, 1.5]$. Each family contains 32 designed pairs, giving 64 prompts per family. Each sentence in a candidate pair must contain 3–80 whitespace words.

AG News. AG News uses the four dataset labels: `business`, `sci_tech`, `sports`, and `world`. Article candidates are filtered to 6–120 whitespace words and at most 192 GPT-2 tokens.

The AG News same-topic family (`ag_news_same_topic`) contains 32 designed pairs sampled in label-balanced round-robin order, with both articles in each pair drawn from the same label.

The AG News cross-topic family (`ag_news_cross_topic`) contains 32 designed pairs sampled by cycling over the six unordered label combinations. Each pair contains articles from two different labels.

The AG News stratified random cloud (`ag_news_random_64_stratified`) contains 64 articles, exactly 16 from each AG News label. Its primary pair surface is all unordered pairs among the 64 prompts, giving $\binom{64}{2} = 2016$ pairs. When a within-label versus between-label split is reported, it is a post-hoc split of this same 2016-pair cloud: 480 within-label pairs and 1536 between-label pairs.

MIND. For MIND, the prompt text is the news title followed by the abstract when an abstract is present. Candidate prompts are filtered to 6–120 whitespace words and at most 192 GPT-2 tokens.

The JSONL field named `category` stores the MIND subcategory. The MIND top-level category is stored separately as `macro_domain`. Therefore, paper-facing descriptions use "subcategory" when referring to these MIND families.

The MIND same-subcategory news pairs (`mind_news_same_category`) contains 32 designed pairs. Each pair contains two articles from the same MIND subcategory.

The MIND subcategory-disjoint news cloud (`mind_news_random_64_category`) contains one article from each of 64 selected MIND subcategories. Its pair surface is subcategory-disjoint all-pairs. Because there is only one prompt per selected subcategory, this gives $\binom{64}{2} = 2016$ unordered pairs. These pairs are subcategory-disjoint, not necessarily top-level-category-disjoint.

SimLex-999 word controls. The SimLex high-similarity word pairs (`word_correlated`) uses SimLex-999 pairs with similarity score at least 7.5. It contains 16 designed pairs, and each prompt is a single word from the source pair.

The SimLex low-similarity word cloud (`word_random`) is built from SimLex-999 pairs with similarity score at most 2.0. For each selected low-similarity pair, one of the two words is sampled as a single-word prompt. The resulting family has 32 single-word prompts. Here, "source-pair-disjoint" means that two prompts do not come from the same source low-similarity pair. Since the selected source-pair labels are unique, the current family uses all $\binom{32}{2} = 496$ unordered pairs.

No prompt family uses an instruction template. STS-B prompts are plain sentences, AG News prompts are article text, MIND prompts are title/abstract text, and word-control prompts are single words.

A.4. Metric Definitions

Let p_i denote prompt i, let m denote a model, let e denote the extraction mode, and let ℓ denote the layer index. The analyzed vector is

$$h_\ell^{m,e}(p_i) \in \mathbb{R}^{d_m}.$$

When the model and extraction mode are clear from context, we write $h_\ell(p_i)$.

For a prompt family f with prompt set $\mathcal{P}_f$, the layerwise family centroid is

$$\mu_{f,\ell}^{m,e} = \frac{1}{|\mathcal{P}_f|} \sum_{p_i \in \mathcal{P}_f} h_\ell^{m,e}(p_i).$$

The centroid is computed separately for each model, layer, extraction mode, and prompt family. It is not shared across datasets or extraction modes. Thus, `mean_pool` and `last_token` have separate centroids.

For the main pairwise centered-cosine result, the centroid is computed over the two prompt groups being compared. For a contrast between a related prompt set $\mathcal{P}_r$ and a controlled prompt set $\mathcal{P}_c$, the contrast centroid is

$$\mu_{\mathrm{ctr},\ell}^{m,e} = \frac{1}{|\mathcal{P}_r \cup \mathcal{P}_c|} \sum_{p_i \in \mathcal{P}_r \cup \mathcal{P}_c} h_\ell^{m,e}(p_i).$$

The primary pairwise metric is centered cosine distance:

$$d_{\mathrm{ccos},\ell}^{m,e}(i,j) = 1 - \cos\left(h_\ell^{m,e}(p_i) - \mu_{\mathrm{ctr},\ell}^{m,e}, h_\ell^{m,e}(p_j) - \mu_{\mathrm{ctr},\ell}^{m,e}\right).$$

The centering-scope comparison in Figure 13 and Table 8 compares this contrast centroid with two alternatives. The family-centered baseline uses $\mu_{f,\ell}^{m,e}$, where f is the prompt family being summarized. For the AG News random cloud in the family-centered baseline, the within-label and between-label comparisons are both centered using the full `ag_news_random_64_stratified` family centroid, because they are splits of the same random-cloud surface.

The suite centroid uses all prompts in the rough probe suite,

$$\mu_{s,\ell}^{m,e} = \frac{1}{|\mathcal{P}_s|} \sum_{p_i \in \mathcal{P}_s} h_\ell^{m,e}(p_i).$$

The contrast centroid is the more conservative presentation because the related and controlled groups share the same centering surface. The family and suite centroids are retained as sensitivity checks.

The raw metrics are:

$$\mathrm{raw_norm}_\ell(i) = \|h_\ell(p_i)\|_2,$$
$$\mathrm{raw_L2}_\ell(i,j) = \|h_\ell(p_i) - h_\ell(p_j)\|_2,$$
$$\mathrm{raw_cosine}_\ell(i,j) = 1 - \cos\left(h_\ell(p_i), h_\ell(p_j)\right).$$

For per-prompt norm diagnostics, family-centered radius is defined as

$$\mathrm{family_centered_norm}_\ell(i) = \|h_\ell(p_i) - \mu_{f,\ell}\|_2.$$

The norm-diagnostic scripts also compute a `global_centered_norm` for the meaningful-versus-random suite by centering both included families together. This is a diagnostic comparison only; the main pairwise centered-cosine protocol uses the contrast centroid.

For pairwise L2, subtracting the same family centroid from both endpoints does not change the distance:

$$\|(h_i - \mu_f) - (h_j - \mu_f)\|_2 = \|h_i - h_j\|_2 .$$

Therefore, raw pairwise L2 and centered pairwise L2 are mathematically equivalent unless an additional normalization step is applied. This is why raw L2 expansion is treated as a background representation dynamic rather than as the semantic headline.

Metric or view	Unit	Main use	Interpretation boundary
Raw norm	prompt	Measures hidden-state magnitude growth across layers.	Useful as an expansion diagnostic; not evidence of semantic organization by itself.
Family-centered radius	prompt	Measures within-family dispersion after subtracting the family centroid.	Useful for the meaningful-versus-unigram-random diagnostic; not a pairwise semantic distance.
Raw L2	pair	Measures Euclidean pairwise displacement between hidden-state vectors.	Useful for showing distance growth; scale-sensitive and not a standalone semantic metric.
Centered-normalized L2	pair	Measures Euclidean distance after family centering and unit normalization.	A monotonic angular comparison related to centered cosine; should be interpreted as a controlled pairwise metric, not raw displacement.
Centered cosine	pair	Compares directions after subtracting a shared centroid.	Primary pairwise result uses the contrast centroid; family and suite centroids are reported as sensitivity checks.
Centering-scope probe	pair	Repeats final-layer centered-cosine gaps under family, contrast, and suite centroids.	Useful for checking centering artifacts; the current centered result is a four-checkpoint `mean_pool` pass.

Table 4. Metric roles and interpretation boundaries. The paper treats raw norm and raw L2 as expansion diagnostics, while sentence, topic, and category claims are based on controlled centered angular comparisons. Family and suite centering are reported as sensitivity analyses.

A.5. Aggregation Rules

Designed families are averaged over their predefined metadata pairs: 32 pairs for the STS-B and news designed families, and 16 pairs for the SimLex high-similarity word pairs.

Random-control clouds are averaged over all eligible unordered pairs. In the current suite, this means 2016 pairs for the AG News stratified random cloud, 2016 pairs for the MIND subcategory-disjoint news cloud, and 496 pairs for the SimLex low-similarity word cloud.

The extraction outputs store the mean, median, minimum, maximum, and pair count for each model, extraction mode, layer, family, pair surface, and metric. Unless a figure explicitly says otherwise, layerwise curves use the mean statistic. The core extraction pipeline reports mean, median, minimum, maximum, and pair count. The present figures do not include uncertainty bands; final-layer per-checkpoint values are reported separately to show model-level consistency. If later figures add uncertainty bands, the caption should state whether they are standard deviations, standard errors, or confidence intervals.

Layerwise distance and norm curves are absolute values by default. They are not normalized by layer 0 in the core outputs. Some summaries may additionally report final-to-layer-0 ratios for compact comparison, but those ratios are derived from the absolute layerwise outputs and should be labeled as ratios.

The consistency table in the main text is generated by `scripts/compute_statistical_evidence.py` from retained CSV artifacts under `out/pass_2026-04-29/` and `out/pass_2026-05-08/`. The script writes `out/pass_2026-05-08/statistical_evidence/`, including checkpoint-level sign tests, pair-level bootstrap intervals, deterministic Monte Carlo permutation checks, and a markdown brief. These outputs are intended as consistency and audit summaries. Pair-level resampling treats pair distances as the resampling unit; for random-control clouds this is descriptive because pairs share prompt endpoints. The checkpoint-level sign rows are the cleaner replication-unit summary, but with only four retained checkpoints they should not be interpreted as high-powered population-level tests.

For the MIND meaningful-versus-unigram-random norm diagnostic, the reported per-family quantities are per-prompt raw norm, family-centered radius, and global-centered radius. The summary table stores mean, median, population standard deviation, minimum, and maximum over prompts in each family at each layer.

A.6. Model and Hidden-State Extraction Details

The retained model set is listed in Table 5.

Short name	Hugging Face checkpoint	Model type in this paper
GPT-2	`openai-community/gpt2`	base decoder-only model
Phi-4-mini	`microsoft/Phi-4-mini-instruct`	instruction-tuned decoder-only model
Mistral-7B	`mistralai/Mistral-7B-v0.1`	base decoder-only model
Gemma-3-12B	`google/gemma-3-12b-pt`	pretrained decoder-only model

Table 5. Decoder-only checkpoints used in the experiments. Checkpoint citations: Radford et al. (2019); Jiang et al. (2023); Microsoft (2025); Gemma Team (2025).

Models are loaded through the Hugging Face causal-language-model auto class (Wolf et al., 2020) with `from_pretrained` and `torch_dtype="auto"`. No manual fp16, bf16, or fp32 override is imposed beyond that Hugging Face setting. Extracted prompt vectors are converted to float32 NumPy arrays before metric computation.

Hidden states are obtained from the standard Hugging Face forward pass with `output_hidden_states=True` and `use_cache=False`. No custom pre-normalization or post-normalization residual hook is used in the main results. Layer 0 is the embedding output. Layers 1 through N are the returned transformer hidden-state entries. The final analyzed hidden state is the model output hidden state as returned by Hugging Face before the language-model head. For architectures with an internal final normalization in the model forward pass, this returned final state includes the architecture's standard final normalization.

Two extraction modes are used:

- `mean_pool`: mean of all non-padding token hidden states. This is the primary prompt-level representation.
- `last_token`: final non-padding token hidden state. This is reported as a decoder-position baseline and is more exposed to final-token and output-head effects.

Tokenization uses each checkpoint's Hugging Face tokenizer with padding and truncation enabled. If the tokenizer has no pad token, the EOS token is used as the pad token when available; otherwise a `<|pad|>` token is added. Padding side is not manually overridden, so it follows the tokenizer default. Pooling uses the attention mask, so padding tokens do not contribute to `mean_pool`, and `last_token` selects the final non-padding token.

The main 512-prompt distance runs use `max_length=256` at extraction time. The suite builder also filters candidates to at most 192 GPT-2 tokens during prompt construction. The MIND meaningful-versus-unigram-random norm diagnostic uses `max_length=192` at extraction time. In all cases, prompts are plain text and are not wrapped in a question-answer or instruction template.

A.7. Additional Plot Layouts

This subsection provides the empirical audit trail for the headline comparisons in the main paper. Unless otherwise stated, layerwise plots use the mean over the relevant pair surface and aggregate over the four retained checkpoints. The horizontal axis is normalized layer depth so that checkpoints with different numbers of transformer blocks can be shown on a common scale. The centered cosine results use the full four-checkpoint `mean_pool` pass.

Result	Metric/view	Headline value	Reading
MIND meaningful vs. unigram-random	final raw norm	unigram-random higher in 4/4 models	Raw magnitude does not decrease for incoherent controls.
MIND meaningful vs. unigram-random	final family-centered radius	meaningful higher in 4/4 models	Coherent prompts separate after family centering.
AG News same-topic vs. cross-topic	raw L2, `mean_pool`	4.1259→70.4711 vs. 4.1045→77.7148	Raw L2 expands strongly for both same-topic and cross-topic article pairs.
MIND same-category vs. random-category	raw L2, `mean_pool`	17.18× vs. 18.35×	Raw L2 expansion appears for related and control pairs.
STS-B high- vs. low-similarity	centered cosine gap	0.2730, four-checkpoint `mean_pool`	Low-similarity sentence pairs are farther than high-similarity pairs.
AG News same-topic vs. cross-topic	centered cosine gap	0.1818, four-checkpoint `mean_pool`	Cross-topic article pairs are farther than same-topic pairs.
MIND same-category vs. random-category	centered cosine gap	0.2254, four-checkpoint `mean_pool`	Random-category pairs are farther than same-category pairs.
SimLex correlated vs. random words	centered cosine gap	0.1127, four-checkpoint `mean_pool`; 3/4 models positive	Exploratory lexical control; positive on average but not part of the main claim.

Table 6. Headline final-layer values supporting the main Results section. Arrows denote layer-0 to final-layer raw L2 values. Centered-cosine gaps are control-minus-related final-layer distances from the four-checkpoint `mean_pool` pass.

Contrast	GPT-2	Phi-4-mini	Mistral-7B	Gemma-3-12B	Mean
STS-B low minus high similarity	0.2301	0.2542	0.2996	0.3081	0.2730
AG News cross minus same topic	0.1710	0.1325	0.2046	0.2190	0.1818
MIND random minus same category	0.1630	0.2203	0.2572	0.2609	0.2254
SimLex random minus correlated words	0.0583	-0.0300	0.2209	0.2016	0.1127

Table 7. Per-checkpoint final-layer centered cosine gaps from the full four-checkpoint `mean_pool` pass. Positive values mean the controlled alternative is farther than the related condition after centering both compared groups with their shared contrast centroid. STS-B, AG News, and MIND are positive for all four checkpoints; SimLex is positive for three checkpoints and slightly negative for Phi-4-mini.

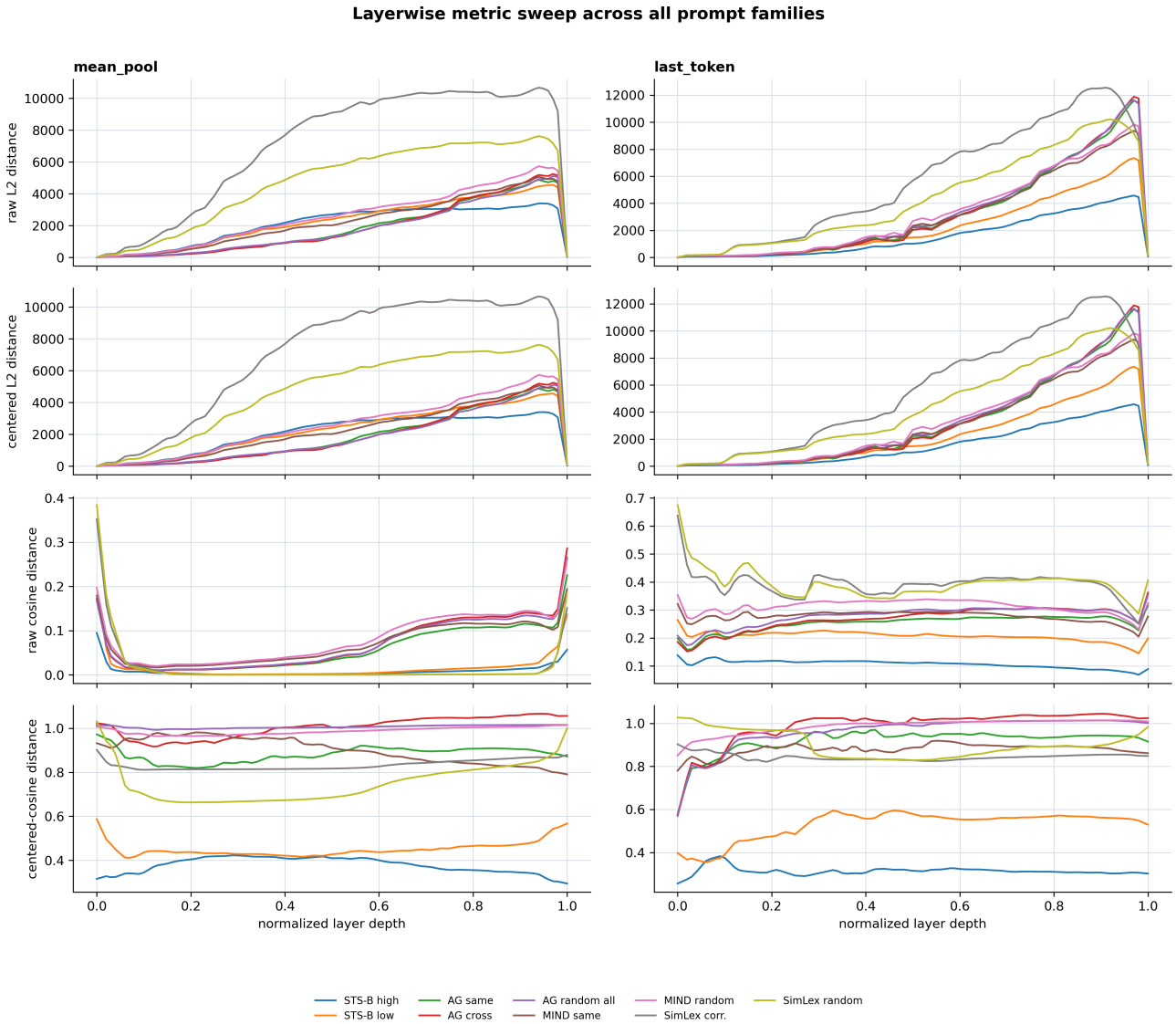

Figure 4. Layerwise metric sweep across all main prompt families and both extraction modes. Rows show raw L2, centered L2, raw cosine, and centered cosine distances; columns show `mean_pool` and `last_token`. The raw L2 and centered L2 rows overlap because subtracting one family centroid from both members of a pair does not change their Euclidean distance. The cosine rows expose the difference between raw anisotropic angular comparisons and centered angular comparisons.

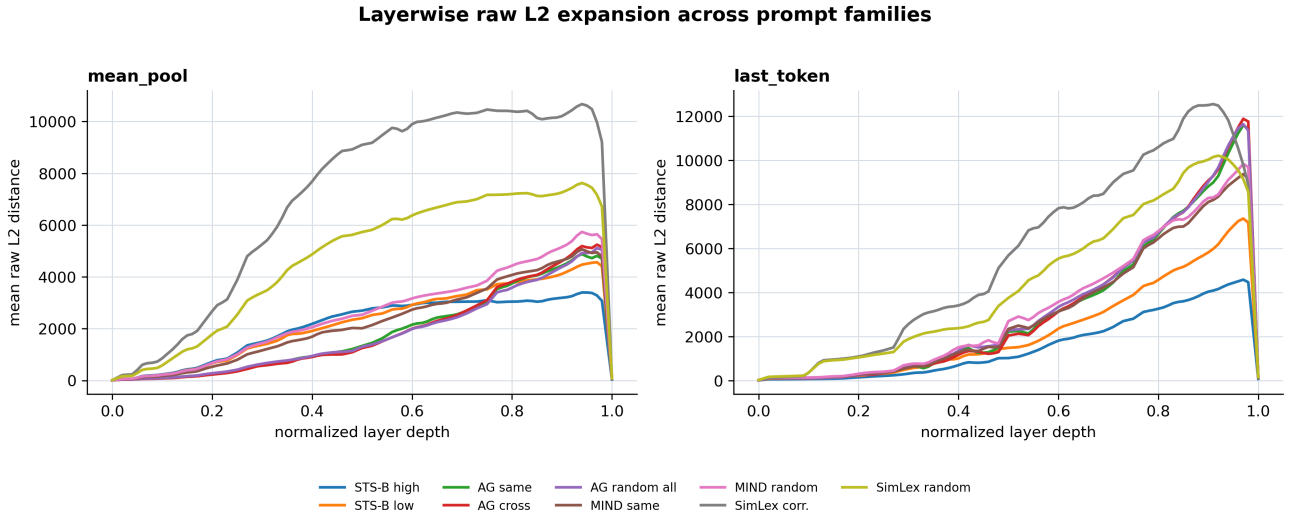

Figure 5. Layerwise raw L2 expansion across all main prompt families. Curves show four-checkpoint aggregate means for `mean_pool` and `last_token` extraction. The plot supports the interpretation of raw L2 as a broad displacement and expansion diagnostic rather than a clean semantic metric.

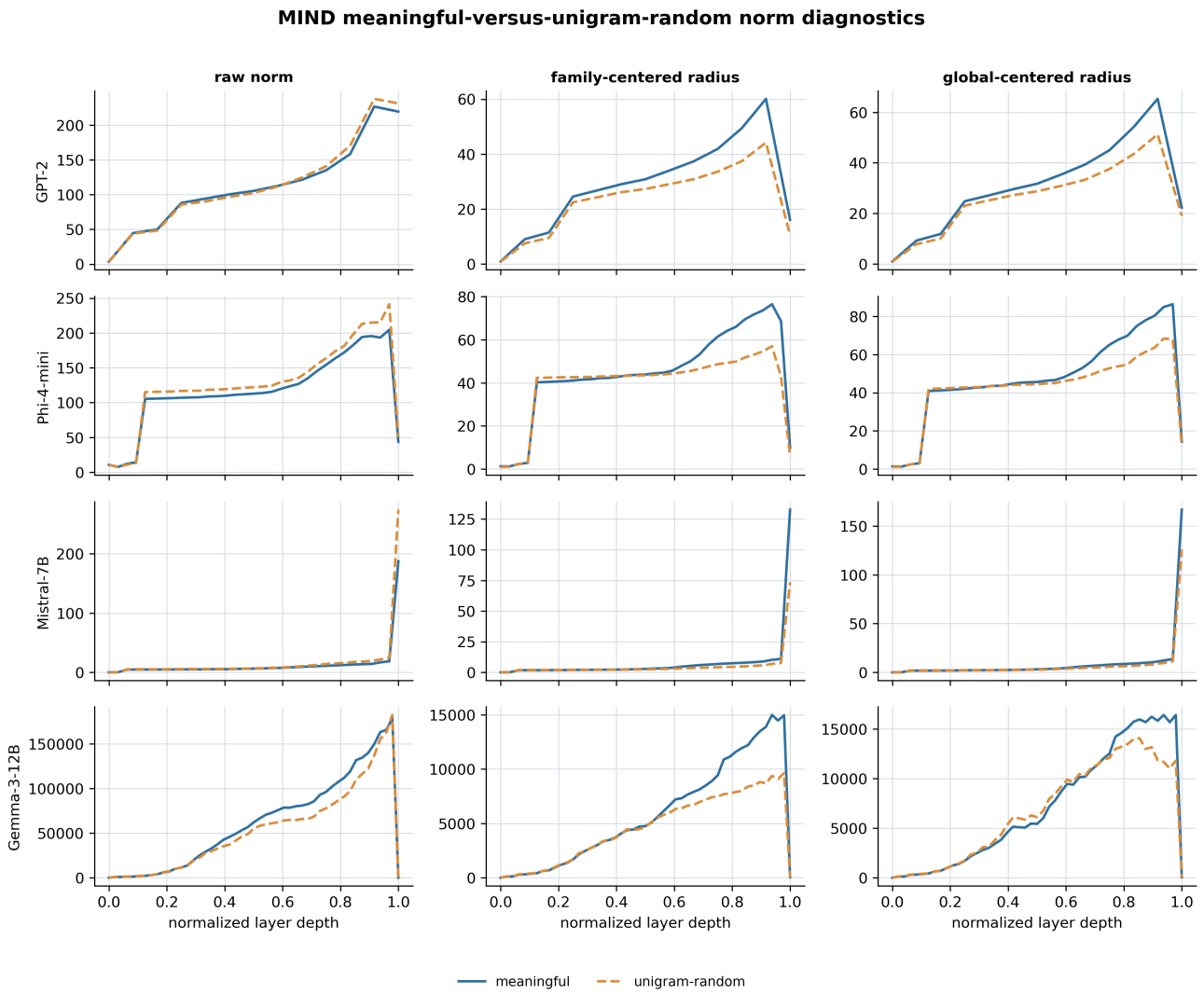

Figure 6. MIND meaningful-versus-unigram-random norm diagnostics for all four checkpoints under `mean_pool` extraction. Columns show raw norm, family-centered radius, and global-centered radius; rows show models. Unigram-random controls have comparable or larger raw norms, while meaningful prompts show larger family-centered radius in the final layer.

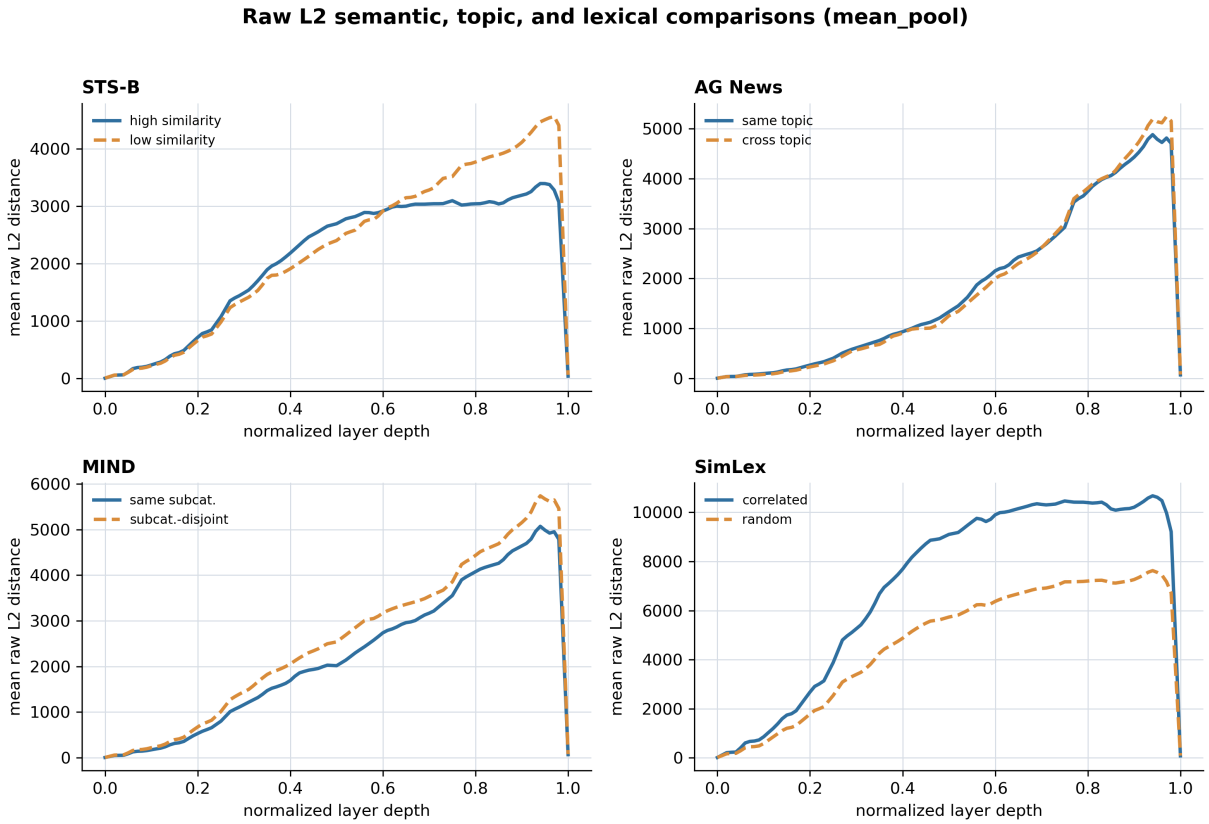

Figure 7. Raw L2 comparisons for semantic, topic, and lexical prompt families under `mean_pool` extraction. Each panel compares a related condition with its controlled alternative. The curves show that raw Euclidean distances grow substantially across both related and control surfaces.

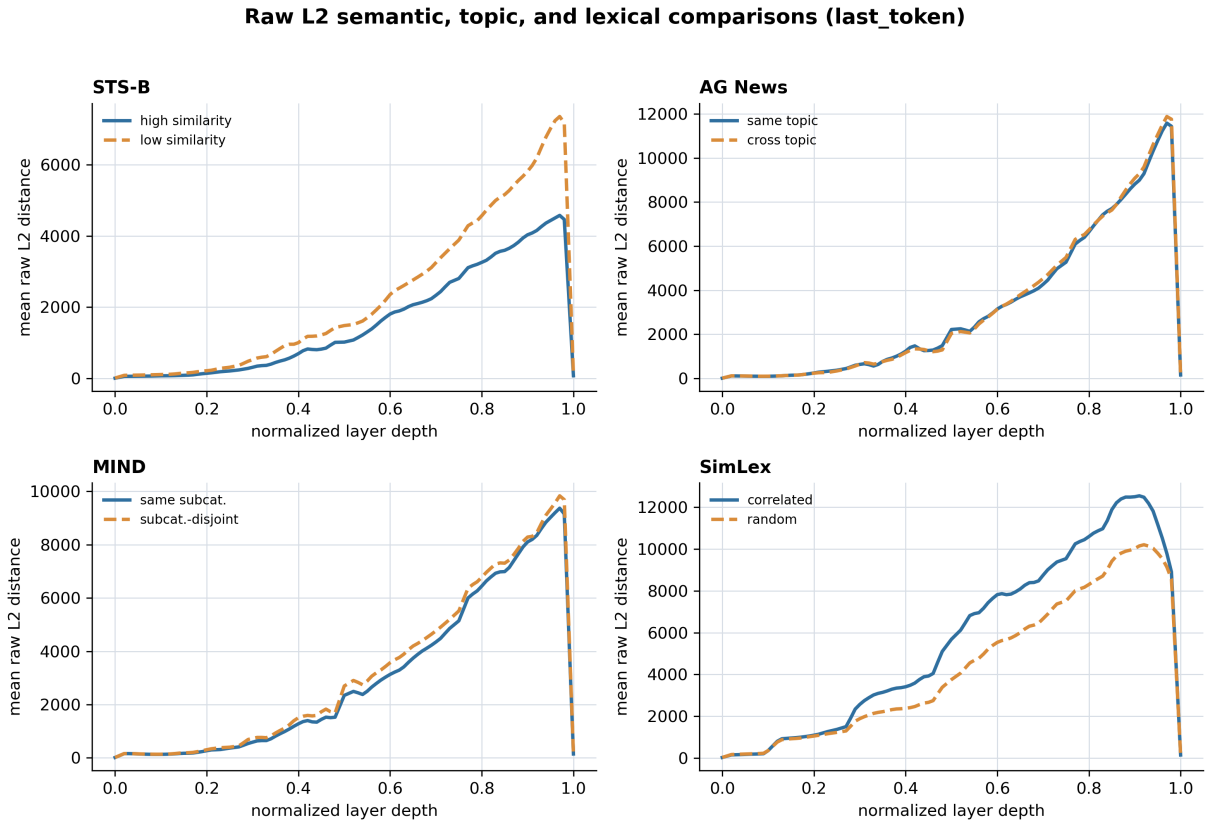

Figure 8. Raw L2 comparisons for the same semantic, topic, and lexical prompt families under `last_token` extraction. The decoder-final-position surface shows the same broad Euclidean expansion pattern, reinforcing that raw L2 is an expansion diagnostic rather than a standalone semantic metric.

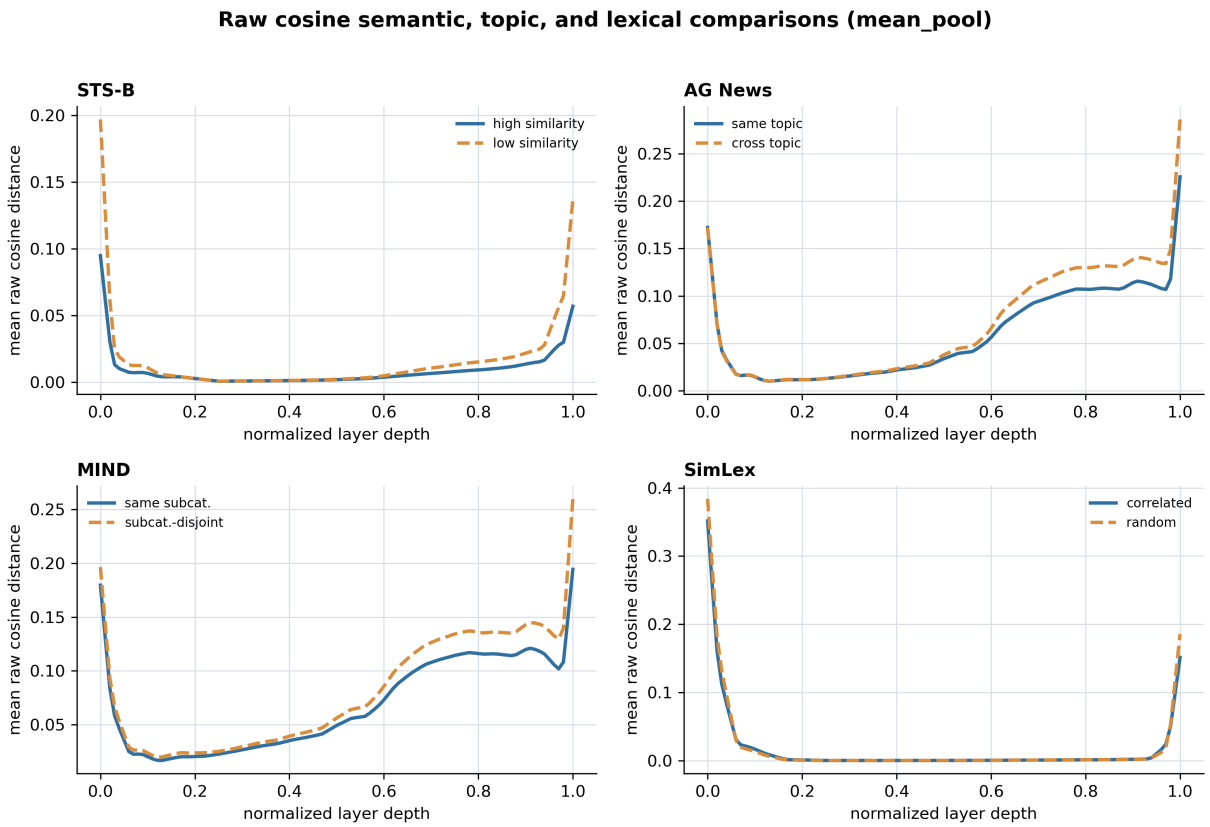

Figure 9. Raw cosine comparisons for semantic, topic, and lexical prompt families under `mean_pool` extraction. These curves show the angular geometry before family centering, where common-direction effects can compress or distort the relation between related and controlled alternatives.

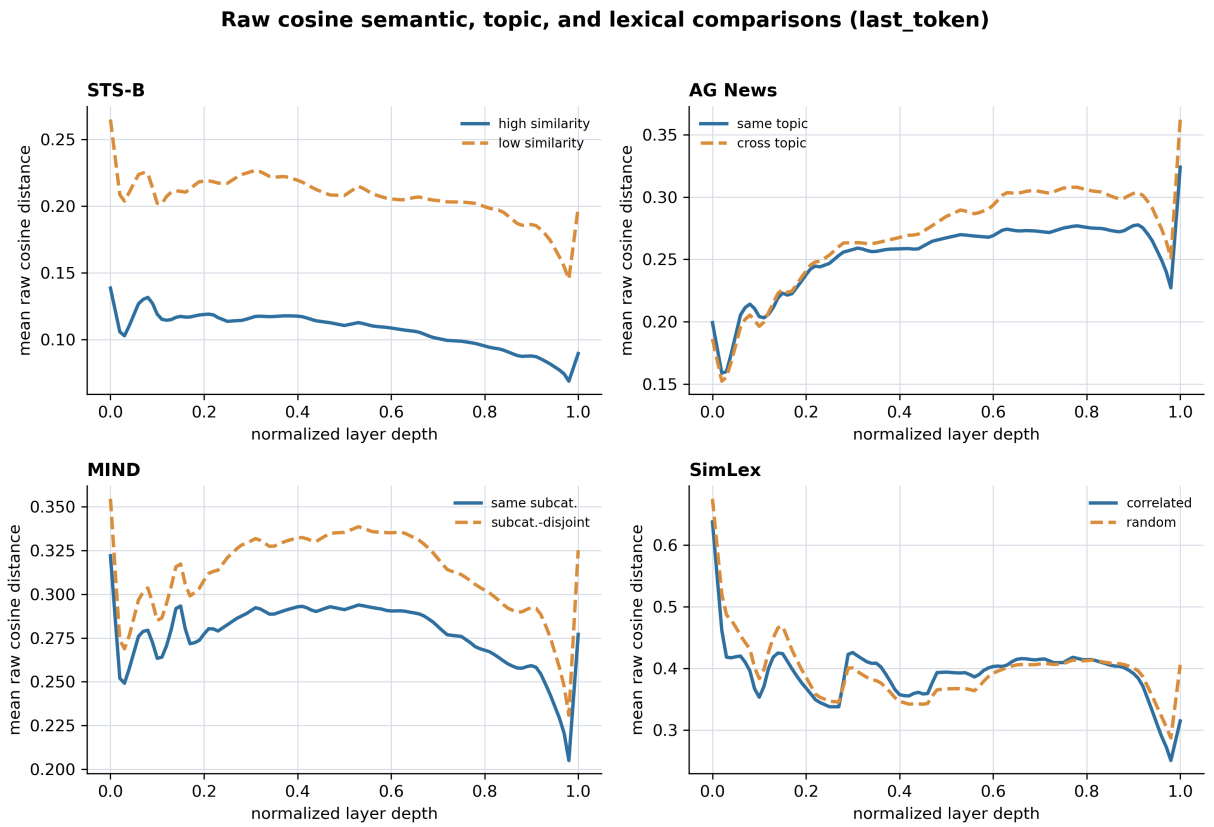

Figure 10. Raw cosine comparisons for semantic, topic, and lexical prompt families under `last_token` extraction. This decoder-position baseline provides the raw angular counterpart to the centered-cosine robustness checks.

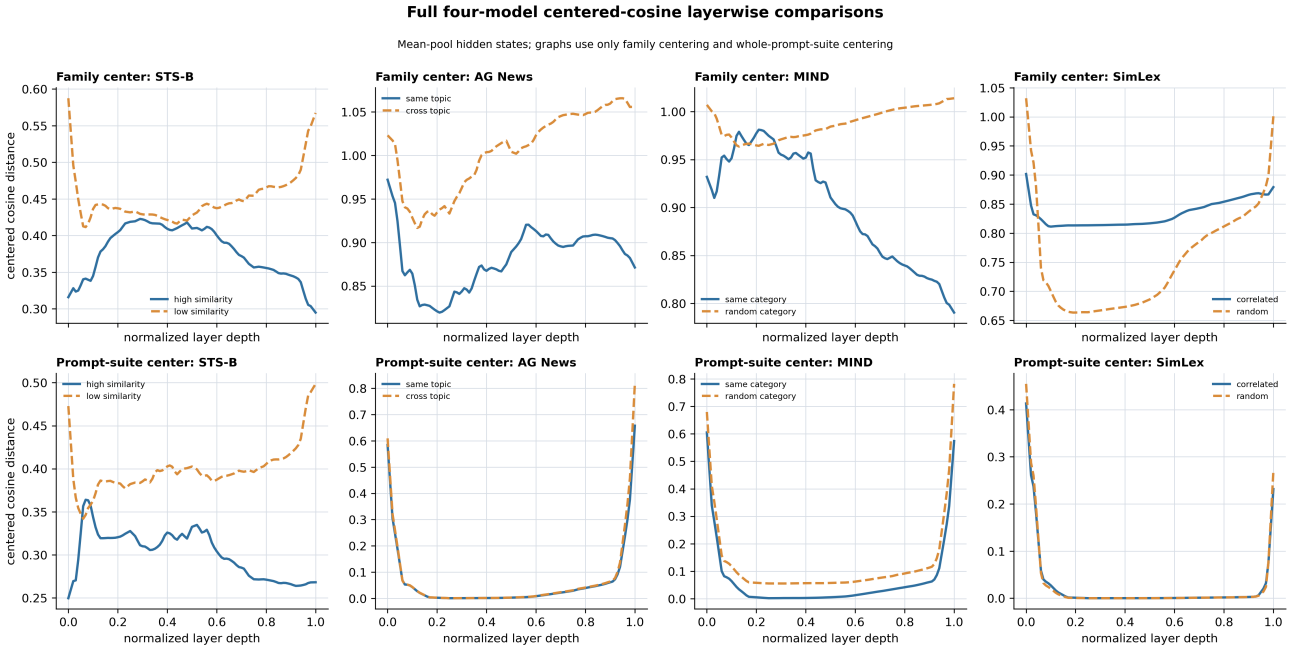

Figure 11. Centered cosine comparisons for the controlled prompt families under `mean_pool` extraction. The plot averages all four retained checkpoints. Related conditions remain closer than controlled alternatives for STS-B, AG News, and MIND after both compared groups share a single contrast centroid; SimLex is positive on average but more model-variable.

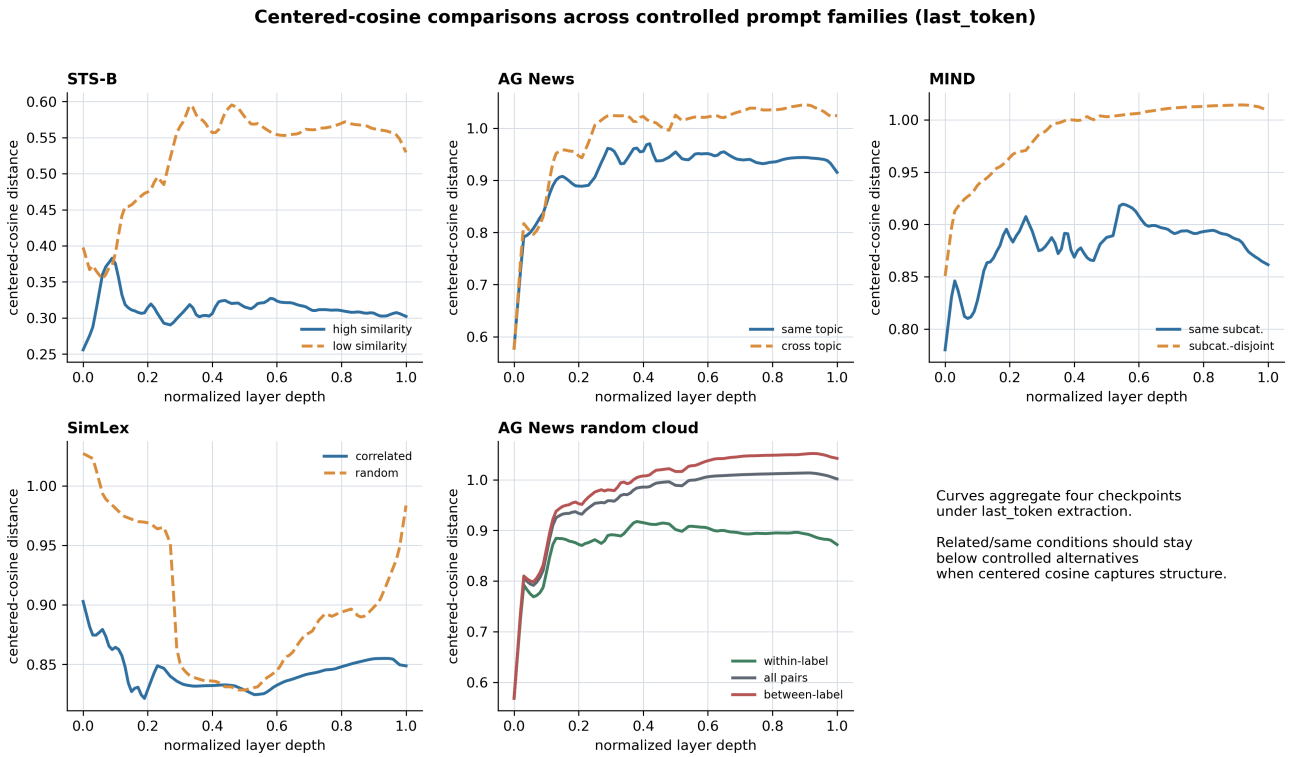

Figure 12. Family-centered cosine comparisons under `last_token` extraction from the earlier four-checkpoint baseline. This decoder-final-position view is retained as a representation-choice reference, while the main centered-cosine presentation uses the centered `mean_pool` result.

The final robustness check varies the centering surface itself. The contrast centroid is the protocol used for the paper's main centered-cosine result. The family centroid centers each prompt family separately, while the suite centroid pools the whole rough probe suite. If a result is positive under contrast centering and comparable to the family-centered value, the separation is less likely to be only a prompt-family centering artifact.

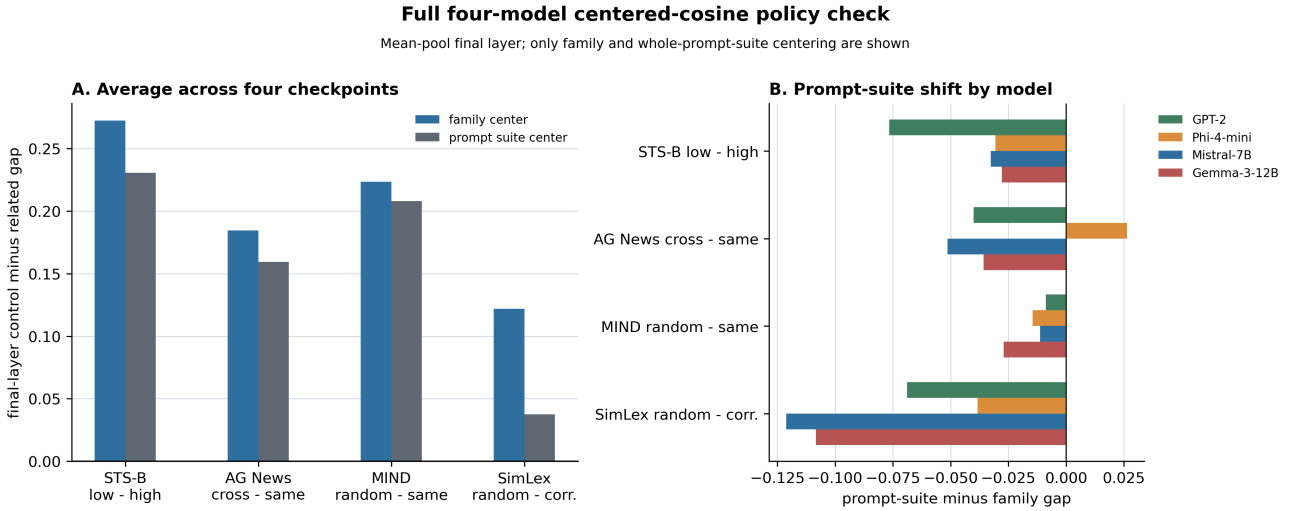

Figure 13. Centering-scope robustness probe for final-layer centered-cosine gaps under `mean_pool` extraction. The probe uses all four retained checkpoints and compares family centering, contrast centering, and suite centering. STS-B, AG News, MIND, and SimLex are positive on average under contrast centering, with SimLex remaining the most model-variable control.

Contrast	Family center	Contrast center	Suite center	Reading
STS-B low minus high similarity	0.2725	0.2730	0.2306	robust
AG News cross minus same topic	0.1846	0.1818	0.1594	robust
MIND random minus same category	0.2235	0.2254	0.2081	robust
SimLex random minus correlated words	0.1218	0.1127	0.0376	model-variable

Table 8. Centering-scope robustness for final-layer centered-cosine gaps in the full four-checkpoint pass. Values average GPT-2, Phi-4-mini, Mistral-7B, and Gemma-3-12B under `mean_pool` extraction. Positive values mean the controlled alternative is farther than the related condition after centering. Contrast centering uses the union of the two compared groups rather than a prompt-family-only centroid.

A.8. Interpretation Boundary

This appendix supports the paper's interpretation boundary. Raw hidden-state norms and raw pairwise L2 distances can expand with depth for both coherent prompts and unigram-random controls. This expansion alone should not be presented as evidence that a model has discovered semantic distance. The semantic question begins after controlling for common geometric effects, which is why the main centered-cosine result is reported using a contrast centroid over controlled prompt comparisons.

The centering-scope comparison adds a second boundary. It supports the sentence, topic, and category results under contrast centering across all four retained checkpoints. STS-B, AG News, and MIND remain positive when the centroid is computed over the two compared groups. SimLex is also positive on average, but its model-level signs are less consistent than the sentence, topic, and category contrasts: Phi-4-mini is slightly negative in the full centered `mean_pool` pass, while GPT-2, Mistral-7B, and Gemma-3-12B are positive. The appendix therefore supports saying that the strongest current evidence is for sentence-level, topic-level, and category structure; lexical word-level structure should be framed as suggestive until a larger lexical-specific suite is run.