WHAT LIES WITHIN: DIAGNOSTICS OF COMPOSITIONAL AND SEQUENTIAL BEHAVIOR IN LONG-CONTEXT TEXT EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embedding evaluation has long relied on indirect external tasks such as retrieval, owing to the opaque, vector-valued nature. We depart from this dependence and present a paradigm that directly analyzes the compositional behavior of the output vector under cumulative synthesis. RSS measures the coherence between a text and its embedding representations, while SRS measures the order sensitivity across positions in text. RSS correlated with downstream retrieval performance (LEMB, Pearson $r = 0.67$, and $r = 0.80$ on NarrativeQA task). SRS likewise correlated with LEMB, with the negative SRS score correlating at $r = 0.64$. RSS required relatively little measurement time compared with SRS and LEMB ($0.52\times$ that of LEMB); it can therefore be used to analyze performance by sequence length quickly. Every embedder exhibited a distinguishable SRS pattern, and these patterns tracked each model's fine-tuning lineage and base model. Measuring and visualizing order sensitivity—as SRS does for single-vector embedders—constitutes, to the best of our knowledge, the first attempt to interpret and structure how the position and order information of a document is preserved at the final representation level. FINESSE, the framework behind these results, will be released as a Python package.

1 INTRODUCTION

For LLMs, context-window verification is feasible because their outputs can be evaluated—by humans, automated metrics, or a stronger LLM (Zhang et al., 2024b; Hsieh et al., 2024; Yen et al., 2024; Modarressi et al., 2025). Embedders, by contrast, need an external task for evaluation. LONGEMBED (Zhu et al., 2024) met this need by extending the retrieval-based evaluation methodology of well-established benchmarks like BEIR and MTEB (Thakur et al., 2021; Muennighoff et al., 2023).

However, retrieval-based extensions have a blind spot. Consider a retrieval corpus in which the key information is concentrated near the beginning of each document: an embedder can then attain a high score without inspecting the document's tail. LONGEMBED levelled this criticism at LOCO (Zhu et al., 2024; Saad-Falcon et al., 2024, Figure 2), yet whether LONGEMBED shares the same bias has not been investigated. Meanwhile, embedders claiming strong long-context capability are being released at a remarkable rate, outpacing the development of robust inspection regimes (Nussbaum et al., 2024; Yu et al., 2024; Babakhin et al., 2025; Zhang et al., 2025a; Akram et al., 2026). This tension between release speed and inspection regime makes it difficult to select an embedder suited to a given task.

For LLMs, moreover, the mechanistic interpretability of outputs is actively applied to practical model analysis, thanks to the residual stream—the multi-layer, multi-token activations of an LLM (Rai et al., 2024; Templeton et al., 2024). Such studies directly analyze how the position and order of meaning within a document affect the LLM's output.

In embedding, by contrast, the information carried by the various positions of a document and their order is compressed into a single vector (Weller et al., 2026). An approach that directly analyzes multi-layer, multi-token activations, as is done for LLMs, is therefore difficult to transplant. As a result, research on the mechanistic interpretability of embeddings has proceeded by way of a restricted analysis (Park et al., 2025; Cai et al., 2026). Separately, embedders are known to be sensitive to word

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

order (Opitz et al., 2025, p.9), and several embedding studies indeed report positional-awareness and retrieval-performance biases (Coelho et al., 2024; Fayyaz et al., 2025; Yu et al., 2026). Prior work, however, concentrates on detecting positional bias. There is consequently little empirical study of how order sensitivity contributes to the final vector.

We therefore propose FINESSE: Framework for In-context Narrative Evaluation of Structural Sensitivity in Embeddings, a new analytical paradigm that analyzes the compositional behavior of the embedding operator under cumulative synthesis. FINESSE comprises RSS and SRS: the former measures the coherence between a text and its embedding representation, and the latter measures position-wise order sensitivity. Analyzing various models, we arrive at two central findings and contributions.

RSS correlates with holistic understanding. The RSS metric shows a strong, selective correlation with LEMB's Real Tasks, which test for structural comprehension (Pearson $r = 0.80$ on NarrativeQA), but not with its Synthetic Tasks, which test for local fact retrieval. RSS further detects the LEMB regression of **nomic-embed-text-v1.5** relative to its predecessor as the only negative score in our model set. Moreover, RSS could be measured quickly (mean evaluation time: LEMB 14.21 min vs. RSS 7.40 min). Thus, practical long-context capability of embedders can be assessed more rapidly through RSS.

SRS predicts through two channels. Forward sequence recognition correlates with Synthetic Tasks (Spearman $\rho = 0.63$ on Passkey), while its negative counterpart correlates with Real Tasks (Pearson $r = 0.64$ on NarrativeQA). SRS patterns were further meaningful in that they visualize a model's order sensitivity as a function of document position. Through this visualization we could trace base-model lineage and the model's fine-tuning origin by pattern similarity.

From these two findings and contributions, we expect FINESSE to prove useful across the full span of research that trains and deploys embedders.

2 Related Works

e5-mistral-7b-instruct (Wang et al., 2024a), which adapted **mistral-7b** (Jiang et al., 2023) into an embedder—initially capped at 4096 tokens, as documented in (intfloat), but extended to 32768 via RoPE and NTK-aware scaling (Su et al., 2024; Peng & Quesnelle, 2023; Zhu et al., 2024). This not only spawned a family of derivative models (Choi et al., 2025; Chen et al., 2024a; Linq-AI-Research; Meng et al., 2024; Camara, 2024) but also established Causal-LM-based embedders as the dominant paradigm, with recent state-of-the-art releases defaulting to 32K+ context windows (Babakhin et al., 2025; Akram et al., 2026; Zhang et al., 2025a).

These developments created a need for long-context embedding benchmarks, yet MTEB's evaluation units extend only to single paragraphs (Muennighoff et al., 2023, p.5). LoCo (Saad-Falcon et al., 2024) and LongEmbed (Zhu et al., 2024)—the latter subsequently absorbed into MMTEB (Enevoldsen et al., 2025, p.5)—addressed this gap (throughout this paper, we use LEMB to refer to the LongEmbed test sub-set as absorbed into MMTEB), but no successor has since emerged: we therefore adopt LongEmbed (Zhu et al., 2024) as our primary downstream benchmark (see Appendix A.2 for LEMB-specific discussions).

Mechanistic interpretability has been applied to practical LLM analysis (Templeton et al., 2024), enabled by the residual stream (Rai et al., 2024). Weller et al. (2026) showed that the current single-vector paradigm of embedding is theoretically limited in representational capacity; the LLM-style, multi-layer analysis therefore does not transfer. Work transplanting SAE onto embedders has remained at interpreting feature contributions within a single embedding, and no prior work shows how each position—and its order—leaves a trace in the final vector (Park et al., 2025, p.4, Figure 2) (Cai et al., 2026, p.9).

The observation of Opitz et al. (2025, p.9) that embedders are sensitive to word order is corroborated by studies measuring positional contribution—through segment-to-whole cosine similarity (Coelho et al., 2024, p.4), by relocating an evidence sentence and testing score differences with a paired t-test (Fayyaz et al., 2025, p.5, Figures 3–4), or by leave-one-out and mirror reversal (Yu et al., 2026). These approaches successfully observe positional bias, however: segment-to-whole and leave-one-out reveal only per-position contribution, while mirror reversal cannot report per-segment contribution

and, in essence, cannot distinguish balanced positional contribution from the total absence of order recognition. Every prior approach is, in short, blind to a bag-of-words embedder.

There's a separate line of research which characterized embedding quality through properties of the vector space itself (Ethayarajh, 2019; Wang & Isola, 2020; Garrido et al., 2023). FINESSE differs in its object of measurement: rather than the static distributional shape of an embedding space, it probes the compositional behavior of the embedding operator under cumulative synthesis.

3 FINESSE: Framework for In-context Narrative Evaluation of Structural Sensitivity in Embeddings

	RSS	SRS
Target	Constituent/non-constituent separation under cumulative composition	Representational sensitivity to order changes
Positive (+)	Constituents are meaningfully closer than non-constituents	Closer to contexts containing the forward insertion
≈ 0	Hard to separate constituents from non-constituents	No clear directional preference
Negative (−)	Non-constituents are closer than constituents	Closer to contexts containing the reversed insertion
Calculation	$(\Phi_{\mathrm{TD}} + \Phi_{\mathrm{BU}}) \times 500$ (raw $[-4, +4]$; scaled $[-2000, +2000]$)	$(\mathrm{median}_{\mathrm{pos}} - \mathrm{median}_{\mathrm{neg}}) \times 1000$ (raw $[-2, +2]$; scaled $[-2000, +2000]$)

FINESSE is an analytical tool that probes the compositional behavior of the embedding operator under cumulative synthesis. It treats the text to be embedded as a sequence of independent semantic atoms—chunks that are individually embeddable yet semantically distinct—and measures how their similarities evolve as these atoms are progressively combined. This methodology builds upon an earlier conceptual exploration of embedding synthesis by Ko (2025).

3.1 Terminology and Overview

Term	Symbol	Definition	Our Setup
Semantic atom	c_i	An n-token text chunk that can be sampled and embedded independently	$n = 500$ tokens
Number of atoms	L	Number of atoms per test; total $L \times n$ tokens	$L_{\mathrm{RSS}} = 4 \cdots 16$, $L_{\mathrm{SRS}} = 8$
Individual embedding	$e_i = \mathcal{S}(c_i)$	Embedding of an atom in isolation	—
Cumulative composition	$s_j = \mathcal{S}(c_1, \ldots, c_j)$	One embedding of the concatenated first j atoms	Up to 8,000 tokens
State matrix	$S^*_{ij} = \mathrm{sim}(e_i, s_j)$	Complete record of similarity changes during composition	$L \times L$

The **cumulative synthesis operator** $\mathcal{S}(c_1, \ldots, c_k)$ implements the abstraction described above. In our experiments, we evaluate two distinct implementations.

- $\mathcal{S}(c_1, \ldots, c_k) = \textbf{Encoder}(c_1 \oplus \cdots \oplus c_k)$ — the *primary* method: chunks are concatenated, then encoded.
- $\mathcal{S}(c_1, \ldots, c_k) = \textbf{mean}(\textbf{Encoder}(c_1), \ldots, \textbf{Encoder}(c_k))$ — the *control* method: chunks are encoded individually, then averaged. We refer to this as the **chunk-embed-then-average** strategy.

All similarities are cosine similarities; we use the quartile operators Q_1, Q_2, Q_3 as robust location summaries. In the experiments we fix **token_per_sample** $= n = 500$ and evaluate RSS over $L \in \{4, \ldots, 16\}$, SRS at $L = 8$. These will be the named defaults of the FINESSE package.

3.2 RSS: Robust Separation Score

Let $\mathbf{c}_i$ be the i-th atom, and let $\mathbf{e}_i = \mathcal{S}(c_i)$ be its individual embedding. The cumulative composition up to atom j is $\mathbf{s}_j = \mathcal{S}(c_1, \ldots, c_j)$. The $L \times L$ **state matrix** S^* collects the similarity between every atom and every composition step:

$$S^*_{ij} = \text{sim}(\mathbf{e}_i,\ \mathbf{s}_j) = \text{sim}(\mathcal{S}(c_i),\ \mathcal{S}(c_1, \ldots, c_j)) \tag{1}$$

The score is the sum of two coherences evaluated along the two directions of the matrix:

$$\text{RSS}(L) = 500 \times (\Phi_{\text{TD}} + \Phi_{\text{BU}}) \tag{2}$$

Top-Down Coherence (Φ_{TD}): asks whether each cumulative composition s_k separates its constituents ($j \le k$) from non-constituents ($j > k$).

$$\Phi_{\text{TD}} = \frac{1}{L-2} \sum_{k=2}^{L-1} \Big(Q_1\big(\{S^*_{jk} \mid j \le k\}\big) - Q_3\big(\{S^*_{jk} \mid j > k\}\big) \Big) \tag{3}$$

The range $k \in [2, L-1]$ avoids ill-defined boundaries. At $k = 1$, the constituent set $\{S^*_{j1} \mid j \le 1\}$ is a singleton, for which the quartile operator Q_1 is undefined. At $k = L$, the non-constituent set $\{S^*_{jL} \mid j > L\}$ is empty, rendering Q_3 undefined.

Bottom-Up Coherence (Φ_{BU}): asks the symmetric question from the atom's side, whether each atom c_k lies closer to the compositions that contain it than to those that do not.

$$\Phi_{\text{BU}} = \frac{1}{L-1} \sum_{k=2}^{L} \Big(Q_1\big(\{S^*_{kj} \mid j \ge k\}\big) - Q_3\big(\{S^*_{kj} \mid j < k\}\big) \Big) \tag{4}$$

$k = 1$ is excluded because the first atom $\mathbf{e}_1$ has no preceding cumulative synthesis, so the member set $\{S^*_{kj} \mid j < 1\}$ is empty and Φ_{BU} is ill-defined.

Both BU and TD correlate with the Final RSS ($r_{\text{TD}} \approx 0.97$, $r_{\text{BU}} \approx 0.96$, Appendix C.3). The theoretical range of RSS is $[-2000, +2000]$, but text embeddings yield effectively non-negative similarities, so values fall within $[-1000, +1000]$ in practice.

3.2.1 Numeric walkthrough for nvidia/llama-embed-nemotron-8b, L=4

This is the RSS walkthrough state matrix ($L = 4$). **Bold** entries mark atoms that are constituents of the column's composition. Calculated value is : $\text{RSS} = 500\,(0.4732 + 0.4350) = \mathbf{454.1}$,[1]

$\text{sim}(e_i, s_j)$	s_1	s_2	s_3	s_4
e_1	**1.00**	**0.70**	**0.51**	**0.37**
e_2	0.06	**0.67**	**0.47**	**0.38**
e_3	0.00	0.03	**0.58**	**0.44**
e_4	0.08	0.14	0.10	**0.60**

Reading columns left to right For each s_k we take Q_1 of constituent similarities minus Q_3 of non-constituent similarities. For s_2: $Q_1\{0.70, 0.67\} - Q_3\{0.03, 0.14\} = 0.5589$; for s_3: $Q_1\{0.51, 0.47, 0.58\} - Q_3\{0.10\} = 0.3875$. Averaging, $\Phi_{\text{TD}} = (0.5589 + 0.3875)/2 = 0.4732$, where we exclude s_1 (the identity case $s_1 = e_1$) and s_4 (no non-constituent remains).

Reading rows top to bottom For e_2: $Q_1\{0.67, 0.47, 0.38\} - Q_3\{0.06\} = 0.3656$; for e_3: $Q_1\{0.58, 0.44\} - Q_3\{0.00, 0.03\} = 0.4611$; for e_4: $Q_1\{0.60\} - Q_3\{0.08, 0.14, 0.10\} = 0.4782$. Averaging, $\Phi_{\text{BU}} = (0.3656 + 0.4611 + 0.4782)/3 = 0.4350$, where e_1 is excluded (no comparison group).

[1]The walkthrough state matrix is a single representative run drawn from the 25 repetitions; Values are rounded to two decimal places, and the arithmetic below uses unrounded values from this run.

3.3 SRS: Sequence Recognition Score

For L atoms, SRS defines $l = 2, \ldots, L-1$ and, for each l, $x = 0, \ldots, L-l$, and computes a sub-score $\mathrm{SRS}(l, x)$ for every pair (l, x); the full result is an (l, x) pattern, displayed as a heatmap.

We first define the **Probe query vector**, as $\mathbf{q}^{(l)} = \mathcal{S}(c_1, c_2, \ldots, c_l)$.

We then generate positive and negative groups combined with distractors. Each context vector $\mathbf{v}$ is generated by inserting the probe chunks $(c_1, \ldots, c_l)$ at position x within a distractor sequence $\mathbf{m} = (m_1, \ldots, m_{L-l})$ drawn from a large memory bank M, and then applying the synthesis operator $\mathcal{S}$ to the entire sequence.

$$V_{\mathrm{pos}}^{(l,x)},\ V_{\mathrm{neg}}^{(l,x)} = \left\{ \mathcal{S}(m_1, \ldots, m_x,\ \mathbf{c}_{\mathrm{fwd/rev}},\ m_{x+1}, \ldots, m_{L-l}) \,\middle|\, \mathbf{m} \in M \right\}. \tag{5}$$

For **Positive context group ($V_{\mathbf{pos}}^{(l,x)}$)**, the probe is inserted in its **forward order**, preserving the same ordinal structure as the query vector. For **Negative context group ($V_{\mathbf{neg}}^{(l,x)}$)**, the probe is inserted in **reverse order**, so that the ordinal structure is inconsistent with that of the query vector.

Let $\mathrm{Sim}_{\mathrm{pos/neg}}^{(l,x)} = \{\mathrm{sim}(\mathbf{q}^{(l)}, \mathbf{v}) \mid \mathbf{v} \in V_{\mathrm{pos/neg}}^{(l,x)}\}$ denote the resulting similarity sets. In this way, a matrix assembling the similarities of the sub-tests can be obtained as the result of $\mathrm{SRS}(L)$.

The cell score is the median contrast:

$$\mathrm{SRS}_{lx}^{(L)} = 1000 \times \left[Q_2\left(\mathrm{Sim}_{\mathrm{pos}}^{(l,x)}\right) - Q_2\left(\mathrm{Sim}_{\mathrm{neg}}^{(l,x)}\right) \right]. \tag{6}$$

The score is defined for $l \in \{2, \ldots, L-1\}$ and $x \in \{0, \ldots, L-l\}$; $l = L$ is excluded as it leaves no distractor. SRS fundamentally asks: *does the positive query have higher similarity than the negative query?* A score near zero indicates order-insensitivity; a clearly positive or negative score represents sensitivity to order at (l, x). As with RSS, the theoretical range is $[-2000, +2000]$[2]. We mask cells with $p \geq 0.05$ (one-sample t-test against zero) with black hatching.

3.3.1 Numeric walkthrough for nvidia/llama-embed-nemotron-8b, L=8, l=5, x=0

We insert the probe $q = \mathcal{S}(c_1, \ldots, c_5)$ forward at $x = 0$ among three distractors (25 contexts, positive group), and the reversed probe $(c_5, \ldots, c_1)$ into the same distractors (25 contexts, negative group). The cell score is the difference between the medians of the two similarity sets: yielding $\mathrm{SRS}(l{=}5, x{=}0) = 1000\,(0.7515 - 0.7854) = \mathbf{-33.9}$[3]

25 $\cos(q, v_{\mathrm{pos}})$ values				
min	Q_1	**median**	Q_3	max
0.7257	0.7428	**0.7515**	0.7584	0.7795
25 $\cos(q, v_{\mathrm{neg}})$ values				
min	Q_1	**median**	Q_3	max
0.7570	0.7774	**0.7854**	0.7922	0.8062

4 Experiments

4.1 Experimental Setup

Models. We evaluate 34 models, fixing **max_ctx** $= 8190$ across all models. This cap reflects the lowest reliably verifiable context window across the model pool—several models declare 32K+ windows but exhibit either training-time configurations well below this or runtime crashes near the declared maximum(see Appendix F.1 for selection criteria, per-model context investigation).

[2]Values fall within $[-1000, +1000]$ in practice.

[3]Single representative run from the 25 repetitions, Median values rounded to four decimals.

LEMB protocol. We follow the official repository of Zhu et al. (2024); dwzhu-pku, adding only minimal compatibility and automation code, and respecting each model's pooling and prefix conventions. Because **max_ctx** is capped at 8190, we exclude lengths $\{16384, 32768\}$ from the Synthetic Tasks.

FINESSE protocol. We use a chunk size of $n = 500$ tokens, balancing long-context coverage with atom-level semantic independence. Atoms are sampled from the English subset of CulturaX (Nguyen et al., 2024; uonlp), specifically **en/en_part_0000*.parquet**. We evaluate RSS across $L \in \{4, \ldots, 16\}$ and SRS at a fixed $L = 8$, with 25 repetitions per length used for statistical testing.

Control baseline. We additionally evaluate **mean(ml-e5-base)**[4], which applies the **chunk-embed-then-average** synthesizer to **multilingual-e5-base** (Wang et al., 2024b), as a baseline for both metrics.

4.2 Experimental Results

4.2.1 Correlation Analysis Between RSS Scores and LEMB Performance

We present the RSS scores at $L = 4, 8, 16$ for selected model groups, together with the $L = 4, 5, \ldots, 16$ average (see Appendix Table 18, Figure 21 for the full RSS scores of the 34 models across $L = 4, 5, \ldots, 16$).

Table 1: RSS scores for selected model groups.

Model	$L = 4$	$L = 8$	$L = 16$	Avg($L4 \ldots L16$)
Top 4 models ranked by avg. score				
llama-embed-nemotron-8b	**456.64**	**263.29**	124.29	**236.16**
F2LLM-v2-4B	398.66	237.36	**140.65**	223.56
F2LLM-v2-1.7B	366.79	225.50	125.68	213.55
F2LLM-v2-8B	383.29	225.48	125.70	211.84
The nomic-embed-text series				
nomic-embed-text-v1	156.18	88.29	44.03	80.32
nomic-embed-text-v1.5	139.39	65.19	-16.12	41.67
Baseline				
mean(ml-e5-base)	99.97	57.22	30.53	53.38

Within Table 1, the model with the highest $L4 \ldots L16$ average is **llama-embed-nemotron-8b**, which exceeds the baseline model by up to 350 points at $L = 4$. However, while this model achieves the top score at $L = 4$ and $L = 8$, at $L = 16$ it is overtaken by **F2LLM-v2-4B**, **F2LLM-v2-8B**, and **F2LLM-v2-1.7B**, and finishes in fourth place. Since all models exhibit a significant decrease in score as L increases at $L4 \rightarrow L8$ and $L8 \rightarrow L16$, ($p < 0.005$; see Appendix E.4), scores at low L and scores at high L must both be taken into account in interpretation.

The most distinctive finding lies among the lower-ranked models in RSS. **nomic-embed-text-v1.5** is the only model across the entire set that records a negative score in the long-context regime ($L > 12$). Let us now examine how the fact that v1.5—the newer version, compared to **nomic-embed-text-v1**—records a negative RSS score actually affects downstream tasks.

In Table 2, **v1.5** consistently receives a lower LEMB score than **v1** (see Appendix E.3 for comparison with other models).

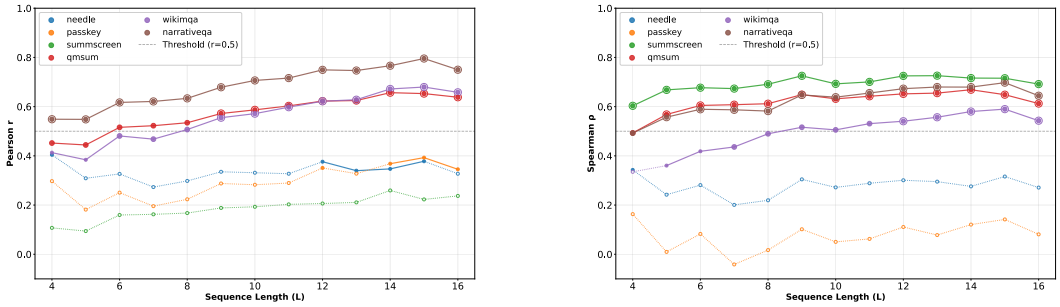
v1.5's only documented architectural change from **v1** is the introduction of Matryoshka Representation Learning (Kusupati et al., 2022), which nomic-ai describe as having negligible impact. RSS exposes this as the only negative score in our model set, and LEMB confirms a substantive regression—suggesting that the representational compression came at a real cost to long-context capability.

[4]Abbreviated from **mean(multilingual-e5-base)** for table space.

Table 2: LEMB results for the nomic-embed-text series.

Model	Synthetic (Acc@1)		Real (nDCG@10)				Avg.
	Needle	Passkey	SummScreenFD	QMSum	WikimQA	NarrativeQA	
nomic-embed-text-v1	0.350	0.593	0.921	0.348	0.737	0.379	0.555
nomic-embed-text-v1.5	0.253	0.483	0.668	0.141	0.363	0.154	0.344
mean(ml-e5-base)	0.293	0.843	0.949	0.378	0.737	0.403	0.601

LEMB and RSS scores thus point in the same direction. To characterize the correlation more precisely, we computed both Pearson and Spearman correlation coefficients between LEMB scores and every RSS value in the range $L = 4, 5, \ldots, 16$.



(a) Pearson Correlation (r)

(b) Spearman Correlation (ρ)

Figure 1: Correlation coefficients between RSS and each LEMB Task across $L = 4, 5, \cdots, 16$. Marker style encodes significance: small open circle ($p \geq 0.05$), filled circle ($p < 0.05$), larger filled circle ($p < 0.01$), and double-ringed circle ($p < 0.001$). Task lines are drawn as dotted segments connecting non-significant points. The horizontal dashed line marks $r = 0.5$ as a reference threshold. (a) Pearson r; (b) Spearman ρ.

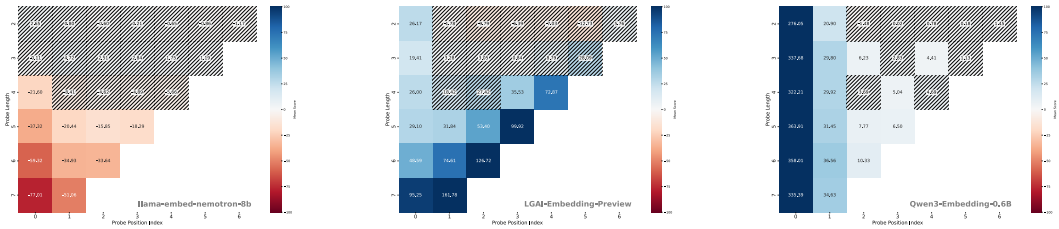
The correlation generally increases from $L = 4$ to $L = 15$, peaking at Pearson $r = 0.6733$, $p < 0.0001$ (Appendix C.1 for the full numerical table). Figure 1 resolves this aggregate trend into its per-task components, revealing a dissociation: Real Tasks (SummScreenFD, QMSum, WikimQA, NarrativeQA) drive the correlation, while Synthetic Tasks (Needle, Passkey) remain below or near the $r = 0.5$ threshold throughout the entire length range. A joint analysis of this dissociation together with the SRS metric follows in Section 4.2.2.

Before turning to SRS, one methodological concern remains: could the RSS-LEMB correlation be an artifact of the specific evaluation corpus? We replicated the RSS($L = 15$) measurements on Wikipedia and obtained Pearson $r = 0.966$ ($p < 10^{-6}$) against the CulturaX scores, confirming that RSS captures model-intrinsic properties rather than corpus-specific features (Appendix C.2). Also, an exploratory extension of RSS to $L = 64$ (32K tokens) for the 11 models with sufficient context windows was conducted; it shows that the positive-to-negative transition point varies systematically across models (see Appendix C.5).

4.2.2 Correlation Analysis Between SRS Scores and LEMB Performance

The full SRS score pattern heatmaps at $L = 8$ for all 34 models are analyzed in Appendix E.2, Appendix Figure 20. Here, we present only the SRS $L = 8$ patterns for selected model groups. For a partial extension to $L = 16$, see Appendix B.2.

Figure 2 shows three representative patterns. **Qwen3-Embedding-0.6B** concentrates sequence sensitivity at $x = 0$, the dominant pattern across the model set—sequence recognition restricted to the early portion of the text. **LGAI-Embedding-Preview** shows the opposite, with positive values along SRS$(7, 1)$, SRS$(6, 2), \ldots,$ SRS$(2, 6)$; Figure 3 shows this late-concentration derived from the **e5-mistral-7b-instruct** base model (Appendix E.2 for further lineage analysis). **llama-embed-nemotron-8b** is the only model with pronounced negative values in every statistically significant



(a) llama-embed-nemotron-8b (b) LGAI-Embedding-Preview (c) Qwen3-Embedding-0.6B

Figure 2: SRS heatmaps for three models at $L = 8$, showcasing distinct patterns of sequence recognition: fully negative (left), late-concentration (middle), and early-concentration (right).

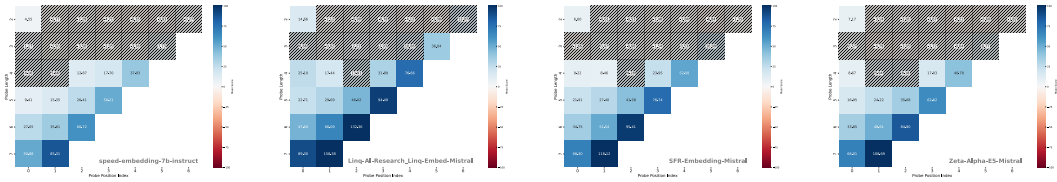


Figure 3: Inheritance of the late-concentration signature. From left to right: speed-embedding-7b-instruct, Linq-Embed-Mistral, SFR-Embedding-Mistral, Zeta-Alpha-E5-Mistral.

cell, and given its top-tier RSS and LEMB performance (Appendix E.1), this fully negative pattern becomes our central anomaly (Appendix E.2.4).

We hypothesize that models with a stronger negative SRS tendency achieve higher LEMB, as exemplified by **llama-embed-nemotron-8b**. Specifically, we denote the negative-only counterpart by $\mathbf{srs}_{\text{neg}}$: the absolute value of the sum of negative SRS scores across cells, divided by the total number of cells. As controls, we additionally compute the positive counterpart $\mathbf{srs}_{\text{pos}}$ and the overall mean $\mathbf{srs}_{\text{mean}}$ using the same total cell count. To situate these against our other primary metric, we jointly analyze them with $\text{RSS}(L = 15)$ in Table 3.

Table 3: Correlation between LEMB task scores and sensitivity metrics ($\mathbf{srs}_{\text{neg}}$, $\mathbf{srs}_{\text{pos}}$, $\mathbf{srs}_{\text{mean}}$, $\text{RSS}_{L=15}$). Significance: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$. Non-significant values ($p\geq0.05$) are shown in gray italics. Exact p-values at full precision are provided in Appendix A.3.

	Pearson r				Spearman ρ			
Task	$\mathbf{srs}_{\text{neg}}$	$\mathbf{srs}_{\text{pos}}$	$\mathbf{srs}_{\text{mean}}$	$\text{RSS}_{L=15}$	$\mathbf{srs}_{\text{neg}}$	$\mathbf{srs}_{\text{pos}}$	$\mathbf{srs}_{\text{mean}}$	$\text{RSS}_{L=15}$
Synthetic Tasks								
Needle	0.3881*	0.3612*	*0.2845*	0.3780*	*0.3271*	*0.3134*	*0.2580*	*0.3160*
Passkey	*0.3359*	0.4381**	*0.3328*	0.3928*	0.3829*	0.6259***	0.4451**	*0.1416*
Real Tasks								
SummScreenFD	*0.3002*	*-0.0407*	*-0.1497*	*0.2229*	0.6193***	*-0.0506*	*-0.2312*	0.7155***
QMSum	0.4574**	*-0.0753*	*-0.1986*	0.6530***	0.5526***	*0.0050*	*-0.1224*	0.6486***
WikimQA	0.3919*	*0.0020*	*-0.1270*	0.6797***	0.4946**	*0.0924*	*-0.0824*	0.5899***
NarrativeQA	0.6412***	*-0.0505*	*-0.2017*	0.7956***	0.6333***	*0.1025*	*-0.0237*	0.6975***
Summary (Average across all tasks)								
Average	0.5399***	*0.1564*	*0.0077*	**0.6733*****	**0.6807*****	*0.2232*	*0.0503*	0.6028***

In Table 3, dissociation is revealed again. On Synthetic Tasks, $\mathbf{srs}_{\text{pos}}$ is the strongest predictor: it reaches Spearman $\rho = 0.626$ ($p < 0.001$) on Passkey, outperforming both $\mathbf{srs}_{\text{neg}}$ and $\text{RSS}(L = 15)$, and remaining competitive on Needle. This alignment is unsurprising—Passkey rewards literal string matching and Needle rewards short-span semantic matching, both of which benefit directly from the forward-probe-to-forward-insertion similarity that $\mathbf{srs}_{\text{pos}}$ measures.

The pattern inverts on Real Tasks. Here $\mathbf{srs}_{\text{pos}}$ collapses—its Pearson correlations turn weakly negative across all four tasks (e.g., Pearson $r = -0.051$ on NarrativeQA)—while $\mathbf{srs}_{\text{neg}}$ and $\text{RSS}(L = 15)$ exhibit their strongest signals: $r = 0.641$ and $r = 0.796$ on NarrativeQA respectively, both at

$p < 0.001$. SummScreenFD is a partial exception for both metrics, showing significance only in Spearman ($\rho = 0.619$, $\rho = 0.716$); we trace this to its narrow score IQR in Appendix C.4. $\mathbf{srs}_{\text{mean}}$, which aggregates both components, correlates with neither task family.

Taken together, Table 3 resolves the aggregate RSS-LEMB correlation reported in Section 4.2.1 into its source: it is driven almost entirely by Real Tasks, where RSS($L = 15$) and $\mathbf{srs}_{\text{neg}}$ consistently align. The two metrics, despite measuring different properties, converge on the same task family. Synthetic Tasks, by contrast, are the exclusive domain of $\mathbf{srs}_{\text{pos}}$. FINESSE thus predicts LEMB through two distinct channels selected by task type, each probing a different aspect of embedding geometry: $\mathbf{srs}_{\text{pos}}$ for literal matching, $\mathbf{srs}_{\text{neg}}$ for holistic understanding. Analysis for both RSS-LEMB and $\mathbf{srs}_{\text{neg}}$-LEMB correlations are further discussed in Appendix A.3 and ablated statistical robustness in Appendix D.1/D.2/D.3, dimensional correlations in Appendix D.4.

4.2.3 Time Cost Analysis Between RSS, SRS, and LEMB

Table 4: Benchmark wall-clock time (minutes) across 34 models.

Metric	n	Mean	Median	Std	IQR	Min	Max
RSS	34	7.40	5.86	5.53	10.56	1.51	18.77
SRS	34	366.00	245.80	266.62	571.38	27.99	745.97
LEMB	34	14.21	10.08	9.09	15.95	4.05	30.62

The evaluation times differed substantially across the three metrics. RSS was markedly lower than the other two metrics on the mean, median, minimum, and maximum alike (its mean was $0.52\times$ that of LEMB). This is especially notable given that, in this work, RSS was not measured in a single run but as 13 consecutive measurements over $L = 4, \ldots, 16$; a practical setting that restricts measurement to a few specific values of L (say, $L \in \{4, 8, 12, 16\}$) would therefore take even less time.

5 Conclusion

We proposed FINESSE, a framework that analyzes the compositional behavior of the embedding operator under cumulative synthesis, comprising RSS (the coherence between text and embedding representation) and SRS (position-wise order sensitivity). Across 34 models, RSS correlated selectively with LEMB's Real Tasks—those probing structural document understanding—and detected, as a negative score, the regression of **nomic-embed-text-v1.5** relative to its predecessor. Moreover, RSS is computationally efficient: computing all RSS values from $L = 4$ to $L = 16$ took, on average, only $0.52\times$ the time of LEMB. RSS can therefore be profitably applied to in-training embedder evaluation. SRS showed that order sensitivity is consistent and structured across models, and that these patterns transfer from the base model and partly persist through fine-tuning. This is significant in that it renders interpretable the traces that per-position and order information leave at the final representation level of a single-vector embedder, and its contribution lies especially in directly interpreting an embedder's document-level order sensitivity. With respect to LEMB, SRS was selective: positive SRS correlated with LEMB's Synthetic Tasks, and negative SRS correlated with LEMB's Real Tasks. This shows that SRS patterns are deeply connected to practical performance.

6 Limitations

SRS was evaluated only at $L = 8$; an extension to SRS($L = 16$) remained exploratory due to compute constraints. Out of an initial pool of roughly seventy candidates, outdated-library compatibility issues and the difficulty of verifying **max_ctx** across model cards reduced our final selection to 34 models with **max_ctx** uniformly fixed at 8190. This both inflated the share of certain families (e.g., F2LLM, $> 20\%$) and prevented us from probing the latent capability of embedders declaring windows beyond 8190—although we partially address the latter by extending L for RSS and SRS in Appendices C.5 and B.2. Evaluation was further restricted to English datasets. Finally, although the theoretical range of RSS and SRS is $[-2000, +2000]$, observed values rarely exceed an absolute value of about 500, since extreme cosine similarities seldom arise in practice—our normalization (by 500 or 1000) follows the theoretical definition rather than the empirical distribution.

7 AI USE STATEMENT

In this work, we used generative AI tools for brainstorming and discussing presentation structure, draft and edit parts of the manuscript to improve readability, and format references. We have not used generative AI tools for any tasks with generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, writing proofs, designing methodology, implementing methods, or interpreting results. We have thoroughly reviewed, verified, and cross-checked all AI-assisted text, mathematical notations, and citations against the raw experimental data and original sources. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

8 REPRODUCIBILITY STATEMENT

All configurations for the models listed in Appendix F.2, including model dimension, parameter size, and prefix style, are catalogued in that appendix. **All tables and figures in this paper** are reproduced by the code and data in Appendix F.3. Main experiments ran on a Colab Pro+ NVIDIA RTX PRO 6000 Blackwell Server Edition and some additional experiments are done on a separate environment such as NVIDIA B200; see Appendix F.4 for details.

REFERENCES

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Gunther, Maximilian Werk, and Han Xiao. Jina-embeddings-v5-text: Task-targeted embedding distillation. *arXiv preprint arXiv:2602.15547*, 2026.

Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, Vignesh P, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, Ken Barker, Martin Franz, Madison Lee, Todd Ward, Salim Roukos, David Cox, Luis Lastras, Jaydeep Sen, and Radu Florian. Granite Embedding R2 models, 2025. URL https://arxiv.org/abs/2508.21085.

BAAI. BAAI/bge-m3. huggingface models repository. URL https://huggingface.co/BAAI/bge-m3/blob/main/config.json. commit hash : b28ce2a6fcc9c75ef1c0619575d0ec19af760082.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-Embed-Nemotron-8B: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*, 2025.

Matthias Blum, Antonina Andreeva, Laise Cavalcanti Florentino, Sara Rocio Chuguransky, Tiago Grego, Emma Hobbs, Beatriz Lazaro Pinto, Ailsa Orr, Typhaine Paysan-Lafosse, Irina Ponamareva, Gustavo A Salazar, Nicola Bordin, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunic, Felipe Llinares-López, Aron Marchler-Bauer, Laetitia Meng-Papaxanthos, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Damiano Piovesan, Catherine Rivoire, Christian J A Sigrist, Narmada Thanki, Françoise Thibaud-Nissen, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, and Alex Bateman. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456, 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1082. URL https://doi.org/10.1093/nar/gkae1082.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. Eurobert: Scaling multilingual encoders for european languages, 2025. URL https://arxiv.org/abs/2503.05500.

Zhixin Cai, Jun Bai, Yang Liu, Jiaqi Li, Yichi Zhang, Taichuan Li, Zhuofan Chen, Zixia Jia, Zilong Zheng, and Wenge Rong. Xetrieval: Mechanistically explaining dense retrieval. *arXiv preprint arXiv:2605.29507*, 2026.

Arthur Camara. Fine-tuning an LLM for state-of-art retrieval: Zeta Alpha's top-10 submission to the MTEB benchmark. Zeta Alpha AI Research Blog, 2024. URL https://www.zeta-alpha.com/post/fine-tuning-an-llm-for-state-of-the-art-retrieval-zeta-alpha-s-top-10-submission-to-the-the-mteb-be.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. Little giants: Synthesizing high-quality embedding data at scale. *arXiv preprint arXiv:2410.18634*, 2024a.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024b.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8602–8615, 2022.

Jooyoung Choi, Hyun Kim, Hansol Jang, Changwook Jun, Kyunghoon Bae, Hyewon Choi, Stanley Jungkyu Choi, Honglak Lee, and Chulmin Yun. Lgai-embedding-preview technical report, 2025. URL https://arxiv.org/abs/2506.07438.

João Coelho, Bruno Martins, João Magalhães, Jamie Callan, and Chenyan Xiong. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 370–377, 2024.

dwzhu-pku. Longembed. github repository. URL https://github.com/dwzhu-pku/LongEmbed. commit hash : 033dc4a18e2da7b87d0e388c5b975cabe19c2fdc.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Marton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, et al. MMTEB: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. Accepted for ICLR.

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 (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9136–9152, 2025.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. COLM 2024.

intfloat. intfloat/e5-mistral-7b-instruct. huggingface models repository. URL https://huggingface.co/intfloat/e5-mistral-7b-instruct/blob/main/README.md. commit hash : d62cc4417be0e3289c03dd459d1ef2d137e2554e.

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

Donghyo Ko. vBERT: Vector sequence merger for limited embedding context extension, 2025. URL https://doi.org/10.5281/zenodo.17641562.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Linq-AI-Research. Linq-ai-research/linq-embed-mistral. huggingface models repository. URL https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. SFR-Embedding-Mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024. URL https://www.salesforce.com/blog/sfr-embedding/.

MoD Team. MoD-Embedding: A fine-tuned multilingual text embedding model, 2025. URL https://huggingface.co/bflhc/MoD-Embedding.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*, 2025. Accepted at ICML 2025.

mteb. mteb/leaderboard. huggingface spaces. URL https://huggingface.co/spaces/mteb/leaderboard. Accessed: 2026-03-24.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4226–4237, 2024.

nomic-ai. nomic-ai/nomic-embed-text-v1.5. huggingface models repository. URL https://huggingface.co/nomic-ai/nomic-embed-text-v1.5/blob/main/README.md. commit hash : e9b6763023c676ca8431644204f50c2b100d9aab.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.

Octen. Octen/octen-embedding-0.6b. huggingface models repository, a. URL https://huggingface.co/Octen/Octen-Embedding-0.6B.

Octen. Octen/octen-embedding-4b. huggingface models repository, b. URL https://huggingface.co/Octen/Octen-Embedding-4B.

Octen. Octen/octen-embedding-8b. huggingface models repository, c. URL https://huggingface.co/Octen/Octen-Embedding-8B.

Octen Team. Octen Series: Optimizing embedding models to #1 on RTEB leaderboard, 2025. URL https://octen-team.github.io/octen_blog/posts/octen-rteb-first-place/.

Juri Opitz, Lucas Moeller, Andrianos Michail, Sebastian Padó, and Simon Clematide. Interpretable text embeddings and text similarity explanation: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22314–22330, 2025.

Seongwan Park, Taeklim Kim, and Youngjoong Ko. Decoding dense embeddings: Sparse autoencoders for interpreting and discretizing dense retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26479–26496, 2025.

Typhaine Paysan-Lafosse, Antonina Andreeva, Matthias Blum, Sara Rocio Chuguransky, Tiago Grego, Beatriz Lazaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Felipe Llinares-Lopez, Laetitia Meng-Papaxanthos, Lucy J Colwell, Nick V Grishin, R Dustin Schaeffer, Damiano Clementel, Silvio C E Tosatto, Erik Sonnhammer, Valerie Wood, and Alex Bateman. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Research*, 53(D1):D523–D534, 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae997. URL https://doi.org/10.1093/nar/gkae997.

Bowen Peng and Jeffrey Quesnelle. NTK-aware scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation, 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have. Posted on r/LocalLLaMA.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Jon Saad-Falcon, Daniel Y Fu, Simran Arora, Neel Guha, and Christopher Ré. Benchmarking and building long-context retrieval models with LoCo and M2-BERT. *arXiv preprint arXiv:2402.07440*, 2024. International Conference on Machine Learning (ICML) 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021. Accepted at NeurIPS 2021 Dataset and Benchmark Track.

uonlp. uonlp/culturax. huggingface datasets repository. URL https://huggingface.co/datasets/uonlp/CulturaX. commit hash : 6a8734bc69fefcbb7735f4f9250f43e4cd7a442e.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, 2024a.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024b.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 11202–11216, 2026.

wikimedia. wikimedia/wikipedia. huggingface datasets repository. URL https://huggingface.co/datasets/wikimedia/wikipedia. commit hash : b04c8d1ceb2f5cd4588862100d08de323dccfbaa.

Wikimedia Foundation. Wikimedia Downloads. URL https://dumps.wikimedia.org.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024. ICLR 2025.

Daegon Yu, SeungYoon Han, and Woomyoung Park. Is position bias in dense retrievers built in-or learned from data? *arXiv preprint arXiv:2605.26578*, 2026.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*, 2024.

Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Wenbo Yang, Daiting Shi, and Xueqi Cheng. Bagging-based model merging for robust general text embeddings. *arXiv preprint arXiv:2602.05787*, 2026a.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1393–1412, 2024a.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. ∞ Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024b.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025a.

Ziyin Zhang, Zihan Liao, Hang Yu, Peng Di, and Rui Wang. F2LLM technical report: Matching SOTA embedding performance with 6 million open-source data. *CoRR*, abs/2510.02294, 2025b. doi: 10.48550/ARXIV.2510.02294. URL https://doi.org/10.48550/arXiv.2510.02294.

Ziyin Zhang, Zihan Liao, Hang Yu, Peng Di, and Rui Wang. F2LLM-v2: Inclusive, performant, and efficient embeddings for a multilingual world, 2026b. URL https://arxiv.org/abs/2603.19223.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. VideoPrism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024. Accepted to ICML 2024.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, 2021.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 802–816, 2024.

A Supplementary Discussions

A.1 Discussion about FINESSE's Further Usecase

FINESSE rests on the existence of *independently embeddable yet semantically distinct* atoms—a property that holds naturally for sufficiently long text chunks but not for a ViT patch or a single protein residue. Cross-domain extension thus requires domain-specific atom definitions. Promising candidates already exist: protein *domains* catalogued in Paysan-Lafosse et al. (2024) and Blum et al. (2024) provide biologically defined atom pools, and frame-level aggregation in video embedders such as Yan et al. (2022) and Zhao et al. (2024) treats each frame as an atom. Verifying whether FINESSE's text-domain findings reproduce under such definitions would extend its diagnostic power into a general account of sequence embeddings.

A.2 Discussions about LongEmbed's Task Dynamics

This appendix focuses on LongEmbed—currently the only long-context embedding benchmark that effectively exists—and in particular the LEMB test, the form in which it was absorbed as a sub-test of MMTEB (Enevoldsen et al., 2025, p.5).

The Synthetic Tasks of LEMB consist of Needle and Passkey. Needle consists of pairs of a short-form question and its answer, while Passkey consists of a numeric value tied to a character name together with the shared query "What is the number of ¡character name¿?". Distractor texts of lengths (256, 512, 1024, 2048, 4096, 8192, 16384, 32768) are then prepared, and the answer sentence is inserted at a random position to construct a test document. The model is then evaluated on whether it can retrieve the test document given the question sentence, using Acc@1 (Zhu et al., 2024, Table 2). Through this procedure, Passkey tests the preservation of literal matching (the preservation of verbatim "identical words"), while Needle tests the preservation of short-semantic matching (the semantic equivalence of short sentences).

The Real Tasks of LEMB take their names from their source datasets. SummScreenFD consists of human-written recaps (summaries) and transcripts (the main text) of screenplays (Chen et al., 2022). QMSum consists of the full transcripts of meetings together with their summaries and query/answer pairs against those summaries (Zhong et al., 2021). WikimQA consists of Wikipedia page titles and QA pairs derived from them (Ho et al., 2020). NarrativeQA consists of the full text of books together with their summaries and QA pairs (Kocisky et al., 2018).

LEMB interprets SummScreenFD and QMSum as tests that ask *how well one can retrieve the main text given its summary*, and WikimQA and NarrativeQA as tests that ask *how well one can retrieve the document that gave rise to a question given that question* (Zhu et al., 2024, p.4-5).

A.3 Discussions about LongEmbed, RSS and SRS correlations

Table 3 already reports the task-level correlations between SRS, RSS, and LEMB in the main text. The two tables below (Table 5, Table 6) provide the exact p-values underlying those correlations at full precision, allowing readers to apply their preferred multiple-comparison correction at their chosen significance level.

A.3.1 Why RSS aligns selectively with Real Tasks.

RSS measures holistic embedding fidelity: whether each cumulative synthesis faithfully represents all of its constituent atoms while remaining clearly separable from distractors. Synthetic Tasks (Needle, Passkey) reward the preservation of localized features and do not exercise this property. Real Tasks do: NarrativeQA requires the embedding to encode the narrative-structural relationships of an entire book, and SummScreenFD, QMSum, and WikimQA likewise probe how well an embedding represents documents in which multiple semantic paragraphs are organically bound (Appendix A.2). RSS's selective correlation with these four corpora thus follows directly from its mathematical definition.

Table 5: Detailed Pearson correlation coefficients and p-values for all metrics.

	$\mathbf{srs}_{\text{neg}}$		$\mathbf{srs}_{\text{pos}}$		$\mathbf{srs}_{\text{mean}}$		$\text{RSS}_{L=15}$	
Task	r	p	r	p	r	p	r	p
Synthetic Tasks								
Needle	0.3881	0.0233	0.3612	0.0358	0.2845	0.1029	0.3780	0.0275
Passkey	0.3359	0.0521	0.4381	0.0096	0.3328	0.0545	0.3928	0.0216
Real Tasks								
SummScreenFD	0.3002	0.0845	-0.0407	0.8190	-0.1497	0.3981	0.2229	0.2050
QMSum	0.4574	0.0065	-0.0753	0.6721	-0.1986	0.2603	0.6530	<0.0001
WikimQA	0.3919	0.0219	0.0020	0.9912	-0.1270	0.4741	0.6797	<0.0001
NarrativeQA	0.6412	<0.0001	-0.0505	0.7766	-0.2017	0.2527	0.7956	<0.0001
Summary (Average across all tasks)								
Average	0.5399	0.0010	0.1564	0.3769	0.0077	0.9657	0.6733	<0.0001

Table 6: Detailed Spearman correlation coefficients and p-values for all metrics.

	$\mathbf{srs}_{\text{neg}}$		$\mathbf{srs}_{\text{pos}}$		$\mathbf{srs}_{\text{mean}}$		$\text{RSS}_{L=15}$	
Task	ρ	p	ρ	p	ρ	p	ρ	p
Synthetic Tasks								
Needle	0.3271	0.0590	0.3134	0.0711	0.2580	0.1406	0.3160	0.0687
Passkey	0.3829	0.0254	0.6259	<0.0001	0.4451	0.0083	0.1416	0.4243
Real Tasks								
SummScreenFD	0.6193	<0.0001	-0.0506	0.7764	-0.2312	0.1884	0.7155	<0.0001
QMSum	0.5526	0.0007	0.0050	0.9774	-0.1224	0.4905	0.6486	<0.0001
WikimQA	0.4946	0.0029	0.0924	0.6031	-0.0824	0.6433	0.5899	0.0002
NarrativeQA	0.6333	<0.0001	0.1025	0.5640	-0.0237	0.8942	0.6975	<0.0001
Summary (Average across all tasks)								
Average	0.6807	<0.0001	0.2232	0.2044	0.0503	0.7777	0.6028	0.0002

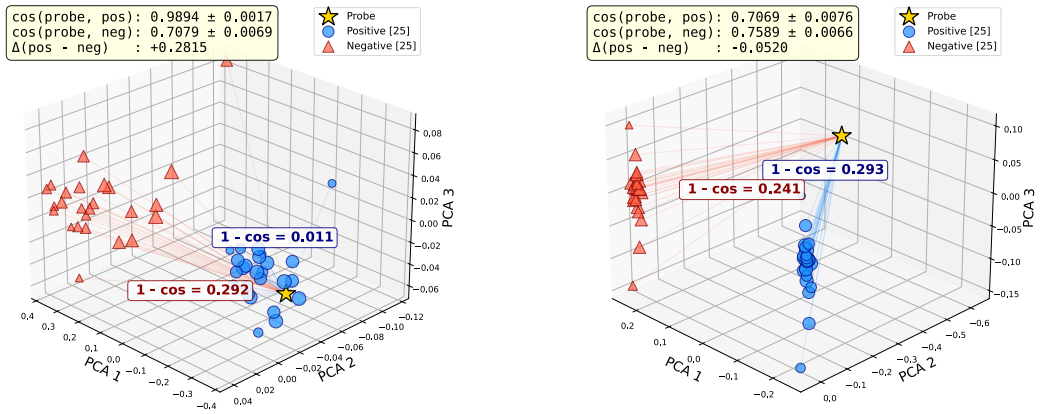
A.3.2 WHY THE NEGATIVE-SRS–REAL-TASK CORRELATION SURVIVES: THE TWO-CLUSTER HYPOTHESIS.

Section 4.2.2 established an asymmetry requiring explanation. The $\mathbf{srs}_{\text{pos}}$–Synthetic Task correlation follows directly from task definition: Passkey rewards literal matching and Needle rewards short-semantic matching, both of which are directly measured by probe-insertion similarity on SRS. The $\mathbf{srs}_{\text{neg}}$–Real Task correlation, however, resists such direct interpretation—holistic structural understanding has no obvious reason to favor negative direction of SRS. We propose a geometric explanation for this second half.

Consider the geometric relation between the probe vector v_p, the cluster C^+ formed by the Positive Group, and the cluster C^- formed by the Negative Group. Two regimes exist: *inclusion*, where v_p lies inside C^+, C^-, or both; and *exclusion*, where all three occupy independent regions of space (See Figure 4 for visualization of the C^+, C^- and v_p).

A positive SRS arises in two ways: (i) v_p lies inside C^+ while C^- is separated (Figure 4a), or (ii) all three are mutually exclusive but v_p happens to be closer to C^+ (Figure 4b). Case (ii) reflects genuine embedding resolution—the model distinguishes order. Case (i) reflects the opposite: the probe is buried inside the positive cluster, meaning the model ignores structural variation in the surrounding context and responds only to v_p itself. In aggregate on Real Tasks, positive scores mix the downstream-favorable case (ii) with the unfavorable case (i), and the correlation washes out.

A negative SRS likewise admits two cases: (i) v_p lies inside C^- while C^+ is separated, or (ii) all three are exclusive and v_p happens to be closer to C^-. Here the asymmetry emerges. Case (i)



(a) Probe buried inside the Positive

(b) Positive & Negative and probe, all separated

Figure 4: The "two-cluster and probe" hypothesis. (a) shows the *inclusion* regime; (b) shows the *exclusion* regime. PCA projection is used for 3D visualization only. Cosine distances shown are computed in the original space. Both are SRS$(l = 7, x = 0)$.

of negative SRS is implausible for a properly trained embedder: if $v_p =$ **ABC**, then C^- consists of dozens of context vectors of the form **(...,CBA,...)**—each carrying a reverse-permuted probe representation surrounded by noise—and the chance of **ABC** landing among them is low (the upper-left region in the quadrant analysis of Appendix B.1, 2.61% of cases). Negative scores therefore reflect almost exclusively case (ii), the regime favorable to embedding resolution, and the correlation with downstream performance survives. When v_p falls inside both clusters simultaneously (the upper-right region in the quadrant analysis of Appendix B.1, 28.69% of cases), the distance differences are negligible and SRS concentrates near zero, contributing to neither sign.

Empirically, a centroid-based nearest-neighbor analysis over the full experimental set confirms the asymmetry: across 22,950 probe-cluster measurements, the probe falls inside C^+ in 21.51% of cases but inside C^- in only 2.61%, with the remaining 75.87% falling outside both clusters or inside both simultaneously (Appendix B.1).

B Details for SRS

B.1 Strengthening the "Two-Cluster and Probe" Hypothesis via Nearest-Neighbor Analysis

B.1.1 Definition of the Centroid-based Rank-in-Cluster

To measure how far a probe vector (v_p) lies from the center of a given vector cluster (C)—that is, its outlier-ness—we define the following rank. A rank close to 1 indicates that the probe lies farthest from the center of the cluster, while a rank close to $N + 1$ indicates that it lies closest to the center.

First, all vectors are L2-normalized ($\hat{x} = x/||x||_2$). From the normalized vector set $\hat{C} = \{\hat{v}_1, ..., \hat{v}_N\}$ of a cluster $C = \{v_1, ..., v_N\}$, we compute the normalized centroid $\hat{c}$.

$$\hat{c} = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{v}_i}{\|\frac{1}{N}\sum_{i=1}^{N} \hat{v}_i\|_2}$$

On this basis, the rank of the probe vector v_p—the Centroid-based Rank-in-Cluster (CRC)—is computed as follows.

$$\text{CRC}(v_p, C) = 1 + \sum_{i=1}^{N} \mathbb{I}(\text{sim}(\hat{v}_i, \hat{c}) < \text{sim}(\hat{v}_p, \hat{c}))$$

The meaning of each term is as follows:

- $\text{sim}(a, b)$: the cosine similarity between two unit vectors a and b, which is equal to the dot product $a \cdot b$.
- $\mathbb{I}(\cdot)$: the indicator function, which returns 1 if the inner condition is true and 0 otherwise.

This formula determines the rank by counting the number of members within the cluster that lie farther from the centroid than the probe, and adding 1 to that count. That is, if the probe lies entirely outside the cluster, $\text{CRC}(v_p, C) = 1$. This provides a quantitative measure of where the probe is located within the cluster's distribution.

B.1.2 Connection to the SRS Computation Procedure

According to Section 3.3, SRS produces one vector and two vector clusters.

1. ($v_p = \mathbf{q}^{(l)}$):
$$\mathbf{q}^{(l)} = \mathcal{S}(c_1, c_2, \ldots, c_l)$$
2. ($C^+ = V_{\text{pos}}^{(l,x)}$):
$$V_{\text{pos}}^{(l,x)} = \{\mathcal{S}(m_1, \ldots, m_x, c_1, \ldots, c_l, m_{x+1}, \ldots, m_{L-l}) \mid \forall \mathbf{m} \in M\}$$
3. ($C^- = V_{\text{neg}}^{(l,x)}$):
$$V_{\text{neg}}^{(l,x)} = \{\mathcal{S}(m_1, \ldots, m_x, c_l, \ldots, c_1, m_{x+1}, \ldots, m_{L-l}) \mid \forall \mathbf{m} \in M\}$$

Therefore, once (model, repetition, l, x) is fixed, $\mathbf{q}$, V_{pos}, and V_{neg} are uniquely determined, and from these a final rank vector can be generated.

$$\mathbf{v}_{\text{rank}} = (\text{CRC}(v_p, V_{\text{pos}}), \text{CRC}(v_p, V_{\text{neg}}))$$

For the 34 models, $\text{SRS}(L = 8)$ is repeated 25 times, and the number of sub-tests (l, x) defined under the $\text{SRS}(L = 8)$ configuration is 27. This yields a total of 22,950 v_{rank} vectors. Since the sizes of the positive and negative context groups are fixed at 25 in our experiments, v_{rank} is defined over the range $(1, 1) \ldots (26, 26)$.

B.1.3 Application to the "Two-Cluster and Probe" Hypothesis

We first represent all 22,950 vectors in the range $(1, 1) \ldots (26, 26)$ as an accumulated 26x26 heatmap in Figure 5. Uncolored cells are missing values, meaning that no v_{rank} vector corresponding to that cell ever appeared.

We now split the two axes of the heatmap in half over the ranges 1-13 and 14-26, dividing it into a total of four blocks. The results can be summarized as follows.

Let us revisit the core premise of the "two-cluster and probe" hypothesis introduced in Section A.3.2:

- **Spatial Relations:** The relationship between the probe vector (v_p) and the positive/negative context clusters (C^+ and C^-) essentially falls into either an **inclusion relation** (v_p is buried inside the clusters) or an **exclusion relation** (all three entities occupy independent regions).
- **The Implausibility of Negative Inclusion:** While positive SRS scores can arise ambiguously from either relation, stable negative SRS scores are primarily linked to the **exclusion relation** (which signals a high-resolution embedding space). The alternative—where v_p is buried specifically inside C^-—is a pathological phenomenon highly unlikely to occur in a properly designed embedder.

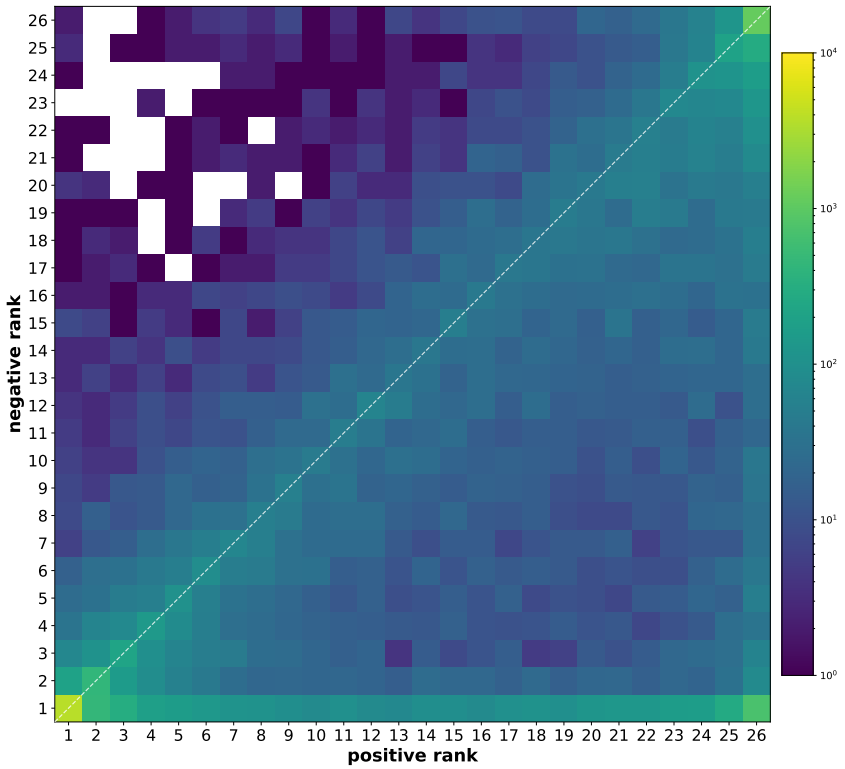


Figure 5: Joint heatmap of the full rank vectors across all 34 models and 25 repetitions. The distribution shows the frequency of probe positions relative to cluster centroids, with most vectors concentrated in the lower-left quadrant (*exclusion* regime).

Table 7: Quadrant analysis results

Quadrant	Vector Count	Proportion
lower-left	10,828	47.18%
upper-right	6,585	28.69%
lower-right	4,937	21.51%
upper-left	600	$\mathbf{2.61}\%$
total	22,950	100%

Consistent with this theoretical expectation, Table 7 shows that most of the v_{rank} vectors were located either in the lower-left quadrant (exclusion relation: the case where both clusters and the probe lie in independent regions of space) or in the upper-right quadrant (inclusion relation: the case where the probe lies simultaneously inside both the pos and the neg clusters)—47.18% and 28.69%, respectively. The lower-right quadrant (inclusion relation: the case where the probe is far from the negative cluster and inside the positive cluster) contained 21.51% of the v_{rank} vectors.

And across all v_{rank} vectors, only 2.61% were located in the upper-left quadrant (inclusion relation: the case where the probe is far from the positive cluster and inside the negative cluster). This quantitatively supports the claim that the probe-in-neg-cluster event occurs extremely rarely in properly designed embedders, and aligns exactly with the hypothesis in the main text.

Table 8: Wilcoxon signed-rank test results

Metric	Value
W-statistic	593.000
p-value	¡ 0.0001
Number of valid models (N_{valid})	34
Number of consistent models ($r_{\text{pos}} > r_{\text{neg}}$)	33 (97.1%)
Number of inconsistent models ($r_{\text{pos}} \leq r_{\text{neg}}$)	1 (2.9%)
Median r_{pos} (overall)	0.2333
Median r_{neg} (overall)	0.0230

B.1.4 Wilcoxon Signed-Rank Test for the Asymmetric Collapse Across Quadrants

To test whether the emptiness of the upper-left quadrant (probe-in-neg-cluster) relative to the lower-right quadrant (probe-in-pos-cluster)—as revealed by the quadrant analysis—is statistically significant, we performed a Wilcoxon signed-rank test on the two quadrants.

For the distribution of the per-model differences $r_{\text{pos}}^{(m)} - r_{\text{neg}}^{(m)}$,

- Null hypothesis H_0: the distribution is symmetric around 0.
- Alternative hypothesis H_1: the distribution is skewed toward the positive direction.

That is, the hypothesis is that, on a per-model basis as well, probe-in-pos-cluster appears more often than probe-in-neg-cluster.

For each model m, we define two test statistics as follows. Here N_m is the total number of trials for model m, and $\mathbb{I}(\cdot)$ is the indicator function.

$$r_{\text{pos}}^{(m)} = \frac{1}{N_m} \sum_{i=1}^{N_m} \mathbb{I}\left(\text{rank}_{pos}^{(m,i)} > 13 \wedge \text{rank}_{neg}^{(m,i)} \leq 13\right)$$

$$r_{\text{neg}}^{(m)} = \frac{1}{N_m} \sum_{i=1}^{N_m} \mathbb{I}\left(\text{rank}_{pos}^{(m,i)} \leq 13 \wedge \text{rank}_{neg}^{(m,i)} > 13\right)$$

As shown in Table 8, the test results reject the null hypothesis at $p < 0.0001$. In 33 of the 34 models (97.1%), probe-in-pos-cluster consistently appeared more often than probe-in-neg-cluster. In terms of the median ratio, r_{pos} (23.33%) is about 10.2 times higher than r_{neg} (2.30%).

B.1.5 Analysis of the Inconsistent and Borderline Models

Table 9: Per-model comparison of r_{pos} and r_{neg}

Model	r_{pos}	r_{neg}	$r_{pos} - r_{neg}$
llama-embed-nemotron-8b	0.010370	0.013333	-0.002963
nomic-embed-text-v1.5	0.004444	0.001481	0.002963
snowflake-arctic-embed-l-v2.0	0.044444	0.008889	0.035556
Qwen3-Embedding-4B	0.050370	0.005926	0.044444
MoD-Embedding	0.099259	0.037037	0.062222
F2LLM-0.6B	0.128889	0.000000	0.128889

As shown in Table 9, the only model for which $r_{\text{pos}} \leq r_{\text{neg}}$ in the Wilcoxon test was **llama-embed-nemotron-8b**. To account for this inconsistency, we analyze r_{pos} and r_{neg} for the six models with the strongest probe-cluster separation tendency, ranked by $(r_{\text{pos}} - r_{\text{neg}})$.

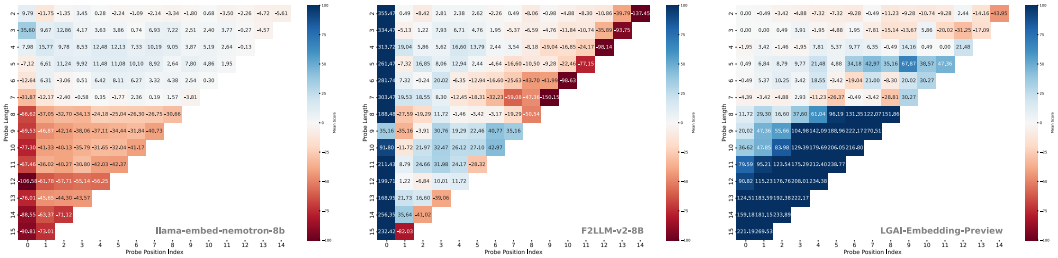
For **llama-embed-nemotron-8b**, $(r_{\text{pos}} - r_{\text{neg}})$ is the only negative value across all models, yet its absolute value is very small—about 1/12 of that of **snowflake-arctic-embed-l-v2.0**, the third-lowest model by $(r_{\text{pos}} - r_{\text{neg}})$. In particular, r_{pos} and r_{neg} for this model are about 1/22.6 and 1/1.7 of the overall medians of r_{pos} and r_{neg}, respectively (Table 8). This indicates that **llama-embed-nemotron-8b** tends not to induce the very situations that r_{pos} and r_{neg} presuppose—in other words, it tends not to induce either probe-in-neg-cluster or probe-in-pos-cluster, meaning it tends not to place the probe vector inside any cluster at all. This probe-cluster separation tendency may be the reason that **llama-embed-nemotron-8b**, though an inconsistent model, nonetheless shows strong downstream capability (LEMB Avg, 0.815).

However, there is another model in which this tendency is even more pronounced than in **llama-embed-nemotron-8b**: **nomic-embed-text-v1.5**. Compared to the overall medians of r_{pos} and r_{neg}, this model exhibits extremely small values of r_{pos} and r_{neg}—about 1/52.8 and 1/15.5, respectively. That is, this model has an even stronger probe-cluster separation tendency than **llama-embed-nemotron-8b**. Yet, as introduced in Section 4.2.1, this model shows a much lower score on downstream evaluations such as LEMB (LEMB Avg, 0.344) than **llama-embed-nemotron-8b**.

What must be noted here is that, while the two models point in the same direction under the quadrant analysis (a strong probe-cluster separation tendency), they point in opposite directions under RSS. **llama-embed-nemotron-8b** is stably located in the positive region with an RSS Avg. of 236.16, whereas **nomic-embed-text-v1.5** is, as reported in Section 4.2.1, the only model among the 34 that exhibits a negative RSS in the long-context regime ($L > 12$).

This observation does not negate the "two-cluster and probe" hypothesis; rather, it clarifies its scope of application. A probe-cluster separation tendency is compatible with strong LEMB performance, but is not sufficient on its own. The quadrant analysis and RSS describe different phenomena in the embedding space, and only models that are sound under both metrics exhibit strong downstream performance. **nomic-embed-text-v1.5** serves as a single-case piece of evidence for the orthogonality of these two metrics, and shows that the quadrant analysis is not a replacement for RSS but a complementary diagnostic tool. We leave the analysis of the integrated geometric mechanism differences between SRS and RSS for these two models to future work.

B.2 Additional Experiments on SRS(L=16)



(a) llama-embed-nemotron-8b (b) F2LLM-v2-8B (c) LGAI-Embedding-Preview

(d) jina-embeddings-v5-text-nano (e) snowflake-arctic-embed-l-v2.0 (f) Qwen3-Embedding-0.6B

Figure 6: SRS heatmaps for six representative models at $L = 16$ (8000 token context).

We analyze the sub-test heatmaps of six representative models shown in Figure 6 (**LGAI-Embedding-Preview**, **jina-embeddings-v5-text-nano**, **llama-embed-nemotron-8b**, **F2LLM-v2-8B**, **snowflake-arctic-embed-l-v2.0**, **Qwen3-Embedding-0.6B**) at SRS($L = 16$), with chunk size as 500 tokens—that is, over a range of 8000 tokens. Here, however, we use the output of a single run rather than of 25 repetitions. Unlike in the main text, cells whose values fail the t-test for significance against zero are not hatched. The experimental environment and resource consumption for this experiment differ from those of the main experiment; for specifics, see Appendix F.4.

For **llama-embed-nemotron-8b**, no clear tendency is observed in SRS(l, x) when l ranges from 2 to 7, but for l from 8 to 15, every cell is negative regardless of x. This is consistent with the tendency we observed at SRS($L = 8$)—values without a clear tendency at low l, and strongly negative values at high l.

Interesting phenomena appeared in **F2LLM-v2-8B** and **LGAI-Embedding-Preview**. **F2LLM-v2-8B**, in addition to the strong positive SRS at $x = 0$ described earlier, showed large negative values at the very last x positions for each probe length l, at cells such as SRS($7, 10$), SRS($6, 11$), ..., SRS($2, 14$). This tendency, however, did not appear clearly for $l \geq 8$. Conversely, **LGAI-Embedding-Preview**, as described earlier, showed strong positive values at SRS($15, 1$), SRS($14, 2$), ..., SRS($8, 8$). Moreover, at SRS($L = 16$), it exhibited a kind of "thick positive diagonal" tendency, with strong positive values appearing not only at the very last x positions but also across the last 4 to 5 chunks. Yet this tendency disappeared entirely for sub-tests with $l \leq 7$, leaving no pattern at all. The two models thus exhibit pattern-loss tendencies at different positions.

For **jina-embeddings-v5-text-nano**, most cells were positive. The maximum was 76.17 and the minimum was -23.19. The strong positive signature at $x = 0$ and $x = 1$ appears to disappear at SRS($L = 16$). By contrast, **snowflake-arctic-embed-l-v2.0** retained its characteristic of having only negligible values: the maximum was 11.72 and the minimum was -9.28.

Qwen3-Embedding-0.6B did not lose its strong positive tendency at $x = 0$. The maximum was 336.43, recorded at SRS($11, 0$). No other notable features were observed, though a tendency for some cells to hold negative values in the tens (below -40) was also noticed. This tendency will need to be statistically verified through repeated experiments.

Overall, even at SRS($L = 16$), the idiosyncratic behaviors of the models were preserved; however, in some models the tendencies were either strengthened or lost within certain ranges of l and x (**F2LLM-v2-8B**, **LGAI-Embedding-Preview**), and in others the idiosyncratic pattern disappeared altogether (**jina-embeddings-v5-text-nano**). This suggests the need for larger-scale studies of RSS and SRS patterns at high L.

C Details for RSS

C.1 Full Correlation of RSS-LEMB Scores

Table 10 presents the full per-length Pearson and Spearman correlation coefficients between RSS($L = x$) and the LEMB average score, complementing the visual summary in Figure 1. The general increase from $L = 4$ ($r = 0.48$) to $L = 15$ ($r = 0.67$) is discussed in Section 4.2.1.

C.2 Cross-Corpus Comparison of RSS(L=15) Scores

For all 34 models used in this study, we measured RSS($L = 15$) after switching the corpus from **CulturaX** (Nguyen et al., 2024; uonlp) to the **Wikipedia** (Wikimedia Foundation; wikimedia) dataset. We then measured the correlation between the RSS($L = 15$) values obtained on **CulturaX** and those obtained on **Wikipedia**. For **Wikipedia**, we used all of the data in the **en** split (**20231101.en/train-000*-of-00041.parquet**).

(Note: **mean(ml-e5-base)** is excluded from the correlation analysis, as it is not a single model but a synthesis strategy for control purposes.)

As a result, the correlation shows in Figure 7 was very high: Pearson $r = 0.966$ and Spearman $\rho = 0.929$ (both at $p < 10^{-6}$). The fact that the models in this study exhibit such a high correlation

Table 10: Per-length Pearson and Spearman correlations between RSS($L = x$) and the LEMB average score, corresponding to the "Average" curve in Figure 1.

RSS($L = x$)	$r_{pearson}$	$p_{pearson}$	$\rho_{spearman}$	$p_{spearman}$
$L = 4$	0.4817	0.0039	0.4298	0.0112
$L = 5$	0.4212	0.0131	0.4377	0.0096
$L = 6$	0.5048	0.0023	0.4973	0.0028
$L = 7$	0.4786	0.0042	0.4509	0.0074
$L = 8$	0.5065	0.0022	0.4808	0.0040
$L = 9$	0.5622	0.0005	0.5481	0.0008
$L = 10$	0.5736	0.0004	0.5172	0.0017
$L = 11$	0.5874	0.0003	0.5346	0.0011
$L = 12$	0.6304	¡ 0.0001	0.5652	0.0005
$L = 13$	0.6186	¡ 0.0001	0.5661	0.0005
$L = 14$	0.6593	¡ 0.0001	0.5792	0.0003
$L = 15$	**0.6733**	¡ 0.0001	**0.6028**	0.0002
$L = 16$	0.6352	¡ 0.0001	0.5367	0.0011

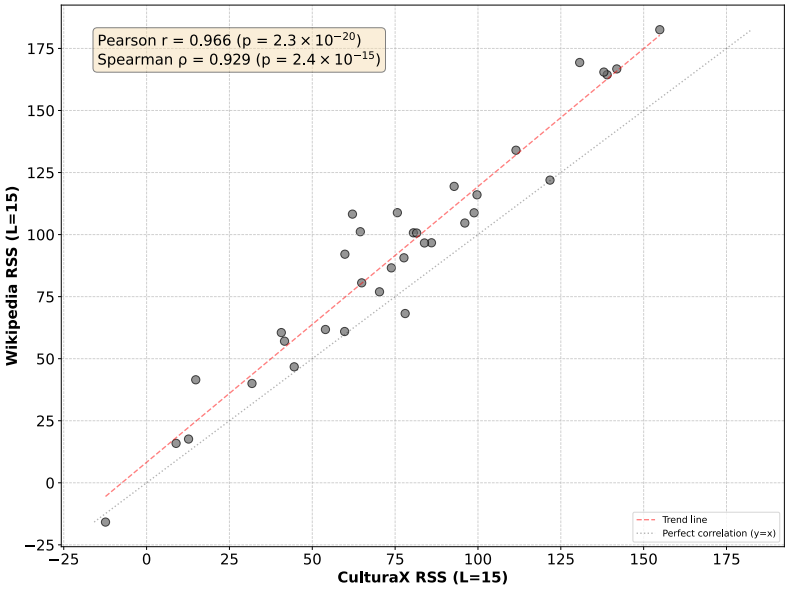


Figure 7: Scatter plot of the correlation between RSS($L = 15$) scores of the 34 models measured on the CulturaX and Wikipedia datasets.

between two RSS($L = 15$) values computed on different domains suggests that the distributional characteristics of the RSS scores measured by FINESSE may not be confined to a specific domain.

A point of interest is the slope of the trend line. Most models received higher scores on **Wikipedia** than on **CulturaX**, and the ranks of some models even changed. Taking this variation into account, the variability of RSS scores across domains will make an interesting subject for future research.

C.3 Correlation Between the Components of the RSS Score and the RSS Score Itself

In Section 3.2, the final RSS score is computed as RSS $= 500 \times (\text{TD} + \text{BU})$. It is therefore essential to analyze how the TD and BU scores correlate with the final RSS score.

Both the TD-RSS and BU-RSS correlations are significant at $p < 0.0001$ across all sequence lengths, with Pearson and Spearman coefficients consistently above 0.9 (Table 11). Averaged over

Table 11: Correlation analysis among TD, BU, and the composite RSS score, across sequence lengths L. All correlations are significant at $p < 0.0001$ ($N = 34$ models, without baseline).

	TD vs. RSS		BU vs. RSS		TD vs. BU	
L	Pearson r	Spearman ρ	Pearson r	Spearman ρ	Pearson r	Spearman ρ
4	0.9694	0.9539	0.9579	0.9383	0.8580	0.8151
5	0.9777	0.9532	0.9732	0.9713	0.9031	0.8851
6	0.9779	0.9725	0.9712	0.9560	0.9000	0.8762
7	0.9799	0.9719	0.9752	0.9716	0.9116	0.9092
8	0.9801	0.9716	0.9744	0.9594	0.9104	0.8921
9	0.9832	0.9704	0.9805	0.9752	0.9281	0.9160
10	0.9713	0.9685	0.9695	0.9780	0.8835	0.9190
11	0.9646	0.9639	0.9579	0.9676	0.8483	0.8982
12	0.9608	0.9697	0.9609	0.9670	0.8464	0.9007
13	0.9544	0.9639	0.9543	0.9697	0.8217	0.8937
14	0.9567	0.9731	0.9576	0.9752	0.8322	0.9123
15	0.9568	0.9737	0.9567	0.9591	0.8306	0.8924
16	0.9403	0.9490	0.9352	0.9343	0.7588	0.8020
Mean	0.9672	0.9658	0.9634	0.9633	0.8640	0.8855

$L = 4, \ldots, 16$, the mean correlations are 0.97 (Pearson) and 0.97 (Spearman) for TD-RSS, versus 0.96 and 0.96 for BU-RSS. Both TD and BU are thus core components of the RSS score.

Meanwhile, the TD-BU correlation ranges from 0.75 to 0.93 (Pearson) and from 0.80 to 0.92 (Spearman), remaining positive across all levels yet never reaching full linear dependence. In other words, TD and BU share a common construct while each retains its own distinctive variance. Conceptually, TD measures whether the embedding of a text faithfully represents its constituent atoms (top-down), whereas BU measures whether the embedding of each atom contributes appropriately to the embedding of its containing text (bottom-up). The directional asymmetry between these two perspectives is what gives rise to the distinctive variance each component carries.

The formula RSS = TD + BU is therefore methodologically justified: only when the two are combined does RSS become a holistic indicator that neither component can provide on its own. This correlation analysis empirically confirms that the sum of the two elements possesses broader diagnostic power than either individual element.

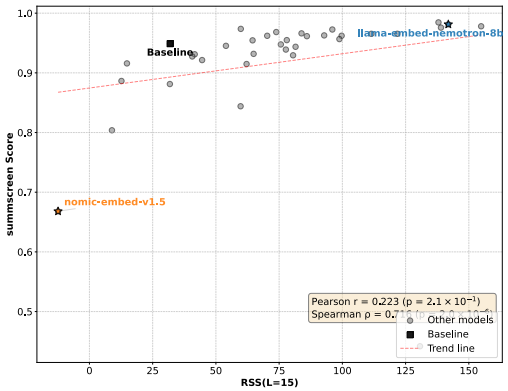
C.4 Additional Analysis of the Correlation Between RSS Scores and the LEMB SummScreenFD Task

Table 12: Score mean and IQR (Q3 - Q1)

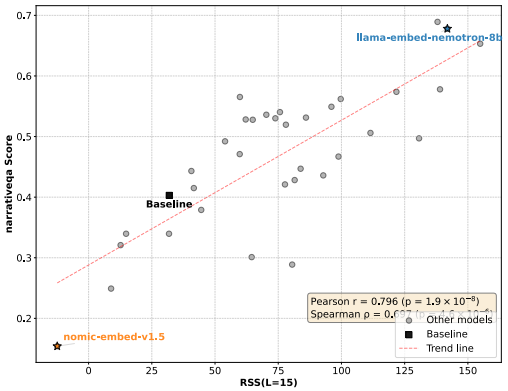
Task	n	Mean	IQR
Needle	34	0.5799	0.1683
Passkey	34	0.9095	0.1117
SummScreenFD	34	0.9182	0.0426
QMSum	34	0.4033	0.0815
WikimQA	34	0.7790	0.1209
NarrativeQA	34	0.4693	0.1228

In Section 4.2.1 (Figure 1), we noted a discrepancy: of the two correlation coefficients between LEMB SummScreenFD and RSS—Pearson and Spearman—only the Spearman correlation was notably high and significant. Here, we analyze the reason.

In the scatter plots (Figure 8), although **mean(ml-e5-base)** was not used in the correlation analysis, we mark only its position with a '*'. This control model has RSS$(L = 15) = 31.88$, placing it 6th from the bottom among the 34 experimental models.



(a) Scatter plot of RSS($L = 15$) versus LEMB SummScreenFD



(b) Scatter plot of RSS($L = 15$) versus LEMB NarrativeQA

Figure 8: Comparison of RSS($L = 15$) correlations with LEMB tasks. (a) SummScreenFD shows weaker Pearson correlation due to narrow score IQR (0.0426), though Spearman remains strong (0.716). (b) NarrativeQA exhibits the strongest correlation (Pearson 0.796) among all tasks.

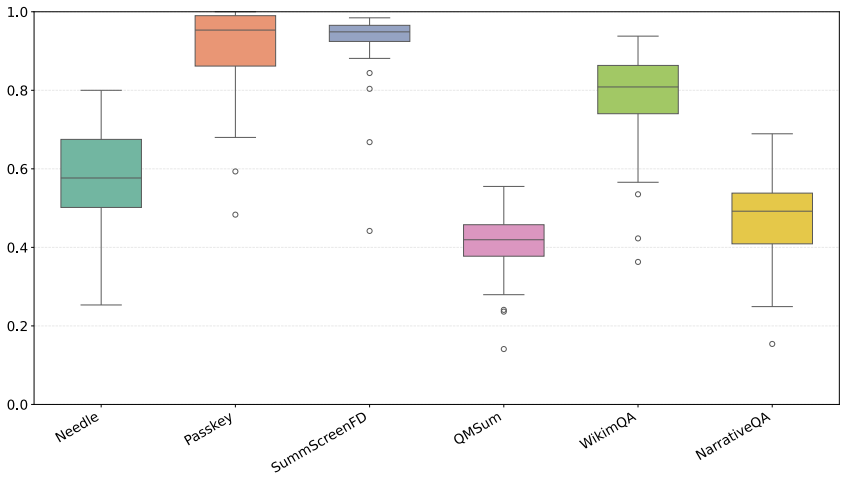


Figure 9: Boxplot of the score distributions of 34 models across 6 LEMB tasks. SummScreenFD exhibits the highest mean (0.918) but narrowest IQR (0.043), while NarrativeQA shows the 2nd spread (IQR 0.123).

Looking at the scatter plots together with the LEMB-score boxplots (Figure 9) for NarrativeQA and SummScreenFD, a substantial difference becomes apparent. The IQR of SummScreenFD scores is 0.0426—the lowest of all—whereas the IQR of NarrativeQA scores is 0.1228, the highest among the four Real Tasks (Table 12). The narrow IQR of SummScreenFD scores, together with its being the highest mean across all Tasks ($mean = 0.9182$), can be inferred to have made it difficult for the Pearson correlation—which considers linear relationships between variables—to hold. Nevertheless, because the Spearman correlation of SummScreenFD scores with RSS($L = 15$) is sufficiently high ($\rho = 0.716$) and significant ($p < 0.001$), we regard the correlation between RSS and the LEMB SummScreenFD Task as meaningful.

C.5 Additional Experiments on RSS(L = 8, 16, ..., 64)

Among our pool of 34 models, 11 support a 32,000-token context window(see Appendix F.1). For these, we conducted an exploratory investigation of RSS over $L \in \{4, 8, \cdots, 64\}$ — a fourfold wider range than the $L \in \{4, \cdots, 16\}$ used in the main experiments. The table above presents a subset of these results, $L \in \{8, 16, \cdots, 64\}$; the full results archived in Appendix F.3.

Table 13: RSS scores across sequence lengths for the $L = 64$ capable 11 models. Cell shading indicates relative magnitude within each column (darker = higher RSS)

Model	L=8	L=16	L=24	L=32	L=40	L=48	L=56	L=64
LGAI-Embedding-Preview	146.21	74.43	38.47	21.36	5.56	-0.28	-6.95	-10.59
MoD-Embedding	70.38	68.31	47.40	25.76	22.47	8.66	2.06	-6.50
Octen-Embedding-0.6B	89.46	66.40	50.65	31.90	26.93	16.72	7.52	4.19
Octen-Embedding-4B	129.10	88.84	58.95	33.13	28.25	13.13	6.53	-2.43
Octen-Embedding-8B	132.13	98.93	69.36	47.78	38.77	25.60	17.08	7.80
BOOM_4B_v1	84.61	52.57	38.51	26.90	23.34	14.89	11.93	13.50
jina-embeddings-v5-text-small	109.03	50.66	28.76	15.70	10.65	0.52	-4.72	-7.12
llama-embed-nemotron-8b	254.07	123.11	75.50	51.34	37.83	21.78	10.09	3.03
Qwen3-Embedding-0.6B	35.64	34.78	29.44	22.27	19.32	13.39	7.63	4.13
Qwen3-Embedding-4B	63.58	57.88	35.27	15.33	12.46	-0.18	-2.45	-11.23
Qwen3-Embedding-8B	79.95	61.09	30.64	15.02	9.96	-1.54	-5.91	-16.07
mean(ml-e5-base)	58.45	29.74	20.43	13.78	10.32	7.01	5.08	2.69

Across the 11 models, RSS decreases with L in nearly all cases, in agreement with the monotonic degradation we argued for $L \in \{4, \cdots, 16\}$ in Appendix E.4. Some models cross into negative RSS as L grows (e.g., **Qwen3-Embedding-4B**, **Qwen3-Embedding-8B**, **LGAI-Embedding-Preview**), while others do not (e.g., **llama-embed-nemotron-8b**, **BOOM_4B_v1**, **Octen-Embedding-8B**). Notably, models indistinguishable at $L = 8$ (e.g., **BOOM_4B_v1** at 84.61 vs. **Qwen3-Embedding-8B** at 79.95) diverge markedly in their degradation rates: the former remains positive through $L = 64$, while the latter crosses zero by $L = 48$. The positive-to-negative transition point of RSS is therefore likely to be a fruitful object of study for future work.

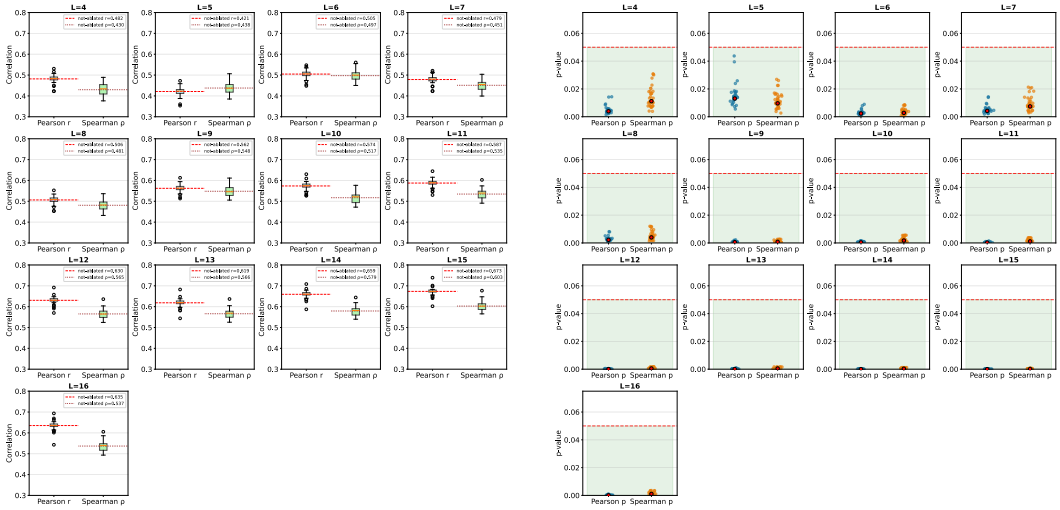
It should be emphasized that our primary experiments were conducted within an $8{,}190$-token context window (see Section 4.1); accordingly, we cannot infer how the distinctive RSS behavior at high L found in this exploratory investigation relates to downstream, practical performance. Nevertheless, capturing the model-specific RSS trajectories that emerge at low and high L is, we believe, sufficiently meaningful on its own.

D Ablation Experiments for the Correlation Analysis

D.1 Jackknife Ablation of the 34 Models on the Correlation Between RSS/SRS and LEMB

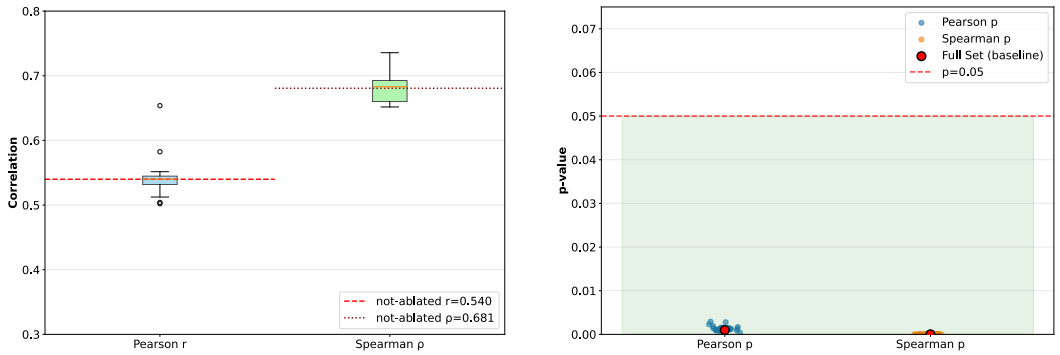
To verify the robustness of the Pearson and Spearman correlation coefficients and their significance for RSS across $L = 4, \ldots, 16$ as described in Section 4.2.2 and Table 3, we conducted a total of 34x13 ablations in which each of the 34 models was excluded one at a time, observing how the correlation changed. Figure 10a shows the boxplot of the Pearson and Spearman correlation coefficients across all 34×13 data points. Figure 10b shows the distribution of the p-values from the same ablation, with the green region indicating the significance zone $p < 0.05$.

See the Q1-Q3 (box region) in Figure 10a. The Q1-Q3 width does not exceed 0.05 for Spearman, and does not exceed 0.02 for Pearson. This shows that the correlation found in the present study is not heavily skewed by any particular model. Moreover, the baseline computed on the full model set (the red line) almost perfectly coincides with the median of the jackknife distribution. This indicates that the full model set reflects the underlying data trend without sampling bias. In addition, the trajectory of the correlation across context lengths (building toward a peak from $L = 8$ to $L = 15$, then dropping at $L = 16$) is observed identically in the jackknife distribution as well, regardless of which individual model is included.



(a) correlation coefficient distribution (b) significance level distribution

Figure 10: Jackknife ablation of RSS-LEMB correlations across 34 models. (a) Boxplots show stable correlation coefficients with Q1-Q3 width < 0.05 for Spearman and Pearson. (b) All data points maintain significance at $p < 0.05$ (green zone), with most at $p < 0.01$ from $L = 9$ to $L = 16$.



(a) correlation coefficient distribution (b) significance level distribution

Figure 11: Jackknife ablation of $\mathbf{srs}_{\text{neg}}$-LEMB correlations. (a) Boxplots show stable correlation coefficients with Q1-Q3 width < 0.05 for Spearman and Pearson. (b) All p-values remain below 0.01.

In Figure 10b, consider the distribution of the p-value scatter. First, every data point was significant ($p < 0.05$). Furthermore, from $L = 9$ to $L = 16$, all jackknife data points except for 1-2 maintained $p < 0.01$.

Similarly, to evaluate the robustness of the correlation between the mean absolute value of the $\mathbf{srs}_{\text{neg}}$ scores and the LEMB scores reported in Section 4.2.2, we performed another 34 jackknife ablations. In Figure 11a, the Q1-Q3 width did not exceed 0.05 for either Spearman or Pearson, and the correlation coefficient computed on the full model set coincided with the median of the jackknife distribution. Looking at Figure 11b, both the Spearman and Pearson p-values are stable at below 0.01.

Taking these results together, we judge that the correlation of RSS and SRS with LEMB is robustly preserved across all 34 models.

Table 14: Changes in correlation coefficients and significance when **nomic-embed-text-v1.5** is excluded.

RSS($L = x$)	$r_{pearson}$	$p_{pearson}$	$\rho_{spearman}$	$p_{spearman}$
$L = 4$	0.4879	0.0040	0.3954	0.0228
$L = 5$	0.4403	0.0103	0.4134	0.0168
$L = 6$	0.5341	0.0014	0.4796	0.0047
$L = 7$	0.4925	0.0036	0.4235	0.0141
$L = 8$	0.5139	0.0022	0.4442	0.0096
$L = 9$	0.5475	0.0010	0.5060	0.0027
$L = 10$	0.5270	0.0016	0.4719	0.0056
$L = 11$	0.5301	0.0015	0.4910	0.0037
$L = 12$	0.5698	0.0005	0.5244	0.0017
$L = 13$	0.5439	0.0011	0.5254	0.0017
$L = 14$	0.5872	0.0003	0.5398	0.0012
$L = 15$	0.6021	0.0002	0.5655	0.0006
$L = 16$	0.5429	0.0011	0.4933	0.0035
$\mathbf{srs}_{\text{neg}}$	0.6537	¡ 0.0001	0.7129	¡ 0.0001

D.2 Investigating the Influence of **nomic-embed-text-v1.5** on the Correlation Between RSS/SRS and LEMB

During the jackknife investigation, however, we found that excluding **nomic-embed-text-v1.5** changes the correlation and significance of both SRS and RSS more than the exclusion of any other model. Let us first examine in detail the result of excluding **nomic-embed-text-v1.5** from the RSS correlation.

In Table 14, both the correlations and their significance have decreased, compared to the original RSS $L = 4, \ldots, 16$ vs. LEMB correlation (Table 3, Appendix Table 10). Nevertheless, the correlation coefficients still lie between 0.4 and 0.6, and all p-values remain below $p < 0.05$. As in the main text, the most significant case was (RSS($L = 15$), LEMB).

The significance of both the Pearson and Spearman correlations here is $p = 0.0002$ and $p = 0.0006$ respectively, which is sufficient to be significant ($p < 0.05$). Excluding **nomic-embed-text-v1.5** therefore does not undermine the correlation.

The effect of **nomic-embed-text-v1.5** on the $\mathbf{srs}_{\text{neg}}$-LEMB correlation is the opposite. The Pearson correlation coefficient rose further to 0.6537, and the Spearman correlation likewise rose ($\rho = 0.7129$). Both correlations became even more significant, at $p < 0.0001$. Viewed alongside Appendix B.1, a rather interesting connection emerges: this model was singled out in Appendix B.1 as an exceptional case under the "two-cluster and probe" hypothesis that interprets the $\mathbf{srs}_{\text{neg}}$-LEMB correlation. In other words, the phenomenon that was noted as a case study in Appendix B.1 has now manifested as a numerical change in the actual ablation experiment on the correlation.

In conclusion, although this model influences the correlation between RSS/$\mathbf{srs}_{\text{neg}}$ and LEMB more than any other model, we judge that the correlation does not depend excessively on this single model.

D.3 Investigating the Influence of F2LLM-Family Models on the Correlation Between RSS/SRS and LEMB

F2LLM-family models account for 7 out of our 34 experimental models, roughly 20% [5]. This raises a concern about potential data bias from a specific model family. We therefore re-examine the correlations and significance after removing them. One caveat in interpretation is that, in this case, the sample size drops substantially from $n = 34$ to $n = 27$. It is difficult to distinguish the increased statistical instability caused by the smaller sample size from actual F2LLM-driven bias.

[5] the other three F2LLM-v2 variants—80M, 160M, 330M—failed to run; see Appendix F.1

Table 15: Changes in correlation coefficients and significance when F2LLM-family models are excluded

RSS($L = x$)	$r_{pearson}$	$p_{pearson}$	$\rho_{spearman}$	$p_{spearman}$
$L = 4$	0.3531	0.0708	0.2021	0.3121
$L = 5$	0.2987	0.1302	0.2216	0.2666
$L = 6$	0.3694	0.0579	0.2552	0.1989
$L = 7$	0.3625	0.0631	0.2204	0.2693
$L = 8$	0.3929	0.0426	0.2308	0.2468
$L = 9$	0.4619	0.0153	0.3156	0.1088
$L = 10$	0.4811	0.0111	0.2875	0.1458
$L = 11$	0.5106	0.0065	0.3083	0.1177
$L = 12$	0.5633	0.0022	0.3376	0.0850
$L = 13$	0.5609	0.0023	0.3565	0.0679
$L = 14$	0.6148	0.0006	0.3748	0.0540
$L = 15$	0.6364	0.0004	0.3901	0.0443
$L = 16$	0.5900	0.0012	0.3107	0.1147
$\mathbf{srs}_{\text{neg}}$	0.4242	0.0274	0.4927	0.0090

In Table 15, compared to Table 3 and Appendix Table 10, the full removal of the F2LLM model family appears to weaken the RSS-LEMB correlation more substantially. The weakening of Spearman significance is particularly pronounced. Nonetheless, if we take the correlation at RSS($L = 15$) as our reference, the Pearson significance is 0.0004—still highly significant—and the correlation is high at 0.6364.

The SRS responds more strongly to the removal of F2LLM-family models. As a result of the removal, the Pearson coefficient changed from 0.5399 to 0.4242, the Pearson p-value from 0.0010 to 0.0274, the Spearman coefficient from 0.6807 to 0.4927, and the Spearman p-value from below 0.0001 to 0.0090. Both correlations remained significant after the removal ($p < 0.05$), but the magnitudes of the correlations decreased. Even so, both correlation coefficients still remained within the $0.4 \sim 0.6$ range.

In conclusion, while the full removal of this model family lowers the correlations between RSS, SRS and LEMB, even under such a harsh condition—the complete removal of the entire model family—the correlations of RSS and SRS still survive.

D.4 Quantitative Dimension-Controlled Analysis for RSS

Table 16: Zero-order correlations between RSS/SRS, embedding dimension, and LEMB.

Metric	n	$r(X, \text{LEMB})$	p	$r(X, \text{dim})$	p	$r(\text{dim}, \text{LEMB})$	p
RSS L=4	34	+0.4817	0.0039	+0.2147	0.2228	+0.4807	0.0040
RSS L=5	34	+0.4212	0.0131	+0.2677	0.1259	+0.4807	0.0040
RSS L=6	34	+0.5048	0.0023	+0.2740	0.1168	+0.4807	0.0040
RSS L=7	34	+0.4786	0.0042	+0.3159	0.0688	+0.4807	0.0040
RSS L=8	34	+0.5065	0.0022	+0.2770	0.1127	+0.4807	0.0040
RSS L=9	34	+0.5622	0.0005	+0.3024	0.0821	+0.4807	0.0040
RSS L=10	34	+0.5736	0.0004	+0.3485	0.0434	+0.4807	0.0040
RSS L=11	34	+0.5874	0.0003	+0.3247	0.0610	+0.4807	0.0040
RSS L=12	34	+0.6304	<0.0001	+0.3495	0.0428	+0.4807	0.0040
RSS L=13	34	+0.6186	<0.0001	+0.3272	0.0589	+0.4807	0.0040
RSS L=14	34	+0.6593	<0.0001	+0.3211	0.0641	+0.4807	0.0040
RSS L=15	34	+0.6733	<0.0001	+0.3398	0.0493	+0.4807	0.0040
RSS L=16	34	+0.6352	<0.0001	+0.3207	0.0644	+0.4807	0.0040
SRS_{neg}	34	+0.5399	0.0010	+0.4073	0.0168	+0.4807	0.0040

Table 17: Dimension-controlled partial correlations between RSS/SRS and LEMB, controlling for model-wise embedding dimension.

Metric	n	$r_{\text{partial}}(P)$	p	$r_{\text{partial}}(S)$	p
RSS L=4	34	+0.4419	0.0100	+0.3899	0.0249
RSS L=5	34	+0.3462	0.0484	+0.3342	0.0573
RSS L=6	34	+0.4424	0.0099	+0.4250	0.0137
RSS L=7	34	+0.3928	0.0237	+0.3647	0.0369
RSS L=8	34	+0.4431	0.0098	+0.4347	0.0115
RSS L=9	34	+0.4987	0.0031	+0.4722	0.0055
RSS L=10	34	+0.4940	0.0035	+0.4322	0.0120
RSS L=11	34	+0.5201	0.0019	+0.4655	0.0063
RSS L=12	34	+0.5628	0.0007	+0.4954	0.0034
RSS L=13	34	+0.5567	0.0008	+0.5015	0.0029
RSS L=14	34	+0.6080	0.0002	+0.5255	0.0017
RSS L=15	34	+0.6183	0.0001	+0.5469	0.0010
RSS L=16	34	+0.5791	0.0004	+0.4733	0.0054
SRS_{neg}	34	+0.4297	0.0126	+0.5214	0.0019

FINESSE never compares raw similarity scores directly across models of differing embedding dimension. Nevertheless, because the distribution of cosine similarities can depend on dimensionality, there remains the possibility that the two metrics merely reflect differences in embedding dimension.

The 34 models span output dimensions from 384 to 4096. We computed Pearson and Spearman partial correlations between RSS and LEMB while controlling for model-wise dimension as a covariate (Table 17; each point at $L = 4, \ldots, 16$). Pearson partial correlations remain significant throughout (all $p < 0.05$; at $L = 15$, $r_{\text{partial}} = 0.618$, $p = 0.0001$). Spearman partial correlations remain significant at every length except $L = 5$ ($p = 0.057$); at $L = 15$, $r = 0.547$, $p = 0.001$. Dimension is itself associated with LEMB ($r(\dim, \text{LEMB}) = 0.481$, $p = 0.004$). The decisive point, however, is that the RSS–LEMB correlation persists after controlling for dimension. This means RSS does not reach LEMB merely through dimension: it shares dimension-independent information with LEMB. Moreover, the zero-order correlation between RSS and dimension is non-significant at 10 of the 13 RSS lengths ($p > 0.05$), and small even where significant (e.g., $r = 0.340$, $p = 0.049$ at $L = 15$).

Beyond partial control, the hypothesis that dimension determines RSS is tested directly by examining conditional variance. If dimension determined RSS, scores could not diverge widely among models sharing the same dimension.

Among ten 4096-dimensional models, RSS($L = 4$) ranges from 169.31 to 456.64 (span 287.33). Because this group mixes model families—Llama (Nemotron), Qwen3, and F2LLM—part of this variance could arise from base-model differences rather than dimension. Restricting the analysis to derivatives of the same ‘e5-mistral-7b-instruct‘ base (all dimension 4096) still yields an RSS($L = 4$) range of 169.31–334.09 (span 164.78). Substantive differences in compositional behavior therefore persist even when both dimension and base model are fixed.

Although the evidence above indicates that FINESSE is not described by model dimension alone, dimension and parameter count are very tightly coupled (Pearson $r = 0.992$; Spearman $\rho = 0.982$). A substantial portion of the dimension effect is therefore entangled with model scale, and double control (dimension together with parameter count) is unstable at $n = 34$ owing to multicollinearity. Rather than forcing an artificial separation, we leave the isolation of any residual dimension effect to future work.

E Detailed Experiment Results

E.1 RSS Evaluation Results

Table 18 and Figure 21 present the complete RSS results for all 34 models plus the baseline. Several patterns are visible at this scale. First, the per-column shading confirms a general trend of monotonic degradation, from $L = 4$ to $L = 16$. However, scores remain positive for all models except one: **nomic-embed-text-v1.5**, whose $L = 16$ score turns negative—discussed in Section 4.2.1. Second,

Table 18: RSS scores across sequence lengths for all 34 evaluated models, sorted by Avg($L = 4 \ldots L = 16$) in descending order. Cell shading indicates relative magnitude within each column (darker = higher RSS). Red indicates a negative score. The *mean(ml-e5-base)* entry is a baseline.

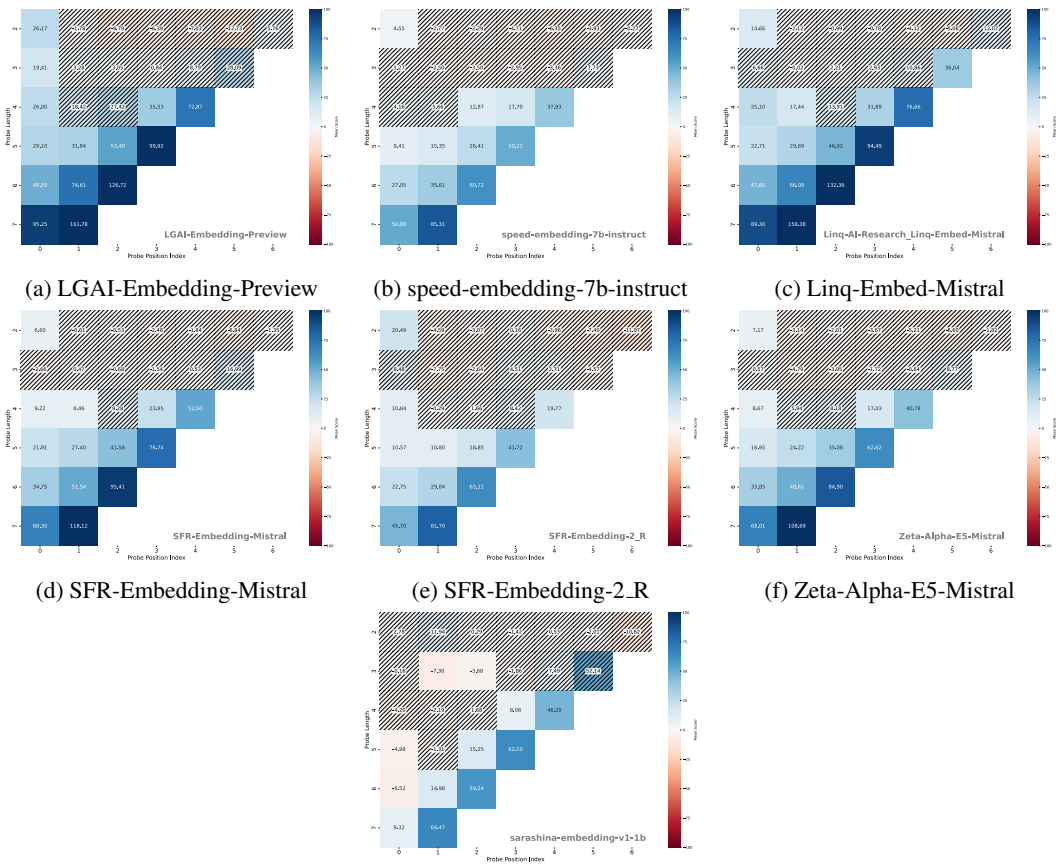
Model	$L = 4$	$L = 8$	$L = 16$	Avg($L4 \ldots L16$)
llama-embed-nemotron-8b	456.64	263.29	124.29	236.16
F2LLM-v2-4B	398.66	237.36	140.65	223.56
F2LLM-v2-1.7B	366.79	225.50	125.68	213.55
F2LLM-v2-8B	383.29	225.48	125.70	211.84
Linq-Embed-Mistral	334.09	206.64	112.38	194.17
snowflake-arctic-embed-l-v2.0	395.53	220.37	89.39	193.71
jina-embeddings-v5-text-nano	352.46	198.29	114.98	192.30
F2LLM-v2-0.6B	335.94	172.20	95.72	169.75
modernbert-embed-base	296.69	171.51	90.76	156.36
Octen-Embedding-8B	268.47	143.47	90.41	143.68
SFR-Embedding-2_R	224.67	146.14	75.25	138.18
F2LLM-4B	296.95	143.00	71.06	136.22
bge-m3	251.35	144.72	80.33	135.94
gte-modernbert-base	241.55	149.70	83.69	133.56
LGAI-Embedding-Preview	231.55	139.09	75.91	131.49
Octen-Embedding-4B	204.83	137.32	87.22	129.42
Zeta-Alpha-E5-Mistral	213.15	129.21	66.05	122.87
bge-m3-unsupervised	243.02	131.81	67.10	120.10
SFR-Embedding-Mistral	213.70	125.90	62.73	118.99
jina-embeddings-v5-text-small	250.39	116.81	47.01	111.60
Octen-Embedding-0.6B	192.75	103.12	59.57	104.32
BOOM_4B_v1	208.72	100.94	46.95	95.09
speed-embedding-7b-instruct	169.31	98.66	47.24	93.12
Qwen3-Embedding-8B	180.35	82.27	51.36	87.59
MoD-Embedding	147.79	88.21	65.52	87.28
nomic-embed-text-v1	156.18	88.29	44.03	80.32
Qwen3-Embedding-4B	131.62	76.94	54.38	74.67
granite-embedding-english-r2	111.31	67.74	41.75	62.96
F2LLM-1.7B	218.78	53.93	5.29	62.30
F2LLM-0.6B	204.00	64.73	5.77	58.46
granite-embedding-small-english-r2	102.02	54.09	32.43	53.59
Qwen3-Embedding-0.6B	98.49	43.21	25.73	46.44
nomic-embed-text-v1.5	139.39	65.19	-16.12	41.67
sarashina-embedding-v1-1b	74.24	32.89	11.72	30.16
mean(ml-e5-base)	99.97	57.22	30.53	53.38

the ranking by $L = 4$ and ranking by $L = 16$ do not coincide: **llama-embed-nemotron-8b** leads at $L = 4$ (456.64) but is surpassed at $L = 16$ by **F2LLM-v2-4B** (140.65 vs. 124.29), indicating that initial sequence sensitivity and long-range retention are partially independent axes. Third, the F2LLM series exhibits a version effect: at 0.6B parameters, v1 collapses to an RSS of 5.77 at $L = 16$ while v2 retains 95.72—a roughly 16$\times$ difference attributable to training methodology rather than model scale. A similar gap appears at 1.7B (5.29 vs. 125.68). These version-level contrasts suggest that RSS is sensitive enough to capture improvements in sequence handling that occur between training iterations of the same architecture.

E.2 SRS Evaluation Results

E.2.1 The Diagonal Pattern and its Link to Model Lineage

This group of models is characterized by a distinctive "diagonal" pattern, where sequence sensitivity is concentrated in the latter part of the sequence, as discussed in Section 4.2.2. This pattern appears



(a) LGAI-Embedding-Preview (b) speed-embedding-7b-instruct (c) Linq-Embed-Mistral

(d) SFR-Embedding-Mistral (e) SFR-Embedding-2_R (f) Zeta-Alpha-E5-Mistral

(g) sarashina-embedding-v1-1b

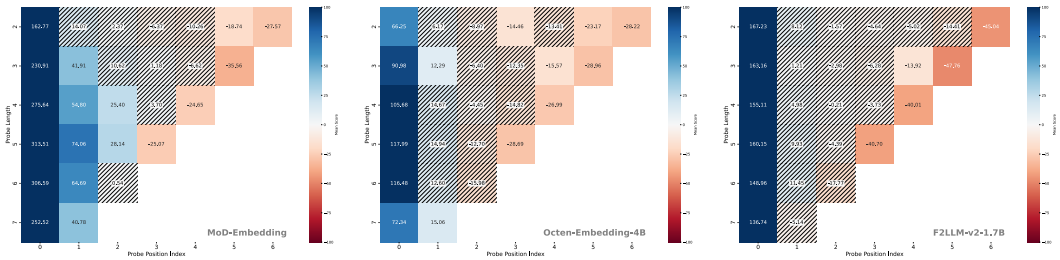
Figure 12: SRS heatmaps for the late-concentration (diagonal) group. Five of the seven models are fine-tuned from **e5-mistral-7b-instruct**; **LGAI-Embedding-Preview** derives directly from **Mistral-7B**, and **sarashina-embedding-v1-1b** belongs to a separate lineage.

to be a strong indicator of model lineage. Of the seven models in this group, six are fine-tuned from a common base model, **e5-mistral-7b-instruct** (Wang et al., 2024a), with **sarashina-embedding-v1-1b** as the only exception. **LGAI-Embedding-Preview** is also an exception: it was derived directly from **Mistral-7B** (Jiang et al., 2023) rather than by further fine-tuning the existing E5-Mistral embedder (Wang et al., 2024a), although the developers note that their work was inspired by E5-Mistral (Choi et al., 2025, p.6).

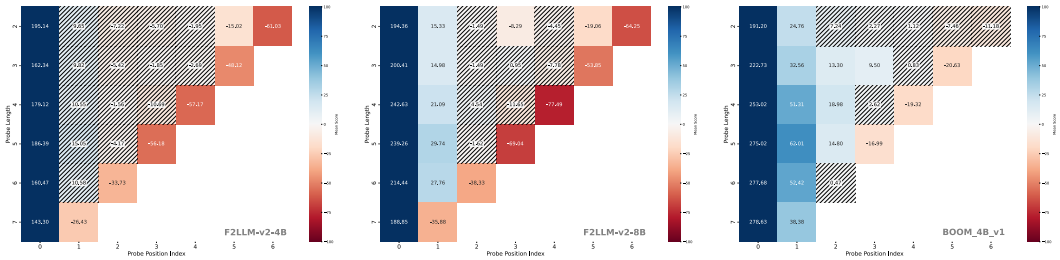
While one might attribute this pattern to general architectural properties of Causal LMs or last-token pooling, this hypothesis is challenged by the fact that other models with the same architecture exhibit entirely different SRS patterns. It is therefore more plausible that this diagonal signature is not a generic architectural artifact but a heritable trait of the **Mistral-7B backbone lineage**: within our model set, it appears both in models fine-tuned from E5-Mistral and in **LGAI-Embedding-Preview**, which derives from **Mistral-7B** directly and bypasses E5-Mistral entirely. **sarashina-embedding-v1-1b** remains the sole exception in this group.

E.2.2 Inheritance of a Hybrid SRS Pattern across the Qwen3 Family

In contrast, Figure 13 shows a group of models exhibiting "hybrid" pattern: sequence recognition is heavily concentrated at the very beginning of the text (i.e., at $\mathrm{SRS}(l, 0)$), while simultaneously displaying a distinct band of negative SRS scores along the trailing diagonal. With the exception of **BOOM_4B_v1** and **MoD-Embedding**, the remaining models in this group belong to distinct multi-parameter model families, which we analyze collectively below.

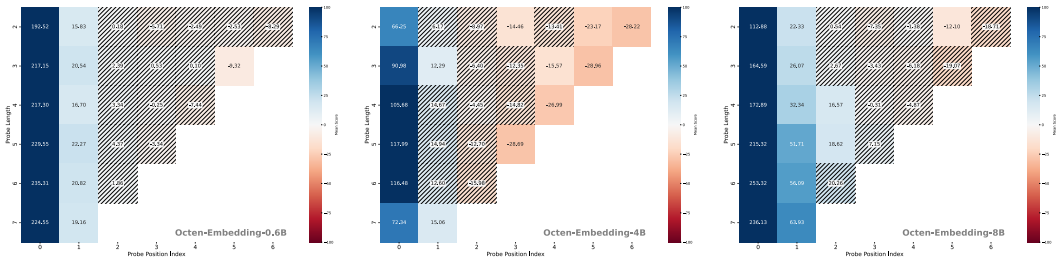


(a) MoD-Embedding (b) Octen-Embedding-4B (c) F2LLM-v2-1.7B



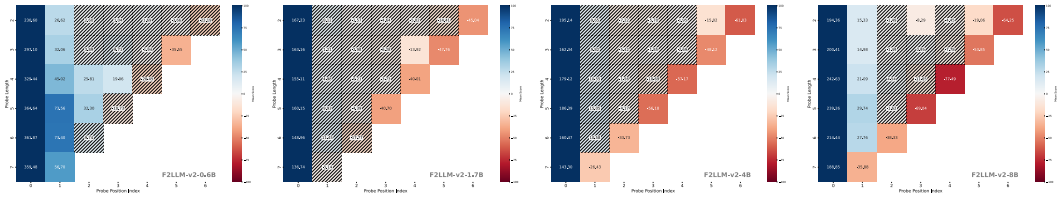
(d) F2LLM-v2-4B (e) F2LLM-v2-8B (f) BOOM_4B_v1

Figure 13: SRS heatmaps for the hybrid group, exhibiting strong early-concentration at SRS$(l, 0)$ together with a band of negative scores along the trailing diagonal.



(a) Octen-Embedding-0.6B (b) Octen-Embedding-4B (c) Octen-Embedding-8B

Figure 14: SRS heatmaps for the Octen model family (0.6B, 4B, 8B), all derived from **Qwen3-Embedding** via second-generation fine-tuning.

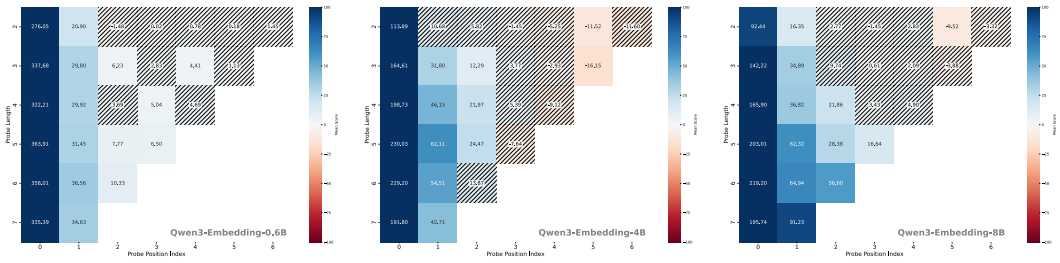


(a) F2LLM-v2-0.6B (b) F2LLM-v2-1.7B (c) F2LLM-v2-4B (d) F2LLM-v2-8B

Figure 15: SRS heatmaps for the F2LLM-v2 model family (0.6B through 8B), fine-tuned directly from the base **Qwen3** LLMs. The negative trailing diagonal intensifies with increasing parameter count.

Examining the **Octen** (Figure 14) and **F2LLM-v2** (Figure 15) model families, we observe that every model across both lineages contains at least one negative SRS cell. The trend is particularly pronounced in the **F2LLM-v2** series: as the parameter count increases, the negative tendency along the trailing diagonal becomes progressively stronger. Crucially, both model families are built upon the **Qwen3** (Team, 2025) base model architecture (e.g., **Qwen3-1.7B-Base**, . . . , **Qwen3-8B-Base**). A subtle distinction exists in their fine-tuning heritage: whereas the **F2LLM-v2** series was fine-tuned

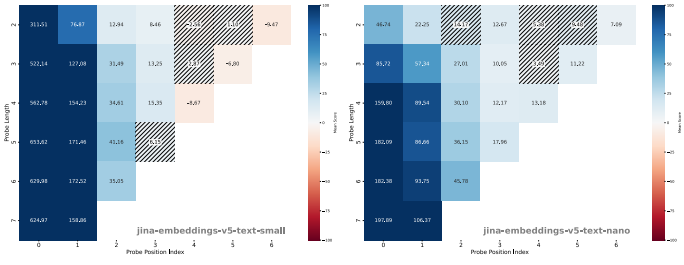
directly from the base **Qwen3** LLMs (Zhang et al., 2026b), the **Octen** series is derived from the **Qwen3-Embedding** (Zhang et al., 2025a) models, which are themselves fine-tuned from **Qwen3** (a second-generation inheritance).



(a) Qwen3-Embedding-0.6B (b) Qwen3-Embedding-4B (c) Qwen3-Embedding-8B

Figure 16: SRS heatmaps for the Qwen3-Embedding family (0.6B, 4B, 8B). Negative cells are absent in the 0.6B variant but emerge in the 4B and 8B variants.

Figure 16 shows that **Qwen3-Embedding** (Zhang et al., 2025a) family itself also exhibits this trait, with negative SRS cells present in both the 4B and 8B variants, although the **Qwen3-Embedding-0.6B** model notably lacks any significant negative cells.



(a) jina-embeddings-v5-text-small (b) jina-embeddings-v5-text-nano

Figure 17: SRS heatmaps for the jina-embeddings-v5 series, serving as a control case for base-model lineage. The **small** variant (Qwen3-0.6B-Base) displays weak negative trailing scores; the **nano** variant (EuroBERT-210m) does not.

Figure 17 shows a compelling control case provided by the **jina-embeddings-v5** (Akram et al., 2026) series. The two models present drastically different SRS heatmaps: the **nano** variant is characterized by exclusively positive scores, whereas the **small** variant clearly displays weak negative scores in the trailing region. This divergence aligns with their respective base models: the **small** variant is founded on **Qwen3-0.6B-Base**, while the **nano** variant is built upon an entirely different lineage, **EuroBERT-210m** (Boizard et al., 2025), and entirely lacks the hybrid signature.

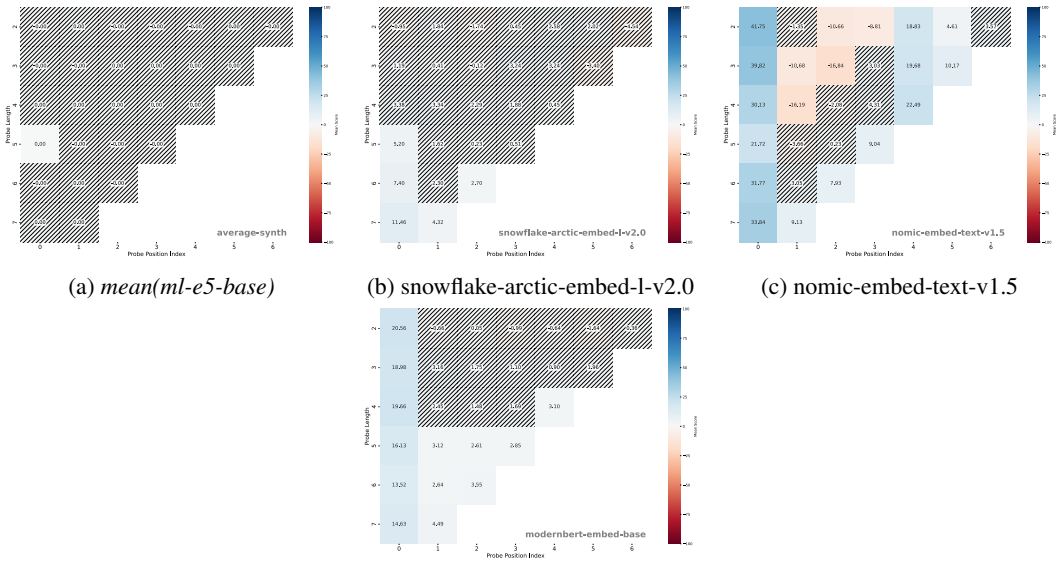
In summary, across multiple independent fine-tuning efforts, models that share the **Qwen3** base model consistently exhibit varying degrees of this negative SRS tendency in their trailing sequences. It is therefore highly probable that this complex hybrid pattern—and specifically the negative trailing diagonal—is a strongly heritable structural fingerprint originating from the **Qwen3** base model itself.

E.2.3 Models with Negligible Sequence Sensitivity

Figure 18 shows the models that have generally low values across their entire heatmaps. Notably, the zero-valued pattern of the baseline control, *mean(ml-e5-base)*, confirms that the **chunk-embed-then-average** strategy is entirely order-agnostic by construction: because each chunk is embedded independently before averaging, no positional information can survive into the final representation.

E.2.4 Explaining the Nemotron Anomaly

As established in Section 4.2.2, **llama-embed-nemotron-8b** is the only model among the 34 evaluated to exhibit a **fully negative** SRS pattern. Across the full model set (Figure 20), the vast majority show



(a) *mean(ml-e5-base)* (b) snowflake-arctic-embed-l-v2.0 (c) nomic-embed-text-v1.5

(d) modernbert-embed-base

Figure 18: SRS heatmaps for models exhibiting negligible sequence sensitivity. The zero-valued pattern of the *mean(ml-e5-base)* baseline confirms that the chunk-embed-then-average strategy is, by construction, entirely order-agnostic.

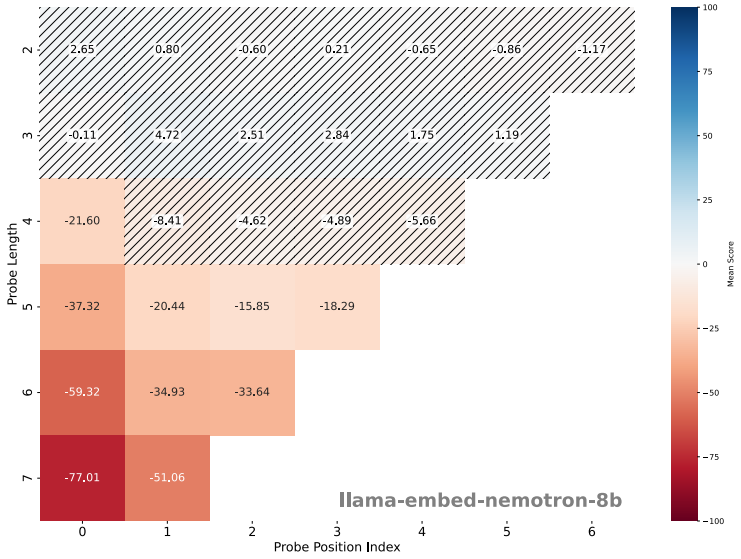


Figure 19: SRS heatmap at $L = 8$ for **llama-embed-nemotron-8b**, the only model in the evaluation set to exhibit a fully negative pattern across all statistically significant cells.

positive scores concentrated at the beginning of the text—i.e., at SRS$(l, 0)$—with the remaining cells containing weak positive or statistically non-significant values.

There is a clear architectural explanation for this anomaly. An examination of the model configuration (Table 22) reveals **llama-embed-nemotron-8b** as the only Causal LM-based model that uses mean pooling, but this observation alone is insufficient. The underlying cause becomes apparent when comparing how **nemotron** and **Qwen3-Embedding** were derived from their respective base models.

According to its technical report of Babakhin et al. (2025, p.2), **nemotron** was obtained from the **Llama-3.1-8B** base decoder by first **removing the causal attention mask** to convert it into a bidi-

rectional encoder, after which fine-tuning commenced. This architectural modification necessitated mean pooling over the final hidden states to produce a single embedding vector. In stark contrast, **Qwen3-Embedding** retained the causal attention of its base **Qwen3** LLM and simply appended an **[EOS]** token, using the hidden state of this single token as the final embedding (Zhang et al., 2025a, p.3).

In short, **nemotron** is not a decoder; it is an **encoder derived from a decoder**. This training methodology is unique among the 34 models evaluated here and offers the most plausible explanation for its distinctive, fully negative SRS pattern.

E.3 LEMB(LongEmbed) Evaluation Results

Table 19: LEMB results for all 34 evaluated models and the *mean(ml-e5-base)* baseline. Synthetic tasks are averaged. Models are sorted by overall average; cell shading indicates relative magnitude within each column.

Model	Synthetic (Acc@1)		Real (nDCG@10)				Avg.
	Needle	Passkey	SummScreen	QMSum	WikimQA	NQA	
llama-embed-nemotron-8b	0.800	0.997	0.981	0.555	0.879	0.678	0.815
F2LLM-v2-8B	0.700	0.997	0.985	0.485	0.938	0.689	0.799
F2LLM-v2-4B	0.667	0.990	0.978	0.477	0.926	0.653	0.782
BOOM_4B_v1	0.663	1.000	0.974	0.512	0.928	0.565	0.774
Octen-Embedding-4B	0.773	0.997	0.973	0.465	0.868	0.549	0.771
MoD-Embedding	0.693	0.990	0.948	0.419	0.891	0.540	0.747
F2LLM-v2-0.6B	0.713	0.933	0.965	0.439	0.901	0.506	0.743
F2LLM-v2-1.7B	0.490	1.000	0.976	0.472	0.911	0.578	0.738
Linq-Embed-Mistral	0.597	0.927	0.965	0.468	0.843	0.574	0.729
Octen-Embedding-8B	0.513	0.980	0.962	0.481	0.819	0.562	0.720
F2LLM-4B	0.683	0.997	0.966	0.408	0.807	0.447	0.718
Qwen3-Embedding-4B	0.603	0.980	0.932	0.377	0.875	0.528	0.716
Zeta-Alpha-E5-Mistral	0.590	0.863	0.968	0.465	0.844	0.530	0.710
jina-embeddings-v5-text-small	0.577	0.967	0.915	0.423	0.797	0.528	0.701
speed-embedding-7b-instruct	0.573	0.940	0.945	0.430	0.807	0.492	0.698
SFR-Embedding-Mistral	0.533	0.860	0.962	0.438	0.827	0.536	0.693
LGAI-Embedding-Preview	0.600	0.923	0.955	0.432	0.666	0.520	0.682
SFR-Embedding-2_R	0.533	0.787	0.961	0.451	0.830	0.531	0.682
Octen-Embedding-0.6B	0.510	0.987	0.939	0.428	0.809	0.421	0.682
granite-embedding-english-r2	0.517	0.950	0.931	0.412	0.858	0.415	0.680
snowflake-arctic-embed-l-v2.0	0.597	0.927	0.963	0.401	0.749	0.436	0.679
F2LLM-1.7B	0.780	1.000	0.916	0.325	0.680	0.340	0.673
Qwen3-Embedding-0.6B	0.460	0.987	0.927	0.393	0.783	0.443	0.666
modernbert-embed-base	0.533	0.800	0.957	0.405	0.812	0.467	0.662
granite-embedding-small-english-r2	0.650	0.953	0.881	0.350	0.787	0.340	0.660
Qwen3-Embedding-8B	0.567	0.980	0.844	0.388	0.535	0.471	0.631
bge-m3-unsupervised	0.493	0.960	0.954	0.393	0.673	0.301	0.629
bge-m3	0.493	0.743	0.944	0.352	0.750	0.428	0.618
gte-modernbert-base	0.323	0.810	0.929	0.426	0.858	0.289	0.606
jina-embeddings-v5-text-nano	0.733	0.943	0.442	0.241	0.743	0.497	0.600
F2LLM-0.6B	0.700	1.000	0.804	0.279	0.566	0.249	0.600
nomic-embed-text-v1	0.350	0.593	0.921	0.348	0.737	0.379	0.555
sarashina-embedding-v1-1b	0.453	0.680	0.886	0.236	0.423	0.321	0.500
nomic-embed-text-v1.5	0.253	0.483	0.668	0.141	0.363	0.154	0.344
mean(ml-e5-base)	0.293	0.843	0.949	0.378	0.737	0.403	0.601

In this Table 19, F2LLM series again illustrates how LEMB performance aligns with the signals identified by FINESSE. While the overall LEMB average suggests non-monotonic scaling within **F2LLM-v2** (0.6B and 1.7B scoring 0.743 vs. 0.738), isolating NarrativeQA—the most holistic of the real tasks—reveals strict monotonic increase with parameter count (0.506 $\rightarrow$ 0.578 $\rightarrow$ 0.653 $\rightarrow$ 0.689 for 0.6B to 8B). This scale effect complements the version effect already observed in RSS (Appendix E.1, where v1 collapses while v2 survives at matched parameter counts), and corresponds

geometrically to the progressive deepening of the trailing negative diagonal in the SRS heatmaps (Figure 15). Together, the three metrics capture orthogonal axes of the same training pipeline: RSS flags the recipe transition between versions, LEMB's holistic tasks track the scale effect within a fixed recipe, and SRS provides the order sensitivity signature, account of why larger models within a consistent recipe generalize better.

E.4 T-TEST VERIFICATION OF RSS DEGRADATION ACROSS L

Table 20: Welch's one-tailed t-tests for RSS degradation at all adjacent length transitions, across 34 models. "# Sig." is the number of models for which the transition reached significance at $\alpha = 0.05$. "Median p" and "Max p" are computed across the 34 per-model p-values. The top block reports all 12 unit-spaced transitions; the bottom block reports the two doubling transitions used as the primary analysis scale in the main text. Note the approximately eight orders of magnitude gap between the median p-values of the two blocks.

Transition	# Sig. / Total	Median p	Max p	Worst-case model
Unit-spaced transitions				
$L4 \to L5$	34 / 34	1.95×10^{-5}	2.79×10^{-2}	Salesforce_SFR-Embedding-2_R
$L5 \to L6$	16 / 34	8.73×10^{-2}	8.12×10^{-1}	codefuse-ai_F2LLM-v2-4B
$L6 \to L7$	29 / 34	5.63×10^{-3}	3.36×10^{-1}	ibm-granite_granite-embedding-english-r2
$L7 \to L8$	16 / 34	5.58×10^{-2}	7.60×10^{-1}	bflhc_MoD-Embedding
$L8 \to L9$	8 / 34	1.84×10^{-1}	8.90×10^{-1}	Qwen_Qwen3-Embedding-0.6B
$L9 \to L10$	23 / 34	7.71×10^{-3}	3.12×10^{-1}	Qwen_Qwen3-Embedding-8B
$L10 \to L11$	19 / 34	4.44×10^{-2}	6.63×10^{-1}	ibm-granite_granite-embedding-small-english-r2
$L11 \to L12$	8 / 34	1.64×10^{-1}	9.83×10^{-1}	Qwen_Qwen3-Embedding-0.6B
$L12 \to L13$	18 / 34	4.63×10^{-2}	4.46×10^{-1}	ibm-granite_granite-embedding-english-r2
$L13 \to L14$	13 / 34	1.77×10^{-1}	8.68×10^{-1}	codefuse-ai_F2LLM-v2-0.6B
$L14 \to L15$	9 / 34	1.47×10^{-1}	8.57×10^{-1}	Qwen_Qwen3-Embedding-4B
$L15 \to L16$	21 / 34	2.73×10^{-2}	7.93×10^{-1}	Alibaba-NLP_gte-modernbert-base
Doubling transitions (primary scale)				
$L4 \to L8$	34 / 34	5.55×10^{-13}	1.90×10^{-6}	bflhc_MoD-Embedding
$L8 \to L16$	34 / 34	5.32×10^{-13}	2.12×10^{-3}	Qwen_Qwen3-Embedding-0.6B

To verify quantitatively that RSS scores decrease as L increases, we conducted Welch's one-tailed two-sample t-tests for each of the 34 models in our evaluation set, across two transitions: $L = 4 \to L = 8$ and $L = 8 \to L = 16$. Each test compares the distributions of 25 measurements obtained at each length, with the samples at L_n and L_{n+1} treated as independent (different evaluation instances are drawn at each length). We used Welch's variant to avoid assuming equal variances across length conditions, and tested $H_0 : \text{RSS}_{L_n} \leq \text{RSS}_{L_{n+1}}$ against $H_1 : \text{RSS}_{L_n} > \text{RSS}_{L_{n+1}}$. This yields a total of $34 \times 2 = 68$ individual t-tests.

Every single one of the 68 tests rejected the null hypothesis at the significance level $\alpha = 0.05$ (see the bottom block of Table 20). The worst-case scenario for each transition still yields p-values well below the conventional threshold. The least significant case overall is **Qwen3-Embedding-0.6B**'s $L = 8 \to L = 16$ degradation at $p = 2.12 \times 10^{-3}$, which is still much smaller than $\alpha = 0.05$.

We additionally examined all unit-spaced adjacent transitions ($L = 4 \to L = 5$, $L = 5 \to L = 6$, $\ldots$, $L = 15 \to L = 16$) under the same Welch's test protocol, reported in the top block of Table 20. At unit granularity the degradation signal is qualitatively weaker and highly non-uniform across transitions: the number of models reaching significance ranges from 34 / 34 at $L = 4 \to L = 5$ down to 8 / 34 at $L = 8 \to L = 9$ and $L = 11 \to L = 12$. By contrast, the two doubling transitions yield median p-values of 5.55×10^{-13} ($L = 4 \to L = 8$) and 5.32×10^{-13} ($L = 8 \to L = 16$) and reach significance for all 34 models without exception. We therefore adopt the $\{4, 8, 16\}$ doubling grid as the primary analysis scale throughout this work: it is the coarsest scale at which the decreasing-RSS trend with increasing L is verified to hold systematically across the entire set of 34 evaluated models.

F Details for Experiment Setup

F.1 Model Selection and Filtering Log

The entire selection process is transparently disclosed at the public link documented in the Supplementary materials (Appendix F.3).

F.1.1 Initial Model Selection Criteria

At the time of writing (**2026-03-24 00:42 UTC+9**), we obtained a list of all 393 models registered on the public MTEB leaderboard (mteb). We then filtered the models according to the following conditions:

1. **Upper bound on context window:** must have a **max_ctx** of at least 8190 tokens. We regarded an input of 8190 tokens as the minimum requirement for evaluating long-context embeddings.
2. **Upper bound on parameters:** must have fewer than 8.0B parameters. Along with experimental efficiency, we sought to evaluate models that are reproducible for many researchers.
3. **Restricted to open-source models:** closed-source models have low reproducibility and accessibility.
4. **Support for Hugging Face transformers library:** must be compatible with our experimental environment.
5. **English support:** we utilized **uonlp/CulturaX** as the test corpus for FINESSE, and the LEMB dataset is likewise an English dataset (Zhu et al., 2024). Embedders must therefore be capable of English embedding.
6. **Allowed only embedders that generate single vector representations:** architectures such as ColBERT are incompatible with FINESSE, which requires computing similarities between single vector representations.
7. **Clarity of license:** we selected only models for which the author has clearly specified a license.

Ultimately, from the 393 models, we selected **74** models.

F.1.2 Secondary Filtering of Models

In the secondary filtering process, the following models were excluded:

- **Removed: LENS-d4000**, **LENS-d8000**: excluded due to lack of max pooling support.
- **Removed: bge-en-icl**: as the product of a special research area, "few-shot in-context learning," a fair evaluation is impossible.
- **Removed: jasper_en_vision_language_v1**: as a multimodal embedder, it was excluded.
- **Removed: Jasper-Token-Compression-600M**: we could not find an inference process for this model using the **transformers** library.
- **Removed: zembed-1**: we could not find an inference process for this model using the **transformers** library.
- **Removed: nomic-embed-text-v1-unsupervised**, **nomic-embed-text-v1-ablated**: the model's author stated that these models were released for research purposes only. We were unable to infer an official inference method.
- **Removed: english_code_retriever**: the license statement was ambiguous.
- **Removed: stella_en_1.5B_v5**, **stella_en_400M_v5**: the author stated that 512 tokens is the de facto **max_ctx**.
- **Removed: opensearch-neural-sparse-encoding-doc-v3-gte**: creates a special "Sparse Vector" embedding. Removed due to compatibility issues.

- **Removed: Tarka-Embedding-350M-V1**: removed because it uses a special library called **lfm1.0**.
- **Removed: C2LLM-0.5B**, **C2LLM-7B**: use a special pooling method called PMA (Pooling by Multi-head Attention), which is not compatible with the current evaluation pipeline.

In addition, the evaluation method for the models below had to be adjusted arbitrarily:

- **Usage of LGAI-Embedding-Preview:** the model card had no specification for pooling method or prefix style. We treat it as a **last** pooling and (**Instruct: \nQuery:**) style embedder.
- **Usage of SearchMap_Preview:** likewise had no explicit specification. We assume **cls** pooling and no prefix.
- **Usage of NV-Embed models:** use a special "Latent-Attention" pooling, which we will regard as **last** pooling.
- **Usage of jina models family:** all **jinaAI** model families had prefixes specified, but the application method varied. We will therefore use the base model without a prefix.

In this process, the final model selection was reduced to **59** models.

F.1.3 Decision to Standardize max_ctx to 8190

In the official **BAAI/bge-m3** Hugging Face repository's **config.json**, we found that despite a reported **max_positional_embeddings** value of **8194** (as of commit hash b28ce2a (BAAI)), the model did not infer correctly in the actual test environment and caused a CUDA kernel crash.

Subsequently, in the official repository's **README.md**, we confirmed that the **Sequence Length** value is specified as **8192**, and that this is the actual maximum context length (**max_ctx**) of the model. This was confirmed identically in both the official announcement and our test environment.

For these reasons, we added the following rule when selecting models:

> **Model Selection Rule:** All models are assumed to have **max_ctx=8190**. We reserve 2 tokens below the 8192 ceiling as a safety margin for special tokens (BOS/EOS), yielding the operational cap of 8190. If a model fails, no further tests are conducted. Since model authors often do not provide a clear **max_ctx**, this is a judgment based on the best information we could verify.

Table 21: Context window investigation for all 59 candidate models surviving secondary filtering, of which 34 passed evaluation. "Original" refers to the value reported in the MTEB leaderboard. "Official" refers to the value we determined after investigating each model's Hugging Face repository. Models marked with a note number required interpretation; see the Notes section below.

Model	Original	Official
gte-Qwen2-7B-instruct	32768	32000 (N1)
F2LLM-v2-8B	40960	Unknown (N8)
Zeta-Alpha-E5-Mistral	32768	4096 (N2)
snowflake-arctic-embed-l-v2.0	8192	8192
jina-embeddings-v2-base-en	8192	8192
Linq-Embed-Mistral	32768	4096 (N2)
llama-embed-nemotron-8b	32768	32768
bge-m3-unsupervised	8192	8192
SFR-Embedding-2_R	32768	4096 (N2)
gte-modernbert-base	8192	8192
F2LLM-v2-0.6B	40960	Unknown (N8)
speed-embedding-7b-instruct	32768	4096 (N3)
gte-multilingual-base	8192	8192
jina-embeddings-v4	32768	32768
jina-embeddings-v2-small-en	8192	8192

Continued on next page

Table 21 – continued

Model	Original	Official
jina-embeddings-v5-text-nano	8192	8192
inf-retriever-v1-1.5b	32768	32768
gte-base-en-v1.5	8192	8192
nomic-embed-text-v1.5	8192	8192
F2LLM-v2-4B	40960	Unknown (N8)
F2LLM-4B	8192	Unknown (N8)
pplx-embed-v1-0.6b	32768	32000 (N1)
Qodo-Embed-1-7B	32768	32000 (N1)
snowflake-arctic-embed-m-v2.0	8192	8192
F2LLM-v2-1.7B	40960	Unknown (N8)
NV-Embed-v1	32768	4096 (N4)
F2LLM-v2-330M	40960	Unknown (N8)
bge-m3	8194	8192 (N6)
Qwen3-Embedding-0.6B	32768	32000 (N1)
QZhou-Embedding	8192	8000 (N7)
BOOM_4B_v1	32768	32000 (N1)
jina-embeddings-v3	8194	8192 (N6)
Octen-Embedding-8B	32768	32768
gte-Qwen2-1.5B-instruct	32768	32000 (N1)
sarashina-embedding-v1-1b	8192	8192
jina-embeddings-v5-text-small	32768	32768
F2LLM-1.7B	8192	Unknown (N8)
LGAI-Embedding-Preview	32768	32000 (N1)
Octen-Embedding-4B	32768	32768
SearchMap_Preview	8192	Unknown (N10)
granite-embedding-small-english-r2	8192	8192
Octen-Embedding-0.6B	32768	32768
F2LLM-0.6B	8192	Unknown (N8)
granite-embedding-english-r2	8192	8192
Qodo-Embed-1-1.5B	32768	32768
F2LLM-v2-80M	40960	Unknown (N8)
GeoEmbedding	32768	Unknown (N9)
nomic-embed-text-v1	8192	8192
F2LLM-v2-160M	40960	Unknown (N8)
MoD-Embedding	32768	32768
Qwen3-Embedding-8B	32768	32000 (N1)
Qwen3-Embedding-4B	32768	32000 (N1)
NV-Embed-v2	32768	32768 (N5)
modernbert-embed-base	8192	Unknown (N11)
Solon-embeddings-mini-beta-1.1	8192	Unknown (N12)
gte-Qwen1.5-7B-instruct	32768	32000 (N1)
pplx-embed-v1-4b	32768	32000 (N1)
SFR-Embedding-Mistral	32768	4096 (N2)
inf-retriever-v1	32768	32768

The following notes explain how we determined the official **max_ctx** value for models whose repositories did not provide a direct numeric answer. Since our experiments fix **max_ctx** at 8190 for all models, these interpretive decisions do not affect the evaluation outcomes, but are recorded here for transparency and reproducibility.

N1: "32k" interpreted as 32000. For models whose Hugging Face repository uses the informal notation "32k" without specifying an exact value, we interpret this as 32000. Applies to:

- **gte-Qwen2-7B-instruct**
- **gte-Qwen2-1.5B-instruct**
- **gte-Qwen1.5-7B-instruct**
- **Qwen3-Embedding-0.6B**
- **Qwen3-Embedding-4B**

- **Qwen3-Embedding-8B**
- **Qodo-Embed-1-7B**
- **pplx-embed-v1-0.6b**
- **pplx-embed-v1-4b**
- **BOOM_4B_v1**
- **LGAI-Embedding-Preview**

N2: Mistral-family models with 4096 working length. For several Mistral-based embedders, the model card does not explicitly specify a maximum context length. However, the "How to Run" or "How to use" section of each repository provides example code that uses 4096 tokens as the working input length. We therefore treat this value as the practical **max_ctx**. Applies to:

- **Zeta-Alpha-E5-Mistral**
- **Linq-Embed-Mistral**
- **SFR-Embedding-Mistral**
- **SFR-Embedding-2_R**

For **Zeta-Alpha-E5-Mistral** and **SFR-Embedding-Mistral**, the authors' blog posts further confirm training configurations of doc/query = 512/192 and 256/128 tokens respectively, well below 4096.

N3: speed-embedding-7b-instruct. The author explicitly states in the Hugging Face repository: "Using this model for inputs longer than 4096 tokens is not recommended." We therefore adopted 4096 as the official **max_ctx**.

N4: NV-Embed-v1. The "Usage" section of the Hugging Face repository provides example code using 4096 tokens. We adopted this as the official **max_ctx**.

N5: NV-Embed-v2. Unlike its predecessor, the "How to use" section of this repository provides example code using 32768 tokens. We adopted this value as the official **max_ctx**.

N6: bge-m3 family. The official configuration file of **bge-m3** reports a maximum positional embedding of 8194, but the model card and the authors' announcement confirm 8192 as the actual maximum. Attempting to use values above 8192 caused CUDA kernel crashes in our test environment. The same discrepancy applies to **jina-embeddings-v3**. We treat 8192 as the authoritative value for both.

N7: QZhou-Embedding. The repository states the context window as "8k" (interpreted as 8000 per the convention in N1), but the "Usage" section example code uses 8192. We adopted 8000 as the conservative lower bound, which is safely within both interpretations and below the 8190 limit applied uniformly in our experiments.

N8: F2LLM and F2LLM-v2 series (Unknown). The MTEB leaderboard reports 40960 tokens for the **F2LLM-v2** variants and 8192 for the original **F2LLM** variants, but the official model cards do not confirm these values, and we could not locate authoritative documentation. We therefore marked these as Unknown. Applies to:

- **F2LLM-v2-8B**
- **F2LLM-v2-4B**
- **F2LLM-v2-1.7B**
- **F2LLM-v2-0.6B**
- **F2LLM-v2-330M**
- **F2LLM-v2-160M**
- **F2LLM-v2-80M**

- **F2LLM-4B**
- **F2LLM-1.7B**
- **F2LLM-0.6B**

N9: GeoEmbedding. The repository reports **max_ctx** of 32768, but the author notes that the model is fine-tuned on **NovaSearch/stella_en_400M_v5**, which itself has **max_ctx** of 512. This inheritance raises concerns about the practical usable context length. We marked this model as Unknown rather than adopting either value uncritically.

N10: SearchMap_Preview. The Hugging Face repository provides no clear specification of **max_ctx**, nor any example usage that implies a concrete value. Marked as Unknown.

N11: modernbert-embed-base. The repository does not explicitly state the maximum context length, and we could not verify it from other official sources. Marked as Unknown.

N12: Solon-embeddings-mini-beta-1.1. Similar to N11, no clear **max_ctx** specification was found in the repository. Marked as Unknown.

F.1.4 Models Excluded Due to LEMB Evaluation Failures

Most appeared to suffer from compatibility issues—typically, model-side code that has not been updated for the newer transformers version.

The following two models caused a **CUDA error: device-side assert triggered**, contaminating the execution instance and potentially causing other, normally functioning models to raise the same exception.

- **gte-base-en-v1.5**
- **gte-multilingual-base**

In addition, we specify the models that could not be run, grouped by error type:

- Error:
  No module named 'transformers.models.qwen2.tokenization_qwen2_fast'
  - **gte-Qwen2-1.5B-instruct**
  - **gte-Qwen1.5-7B-instruct**
  - **gte-Qwen2-7B-instruct**
  - **inf-retriever-v1-1.5b**
  - **inf-retriever-v1**
  - **Qodo-Embed-1-7B**
  - **Qodo-Embed-1-1.5B**
- Error:
  'num_hidden_layers' (x) must equal the number of layer types (y)
  - **F2LLM-v2-80M**
  - **F2LLM-v2-160M**
  - **F2LLM-v2-330M**
- Error:
  No module named 'transformers.onnx'
  - **jina-embeddings-v2-base-en**
  - **jina-embeddings-v2-small-en**
- Error:
  please install xformers
  - **snowflake-arctic-embed-m-v2.0**
  - **SearchMap_Preview**

- Error:
  list index out of range
  - **pplx-embed-v1-4b**
  - **pplx-embed-v1-0.6b**
- Miscellaneous Errors:
  - **GeoEmbedding**: Error:
    'MistralDualModel' object has no attribute '_update_causal_mask'
  - **jina-embeddings-v3**: Error:
    '¡' not supported between instances of 'str' and 'int'
  - **jina-embeddings-v4**: Error:
    cannot import name 'SlidingWindowCache' from 'transformers.cache_utils'
  - **QZhou-Embedding**: Error:
    'Qwen2Config' object has no attribute 'rope_theta'
  - **NV-Embed-v1**: Error:
    cannot import name 'MISTRAL_INPUTS_DOCSTRING' from 'transformers-.models.mistral.modeling_mistral'
  - **NV-Embed-v2**: Error:
    'NVEmbedModel' object has no attribute 'all_tied_weights_keys'
  - **Solon-embeddings-mini-beta-1.1**: Error:
    'default'

Finally, we were able to confirm **34** models.

F.2 Configuration of Selected Models

The following table details the configuration for each of the 34 models evaluated in this study, including parameter count, pooling method, and instruction prefix usage.

Table 22: Configuration of all 34 models used in our experiments: total parameter count (in billions), embedding dimension, pooling method, whether the model uses instruction-tuned query/document prefixes, and the actual prefix strings (if any). **Inst.** indicates whether the model is designed to accept instruction-style prefixes. All models were evaluated with **max_ctx** fixed at 8190 tokens. "n/a" in the prefix columns indicates that no prefix was applied—either because the model does not require one, or because prefix usage was deliberately omitted during evaluation (see Appendix F.1).

Model	Dim	Params (B)	Pooling	Inst.	Query Prefix	Doc Prefix
gte-modernbert-base(Zhang et al., 2024a; Li et al., 2023)	768	0.149	cls	no	n/a	n/a
LGAI-Embedding-Preview(Choi et al., 2025)	4096	7.11	last	yes	n/a	n/a
bge-m3(Chen et al., 2024b)	1024	0.568	cls	no	n/a	n/a
bge-m3-unsupervised(Chen et al., 2024b)	1024	0.568	cls	no	n/a	n/a
MoD-Embedding(MoD Team, 2025)	2560	4.022	last	no	n/a	n/a
Octen-Embedding-0.6B(Octen Team, 2025; Octen, a)	1024	0.596	last	no	n/a	n/a
Octen-Embedding-4B(Octen Team, 2025; Octen, b)	2560	4.022	last	no	n/a	n/a
Octen-Embedding-8B(Octen Team, 2025; Octen, c)	4096	7.567	last	no	n/a	n/a
F2LLM-0.6B(Zhang et al., 2025b)	1024	0.596	last	yes	n/a	n/a
F2LLM-1.7B(Zhang et al., 2025b)	2048	1.721	last	yes	n/a	n/a

Continued on next page

Table 22 – continued

Model	Dim	Params (B)	Pooling	Inst.	Query Prefix	Doc Prefix
F2LLM-4B(Zhang et al., 2025b)	2560	4.022	last	yes	n/a	n/a
F2LLM-v2-0.6B(Zhang et al., 2026b)	1024	0.596	last	yes	n/a	n/a
F2LLM-v2-1.7B(Zhang et al., 2026b)	2048	1.721	last	yes	n/a	n/a
F2LLM-v2-4B(Zhang et al., 2026b)	2560	4.022	last	yes	n/a	n/a
F2LLM-v2-8B(Zhang et al., 2026b)	4096	7.568	last	yes	n/a	n/a
speed-embedding-7b-instruct(Chen et al., 2024a)	4096	7.111	last	yes	n/a	n/a
granite-embedding-english-r2(Awasthy et al., 2025)	768	0.149	cls	no	n/a	n/a
granite-embedding-small-english-r2(Awasthy et al., 2025)	384	0.048	cls	no	n/a	n/a
BOOM_4B_v1(Zhang et al., 2026a)	2560	4.022	last	yes	n/a	n/a
jina-embeddings-v5-text-nano(Akram et al., 2026)	768	0.212	last	no	n/a	n/a
jina-embeddings-v5-text-small(Akram et al., 2026)	1024	0.596	last	no	n/a	n/a
Linq-Embed-Mistral	4096	7.111	last	yes	n/a	n/a
modernbert-embed-base(Nussbaum et al., 2024)	768	0.149	mean	no	search_query:	search_document:
nomic-embed-text-v1(Nussbaum et al., 2024)	768	0.137	mean	no	search_query:	search_document:
nomic-embed-text-v1.5(Nussbaum et al., 2024)	768	0.137	mean	no	search_query:	search_document:
llama-embed-nemotron-8b(Babakhin et al., 2025)	4096	7.505	mean	yes	n/a	n/a
Qwen3-Embedding-0.6B(Zhang et al., 2025a)	1024	0.596	last	yes	n/a	n/a
Qwen3-Embedding-4B(Zhang et al., 2025a)	2560	4.022	last	yes	n/a	n/a
Qwen3-Embedding-8B(Zhang et al., 2025a)	4096	7.567	last	yes	n/a	n/a
SFR-Embedding-2_R(Meng et al., 2024)	4096	7.111	last	yes	n/a	n/a
SFR-Embedding-Mistral(Meng et al., 2024)	4096	7.111	last	yes	n/a	n/a
sarashina-embedding-v1-1b	1792	1.224	last	no	n/a	n/a
snowflake-arctic-embed-l-v2.0(Yu et al., 2024)	1024	0.568	cls	no	query:	n/a
Zeta-Alpha-E5-Mistral(Camara, 2024)	4096	7.111	last	yes	n/a	n/a

F.3 Code and Data Availability

All code, raw evaluation data, and detailed logs referenced throughout this paper—including all tables and figures, which can be regenerated by the deterministic script—are attached as Supplementary materials.

Supplementary materials additionally contain the dataset and evaluation scripts for **SFfD** (Summary Finding from Documents), an exploratory benchmark that falls outside the scope of the present paper. We leave its detailed investigation for future work.

F.4 EXPERIMENTAL ENVIRONMENTS

All FINESSE and LEMB evaluations were conducted on a Google Colab Pro+ instance, with the total evaluation time summing to approximately 234 hours and 23 minutes. The detailed specifications of this primary environment are provided in Table 23.

An exception was the exploratory SRS($L = 16$) heatmap analysis and RSS($L \in \{4, 8, \cdots, 64\}$) extension presented in Appendix B.2/Appendix C.5, were run on a separate environment.

Table 23: Evaluation Environment Specifications

Component	Primary Environment (FINESSE & LEMB)	SRS($L = 16$)	RSS($L \in \{4, 8, \cdots, 64\}$)
Source	Google Colab Pro+ (G4)	vast.ai	vast.ai
CPU Memory	189.9 GB Total	2.14 TB Total	2.79 TB Total
GPU	NVIDIA RTX PRO 6000	NVIDIA B200	NVIDIA B200
GPU Memory	94.97 GB Total	178.35 GB Total	178.34 GB Total
CUDA Version	12.8	12.8	12.8
Torch Version	2.10.0+cu128	2.10.0+cu128	2.10.0+cu128
Transformers	5.0.0	5.0.0	5.0.0
Total Time	**234h 23m**	**17h 42m**	**15h 41m**

Further details, including specific package versions and evaluation scripts for each model, are available in the Supplementary materials (Appendix F.3).

G FINESSE EVALUATION FIGURES

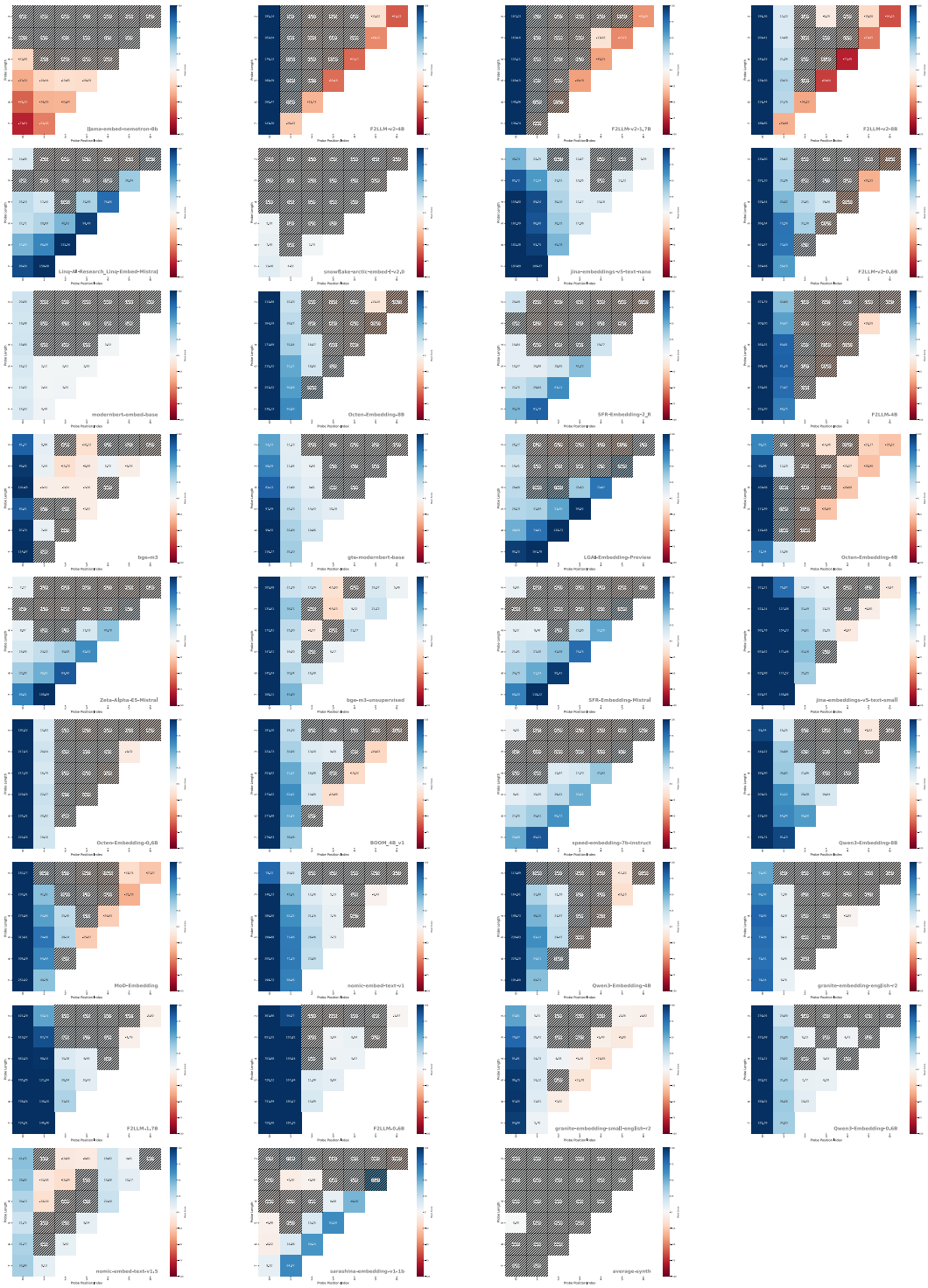


Figure 20: SRS($L = 8$) heatmap for all 34 evaluated models plus the baseline. The order matches Table 18.

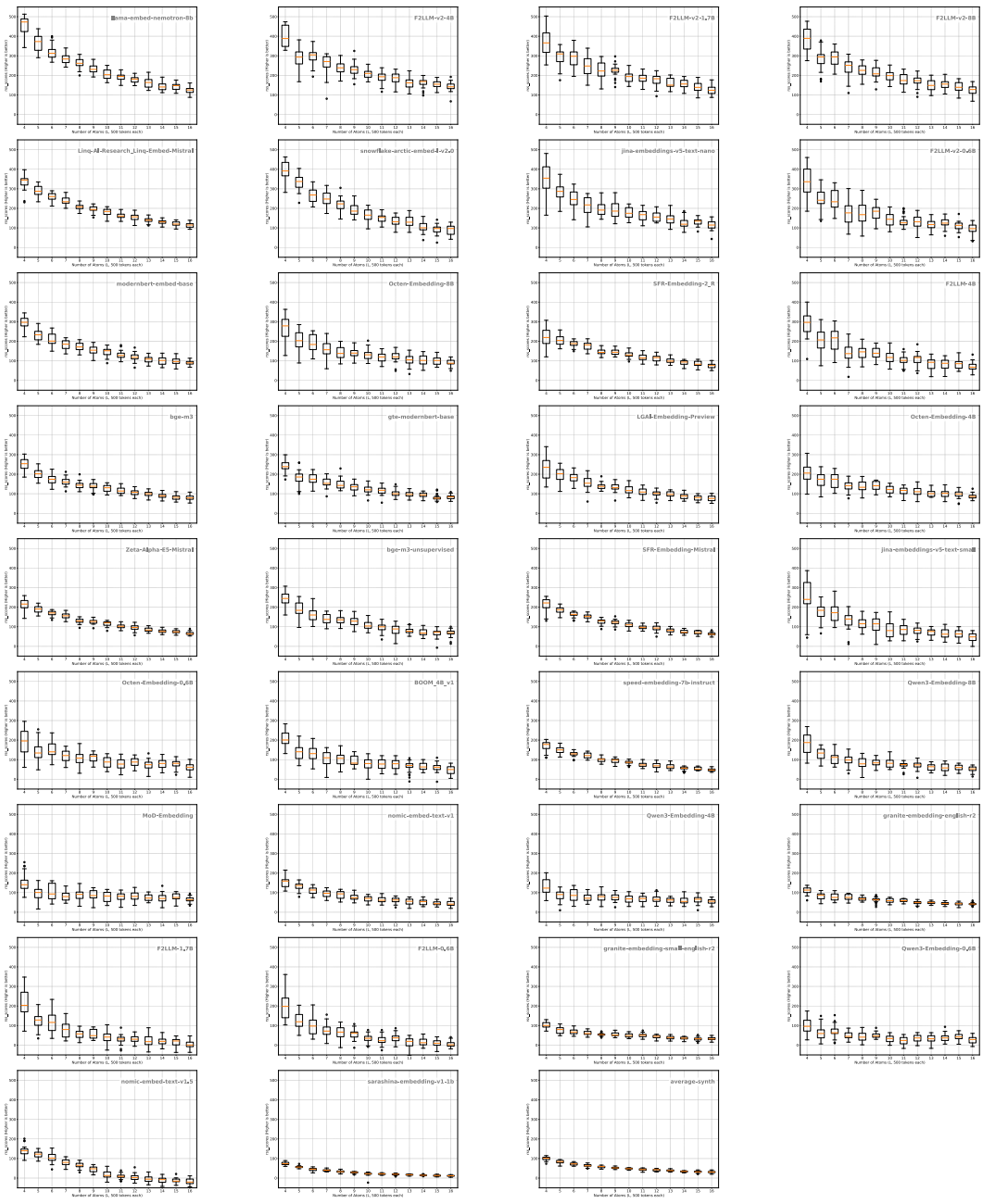


Figure 21: RSS score curves for all 34 evaluated models plus the baseline, arranged in descending order by average RSS score across $L = 4, 8, \ldots, 16$. Each subplot shows the degradation of RSS as sequence length increases from $L = 4$ to $L = 16$. The order matches Table 18.