Quantifying Positional Biases in Text Embedding Models

Reagan J. Lee
University of California, Berkeley
Berkeley, CA 94704
reaganjlee@berkeley.edu

Samarth Goel
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94704
sgoel9@berkeley.edu

Kannan Ramchandran
University of California, Berkeley
Berkeley, CA 94704
kannanr@eecs.berkeley.edu

Abstract

Embedding models are crucial for tasks in Information Retrieval (IR) and semantic similarity measurement, yet their handling of longer texts and associated positional biases remains underexplored. In this study, we investigate the impact of content position and input size on text embeddings. Our experiments reveal that embedding models, irrespective of their positional encoding mechanisms, disproportionately prioritize the beginning of an input. Ablation studies demonstrate that insertion of irrelevant text or removal at the start of a document reduces cosine similarity between altered and original embeddings by up to 12.3% more than ablations at the end. Regression analysis further confirms this bias, with sentence importance declining as position moves further from the start, even with with content-agnosticity. We hypothesize that this effect arises from pre-processing strategies and chosen positional encoding techniques. These findings quantify the sensitivity of retrieval systems and suggest a new lens towards embedding model robustness.

1 Introduction

Embedding models are increasingly used to encode text in critical applications like document search systems. However, their effectiveness diminishes when dealing with long-context inputs, particularly in larger documents that cannot entirely fit into these models' context windows. To address these limitations, techniques such as document chunking are used to segment large documents into smaller pieces of text as model inputs [37]. Despite its utility, research into optimal chunking strategies is still an emerging field and improvements can often be highly domain-specific or underexplored in practical environments. [34].

In this study, we investigate the influence of content position and input size on the resulting text embedding vector from eight embedding models. Our findings reveal a systematic bias in which embedding models, regardless of their positional encoding mechanisms, disproportionately weigh the

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

beginning of a text input. This results in greater importance being assigned to the initial sentences of multi-sentence or long-context inputs. To demonstrate this, we conducted two types of ablation studies: one involving the insertion of irrelevant text ("needles") at different positions in the document [11], and another involving the removal of varying text chunks. We observe that inserting irrelevant text at the beginning of a document reduces the cosine similarity between the altered and original document embeddings by up to 8.5% more than when inserted in the middle, and 12.3% more than when inserted at the end. Similarly, removal experiments show that the largest decreases in similarity occur when text is removed from the beginning of the document.

To further explore this bias, we employ regression analysis to measure sentence-level importance on a complete document-level embedding, isolating model position bias from human writing patterns. Our analysis shows a significant decline in regression coefficients as the sentence position moves further from the beginning of the document, reinforcing the bias toward earlier content. To rule out dataset-specific effects, we repeat all experiments with randomly shuffled sentences and obtain similar results, confirming that this bias arises from the model's internal mechanisms rather than document structure.

We hypothesize that this bias stems from common pre-processing strategies, particularly truncation, used during training when the input exceeds the model's context window [16, 32]. This has important implications for real-world retrieval tasks, where documents with key information located later in the text may be overlooked due to the model's disproportionate weighting of early content [2].

We conclude by discussing the broader implications of these biases in embedding models and highlight the need for future research to develop methods that can better handle the entirety of long-context inputs without disproportionately prioritizing the beginning.

2 Background

2.1 Bidirectional encoding in embedding models

Embedding models, particularly those utilizing transformer encoder architectures [29], employ layers of bidirectional self-attention blocks to process text [6]. These models are distinct from decoders in that they generate a fixed-length vector representing the entire input text. This is achieved by producing an output matrix $L \times D$ (where L is the sequence length and D is the dimensionality of the embeddings), and then applying either mean or max pooling across the L dimension [21]. Such pooling operations are position-invariant, theoretically suggesting an unbiased treatment of input positions in terms of attention and representation [24].

We use cosine similarity to compare the output embeddings from these models, especially to study the effects of textual modifications such as insertions or deletions. Cosine similarity measures the cosine of the angle between two vectors, thus providing a scale- and orientation-invariant metric to assess the similarity between two text representations [15]. Due to the invariance of the architecture and similarity measurement we employ, the last systematic source of bias stems from learned positional embeddings used in our models and the models' training methodology, which are heavily connected.

2.2 Positional Encoding Techniques

Absolute Positional Embedding (APE) assigns fixed position-specific vectors based off of position id to each token embedding. This was first popularized by BERT [6] and remains the most common technique to add positional information in encoder-style models today.

Rotary Positional Embedding (RoPE): RoPE encodes positions by applying a rotation to each token's embedding in the 2D subspaces of the embedding space. For each embedding vector x, it applies a rotation matrix $R(\theta)$ based on the position pos:

$$\mathbf{x}_{\text{pos}}^{(2i)} = \mathbf{x}^{(2i)} \cos(\theta_{\text{pos}}) - \mathbf{x}^{(2i+1)} \sin(\theta_{\text{pos}})$$

$$\mathbf{x}_{\text{pos}}^{(2i+1)} = \mathbf{x}^{(2i)} \sin(\theta_{\text{pos}}) + \mathbf{x}^{(2i+1)} \cos(\theta_{\text{pos}})$$

where $\theta_{\text{pos}} = \frac{\text{pos}}{10000^{2i/d}}$, i indexes the embedding dimensions, and d is the dimensionality.

3. **Attention with Linear Biases (ALiBi)**: ALiBi introduces a relative bias into the attention scores rather than modifying the embeddings. The bias is linear with respect to the distance between tokens. The attention score $A(i, j)$ between token i and token j is modified by adding a bias term $m(|i - j|)$, where $|i - j|$ is the distance between tokens:

$$A(i, j) = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}} + m(|i - j|)$$

where $m(|i - j|)$ is a linear function of the relative distance between tokens i and j, and d_k is the dimensionality of the key vectors.

2.3 Noise from Document Chunking for IR Tasks

In practical applications, documents often exceed the context length capabilities of embedding models, necessitating chunking strategies like naive, recursive, or semantic chunking [7, 8]. This process divides a document into smaller pieces that fit within a model's context window, then embeds each chunk separately for insertion into a vector database [13] and downstream use in Retrieval-Augmented Generation (RAG) [14] tasks. This causes an unintentional, outsized amount of noise in the beginning and end of documents as a function of selected chunking strategies.

2.4 Embedding Models Robustness

The performance of decoder models has been shown to vary significantly with the position of content within the model's context window, with pronounced degradation observed for inputs that exceed the context length seen during training [17]. Positional encoding methods have been studied to address these challenges from both decreasing the effect of content position within training context length[36], and generalizing to longer contexts from itself[1]. However, these works exhibit limitations: The former provides limited analysis of diverse encoding mechanisms, and the latter emphasizes generalization to longer inputs rather than robustness to positional shifts.

Moreover, both studies focus exclusively on decoder-only architectures, whose causal attention mask provides the ability for the model to generalize without explicit positional information itself[1], and remains underexplored as a research direction. Existing work on embedding model robustness predominantly centers on improving training data quality or diversity[33], with relatively little attention paid to architectural components such as positional encoding mechanisms.

3 Effect of sentence-level positioning in embedding output

We explore how the position and size of a sentence in a text influence a document's final embedding vector. Our methodology adapts the needle-in-a-haystack test [11], traditionally used for generative models in information retrieval [26], to evaluate embedding models.

3.1 Experimental setup

We investigate the impact of adding irrelevant or adversarial text ("needle") to a document. After inserting the needle, we generate a new embedding for the altered text and compare it to the original using cosine similarity. We vary the needle's length (5%, 10%, 25%, 50%, and 100% of the original text's token count) and position (beginning, middle, end) across 15 experimental conditions. We use an extended version of Lorem Ipsum placeholder text [27] that exceeds the length of our longest datapoint and is structured in paragraph format to achieve a needle with structural similarity to our data while avoiding a confounding effect on the embedding model.

In a parallel experiment, we remove portions of text (10%, 25%, 50% of sentences, rounded up) from different positions (beginning, middle, end) in the document. The resulting text is then embedded, and its similarity to the original embedding is measured using cosine similarity. We test various models, segmented by their positional encodings, to demonstrate the consistency of our results across multiple popular embedding models. We used six open-source models utilizing various positional encoding methods - BGE-m3 [3] and E5-Large-V2 [31] using APE; Nomic-Embed-Text-v1.5 [18] and E5-RoPE base [37] using RoPE; and Jina-Embeddings-v2-Base [12] and Mosaic-Bert-Base

(sequence length 1024) [19] using ALiBi. We additionally test Cohere's Embed-English-v3.0 [20] due to their popularity and real-world applicability. Although we picked these models due to their varying positional encoding methods and performance, we acknowledge these may not generalize to other architectures and datasets. Context lengths or additional information such as parameter counts or benchmark performance for these models can be found in Appendix A. For texts exceeding these limits, we truncate from the end to fit the models' context windows. For datasets, we use 200 examples each from the PubMed Publications [4], Paul Graham Essay Collection [10], Amazon Reviews [35], Argumentative Analysis [30], and Reddit Posts [9] datasets, selected for their range of writing categorizations and lengths. More details on these datasets can be found in appendix B.

3.2 Results and discussion

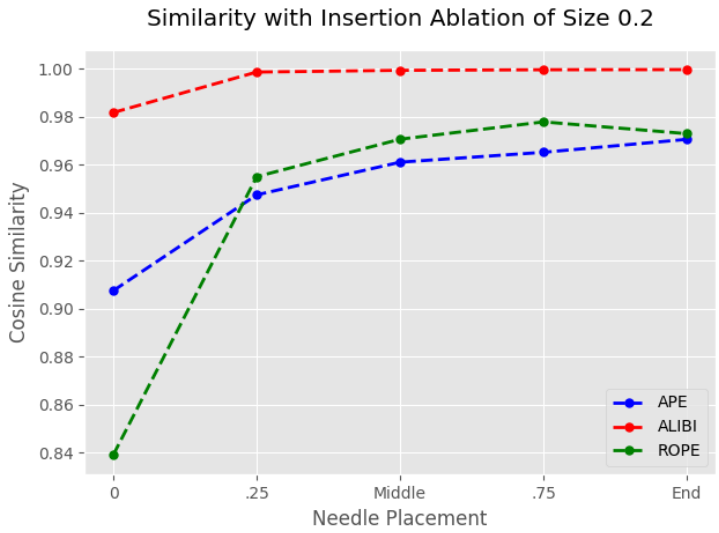


Figure 1: Cosine similarity vs. needle size and position

Our results indicate a pronounced drop in similarity when irrelevant text is inserted at the beginning of documents, with less impact observed when additions occur in the middle or end. Specifically, for APE models, introducing an insertion equal to 20% of the total content at the beginning results in an average cosine similarity of 0.885, compared to 0.963 at the end—a relative decrease of approximately 8%. RoPE-based models show a stronger sensitivity to this disruption, with cosine similarity dropping to 0.819 at the beginning, a 15.4% decrease compared to the 0.968 similarity at the end. By contrast, AliBi models are the most robust, maintaining a high cosine similarity of 0.981 at the beginning and 0.999 at the end, reflecting only a 1.8% decrease. This suggests that earlier positions in the input sequence play a more critical role in model performance, and different positional encoding methods vary in their resilience to this type of input perturbation.

This trend persists across all insertion sizes, with larger insertions intensifying the drop in similarity. Even though the magnitude of the degradation varies by model, we find the trend robust to model differences. Across all five models tested, the average decrease in cosine similarity is approximately 7%, indicating a consistent pattern of sensitivity to input alterations at the beginning of the sequence.

Additionally, we observe that removal ablations yield similar results, although the overall similarity scores are higher in comparison to insertion ablations. This suggests that while the models are affected by both insertion and removal disruptions, the impact of irrelevant insertions at the beginning of sequences may introduce greater noise into the representations.

Similar trends are observed in the removal experiments, where the largest impacts on similarity occur when sentences are removed from the beginning. Removing half of the sentences from the beginning results in a median similarity that is 10.6% lower than when sentences are removed from the end, with no significant difference between middle and end removals—unlike the insertion experiments. Interestingly, even a 50% text removal from the middle maintains a median similarity

of 95%, corroborating our findings from the insertion experiments, where a large drop in similarity was expected but not observed. These results suggest that while the position of removed content has a clear impact, it is somewhat less disruptive than insertions.

4 Analysis of embedding decomposition

Recent advancements in embedding interpretability have demonstrated that certain dimensions in high-dimensional semantic spaces may correspond to specific linguistic or semantic features, such as sentiment or subject matter [5]. Further research has shown that vector operations, such as adding embeddings, can produce new vectors that represent the semantic meaning of their components [22].

Building from these works, we explore the impact of sentence-level positioning on the final document embedding vector through regression analysis, which offers a more direct method to quantify the contribution of individual sentences to a document's embedding representation.

Human writing often emphasizes key information at the beginning and end of documents, a technique that may introduce biases in datasets and reason for embeddings to skew towards these positions. To address these, we employ additional data augmentation and ablation techniques aimed at isolating and understanding these effects, to ensure that our findings more accurately reflect model behavior rather than dataset peculiarities.

4.1 Reconstructing embedding vectors through linear combinations of constituents

To start, we wanted to validate the assumption that the sentence embeddings of a larger document can meaningfully be used as a proxy for the original document embedding [28]. To test this, we wanted to determine how much reconstruction loss we would incur from using an optimal linear combination of sentence embedding vectors instead of a full multi-sentence embedding vector. Optimizing for train R^2, we use Ordinary Least Squares (OLS) regression to reconstruct the document embedding from its sentence embeddings, with the multi-sentence embedding vector as our response and each sentence vector as a predictive datapoint for our regression. Our model choice is notable for its direct interpretability [25], though we acknowledge and check for potential issues posed by OLS, such as multicollinearity. Our regressions use normalized embeddings (L2 norm of 1) to ensure scale invariance [23]. We separate our data points into their component sentences by use of punctuation such as periods, and new lines.

When we regress the sentence embedding vectors onto the multi-sentence embedding vector, we find that our train R^2 across the eight models and five datasets we used ranges from 0.75 to 0.99, with an average R^2 or 0.876 when reconstructing the multi-sentence embedding vector. This result indicates that approximately 87.6% of the variance in a long-content document embedding can be accounted for by analyzing the embeddings of the individual sentences constituting the document. The Mean Squared Error (MAE) summed over all dimensions of this reconstruction across all models and datasets ranged from 0.001 and 0.01 with an average of 0.0069, suggesting minimal deviation in the reconstructed vectors.

4.2 Analyzing regression coefficients as importance weights

Given the high explanatory power of our regression models, the coefficients given to each sentence (datapoint) in our regression are strong indicators to determine their relative importance to the total document. To standardize our comparisons across documents, we standardized each coefficient vector by its L2 norm. One potential issue to note with this approach is the presence of negative coefficient values, but these tended to be rare and very low in magnitude, with very little influence on our final analysis.

We judge the importance of a sentence by its regression coefficient. For example, if a regression on a two-sentence document yielded weights 0.8 and 0.6, we conclude that the first sentence is 33.3% more important to the final semantic meaning of the text than the second sentence.

As shown in Figure 2, there is a downward trend in coefficient values with increasing sentence position, suggesting a positional bias where earlier sentences generally have a greater impact on the document's overall semantic representation. To quantify this observation, we plot regression coefficients against sentence positions over all the documents in our dataset.

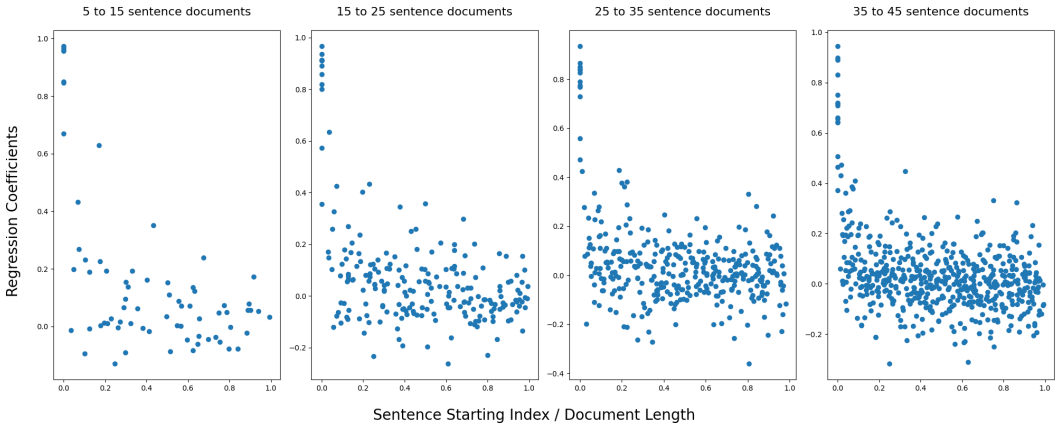


Figure 2: Regression coefficients vs. sentence position, bucketed by document length

Table 1: Correlation and statistical significance of sentence position against shuffled text

Positional Encoding	**Correlation**	**P-value**
APE	-0.127657	2.233374e-103
RoPE	-0.115861	2.259581e-85
ALiBi	-0.07615	9.205763e-38

4.3 Embedding positional bias is robust to human-level writing bias

To validate that this observed bias is not solely a byproduct of dataset-specific characteristics, namely human-level writing bias, we conducted additional regression experiments where all sentences from the above pre-processing steps were shuffled before their embeddings were generated. Using these new embeddings, remarkably, the results mirrored the original findings, with the randomly selected first sentence in the shuffled document consistently receiving a higher weight, thereby disambiguating our results from potential dataset biases.

More specifically, we expect the weight assigned to the first sentence to follow a uniform weight of $\frac{1}{\text{num\_sentences}}$. However, this analysis shows a distinct negative correlation between sentence position and importance score, with significant deviations from the expected uniform distribution ($\alpha \ll 0.001$), confirming a systematic positional influence within document embeddings as shown in table 1. These findings suggest that the embedding models may inherently prioritize the initial information presented in any text sequence, irrespective of its original position in the document.

5 Isolating the role of training methodology in model biases

During training, input data is processed sequentially, starting at the beginning of the context window. Variable-length training samples are packed into this fixed window, often necessitating truncation when the input exceeds the window's length. Truncation typically discards content from the end, leading to a systematic bias where earlier positions in the sample receive disproportionate attention.

For a given position $i \in$ [0, N] within a context window of length N, the model observes t_i, the number of non-padding tokens encountered at position i. The importance of position i can then be modeled as $imp(t_i) = u(t_i)$, where $u(\cdot)$ represents the model's updates based on the presence of non-padding tokens at t_i.

As traditional truncation favors earlier positions, the frequency with which tokens are seen at the beginning of the context window is inherently higher than at the end. This can be modeled as a monotonically decreasing function, where the quantity of non-padding tokens at t_i diminishes as i increases. As a result, the relative importance of earlier positions $imp(t_1) \geq imp(t_2) \geq \cdots \geq imp(t_N)$ is systematically higher, introducing an implicit bias that prioritizes early context over later content.

Although this monotonic impact on position can theoretically be removed by maintaining an equal number of effective updates throughout the context, it is unknown what the impacts on computational costs, and model performance would be. Future pre-training, as well as employing novel context-length enhancement methods, with this bias in mind will require additional research to fully understand the impacts, leading us to believe that this bias will continue in future models.

6 Conclusion

Our study uncovers a positional bias in embedding models, where sentences at the beginning of a document disproportionately influence the resulting embeddings. This bias is consistently observed across various models with different context sizes and datasets and is evident in both text insertion and removal experiments. We further quantified this effect through regression analysis, which highlights the extent of the model's preference for earlier content. Our findings suggest that this bias is intrinsic to the models' training methodologies, particularly the use of truncation strategies, rather than a consequence of dataset-specific patterns.

This positional bias poses challenges in critical applications like information retrieval in document search systems, highlighting the need for alternative positional encoding methods to mitigate these biases and achieve more balanced semantic representations. Additionally, growing research into extending context lengths offers a promising avenue for further exploration of this phenomenon and potential solutions.

References

[1] K. N. R. P. D. S. R. Amirhossein Kazemnejad, Inkit Padhi. The impact of positional encoding on length generalization in transformers, 2023.

[2] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek. Seven failure points when engineering a retrieval augmented generation system, 2024.

[3] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[4] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian. A discourse-aware attention model for abstractive summarization of long documents, 2018.

[5] G. Dar, M. Geva, A. Gupta, and J. Berant. Analyzing transformers in embedding space, 2023.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] W. Fei, X. Niu, P. Zhou, L. Hou, B. Bai, L. Deng, and W. Han. Extending context window of large language models via semantic compression, 2023.

[8] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[9] G. Geigle, N. Reimers, A. Rücklé, and I. Gurevych. Tweac: Transformer with extendable qa agent classifiers, 2021.

[10] S. Goel. paul graham essays (revision 0c7155a), 2024. URL `https://huggingface.co/datasets/sgoel9/paul_graham_essays`.

[11] N. M. Guerreiro, E. Voita, and A. F. T. Martins. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation, 2023.

[12] M. Günther, J. Ong, I. Mohr, A. Abdessalem, T. Abel, M. K. Akram, S. Guzman, G. Mastrapas, S. Sturua, B. Wang, M. Werk, N. Wang, and H. Xiao. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents, 2024.

[13] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus, 2017.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[15] X. Li and J. Li. Angle-optimized text embeddings, 2024.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[17] J. H. A. P. M. B. F. P. P. L. Nelson F. Liu, Kevin Lin. Lost in the middle: How language models use long contexts, 2023.

[18] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.

[19] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022.

[20] N. Reimers. Introducing embed v3, nov 2023. URL `https://cohere.com/blog/introducing-embed-v3`. Accessed: 2024-05-18.

[21] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[22] L. K. Senel, I. Utlu, V. Yucesoy, A. Koc, and T. Cukur. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779, Oct. 2018. ISSN 2329-9304. doi: 10.1109/taslp.2018.2837384. URL `http://dx.doi.org/10.1109/TASLP.2018.2837384`.

[23] H. Steck, C. Ekanadham, and N. Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24. ACM, May 2024. doi: 10.1145/3589335.3651526. URL `http://dx.doi.org/10.1145/3589335.3651526`.

[24] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

[25] T. Słoczyński. Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights, 2020.

[26] G. Team, M. Reid, N. Savinov, D. Teplyashin, Dmitry, Lepikhin, T. Lillicrap, J. baptiste Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. Dai, K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy, J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski, Z. Abbas, N. Schucher, A. Anand, R. Ives, J. Keeling, K. Lenc, S. Haykal, S. Shakeri, P. Shyam, A. Chowdhery, R. Ring, S. Spencer, E. Sezener, L. Vilnis, O. Chang, N. Morioka, G. Tucker, C. Zheng, O. Woodman, N. Attaluri, T. Kocisky, E. Eltyshev, X. Chen, T. Chung, V. Selo, S. Brahma, P. Georgiev, A. Slone, Z. Zhu, J. Lottes, S. Qiao, B. Caine, S. Riedel, A. Tomala, M. Chadwick, J. Love, P. Choy, S. Mittal, N. Houlsby, Y. Tang, M. Lamm, L. Bai, Q. Zhang, L. He, Y. Cheng, P. Humphreys, Y. Li, S. Brin, A. Cassirer, Y. Miao, L. Zilka, T. Tobin, K. Xu, L. Proleev, D. Sohn, A. Magni, L. A. Hendricks, I. Gao, S. Ontanon, O. Bunyan, N. Byrd, A. Sharma, B. Zhang, M. Pinto, R. Sinha, H. Mehta, D. Jia, S. Caelles, A. Webson, A. Morris, B. Roelofs, Y. Ding, R. Strudel, X. Xiong, M. Ritter, M. Dehghani, R. Chaabouni, A. Karmarkar, G. Lai, F. Mentzer, B. Xu, Y. Li, Y. Zhang, T. L. Paine, A. Goldin, B. Neyshabur, K. Baumli, A. Levskaya, M. Laskin, W. Jia, J. W. Rae, K. Xiao, A. He, S. Giordano, L. Yagati, J.-B. Lespiau, P. Natsev, S. Ganapathy, F. Liu, D. Martins, N. Chen, Y. Xu, M. Barnes, R. May, A. Vezer, J. Oh, K. Franko, S. Bridgers, R. Zhao, B. Wu, B. Mustafa, S. Sechrist, E. Parisotto, T. S. Pillai, C. Larkin, C. Gu, C. Sorokin, M. Krikun, A. Guseynov, J. Landon, R. Datta, A. Pritzel, P. Thacker, F. Yang, K. Hui, A. Hauth, C.-K. Yeh, D. Barker, J. Mao-Jones, S. Austin, H. Sheahan, P. Schuh, J. Svensson, R. Jain, V. Ramasesh, A. Briukhov, D.-W. Chung, T. von Glehn, C. Butterfield, P. Jhakra, M. Wiethoff, J. Frye, J. Grimstad, B. Changpinyo, C. L. Lan, A. Bortsova, Y. Wu, P. Voigtlaender, T. Sainath, S. Gu, C. Smith, W. Hawkins, K. Cao, J. Besley,

S. Srinivasan, M. Omernick, C. Gaffney, G. Surita, R. Burnell, B. Damoc, J. Ahn, A. Brock, M. Pajarskas, A. Petrushkina, S. Noury, L. Blanco, K. Swersky, A. Ahuja, T. Avrahami, V. Misra, R. de Liedekerke, M. Iinuma, A. Polozov, S. York, G. van den Driessche, P. Michel, J. Chiu, R. Blevins, Z. Gleicher, A. Recasens, A. Rrustemi, E. Gribovskaya, A. Roy, W. Gworek, S. M. R. Arnold, L. Lee, J. Lee-Thorp, M. Maggioni, E. Piqueras, K. Badola, S. Vikram, L. Gonzalez, A. Baddepudi, E. Senter, J. Devlin, J. Qin, M. Azzam, M. Trebacz, M. Polacek, K. Krishnakumar, S. yiin Chang, M. Tung, I. Penchev, R. Joshi, K. Olszewska, C. Muir, M. Wirth, A. J. Hartman, J. Newlan, S. Kashem, V. Bolina, E. Dabir, J. van Amersfoort, Z. Ahmed, J. Cobon-Kerr, A. Kamath, A. M. Hrafnkelsson, L. Hou, I. Mackinnon, A. Frechette, E. Noland, X. Si, E. Taropa, D. Li, P. Crone, A. Gulati, S. Cevey, J. Adler, A. Ma, D. Silver, S. Tokumine, R. Powell, S. Lee, K. Vodrahalli, S. Hassan, D. Mincu, A. Yang, N. Levine, J. Brennan, M. Wang, S. Hodkinson, J. Zhao, J. Lipschultz, A. Pope, M. B. Chang, C. Li, L. E. Shafey, M. Paganini, S. Douglas, B. Bohnet, F. Pardo, S. Odoom, M. Rosca, C. N. dos Santos, K. Soparkar, A. Guez, T. Hudson, S. Hansen, C. Asawaroengchai, R. Addanki, T. Yu, W. Stokowiec, M. Khan, J. Gilmer, J. Lee, C. G. Bostock, K. Rong, J. Caton, P. Pejman, F. Pavetic, G. Brown, V. Sharma, M. Lučić, R. Samuel, J. Djolonga, A. Mandhane, L. L. Sjösund, E. Buchatskaya, E. White, N. Clay, J. Jiang, H. Lim, R. Hemsley, Z. Cankara, J. Labanowski, N. D. Cao, D. Steiner, S. H. Hashemi, J. Austin, A. Gergely, T. Blyth, J. Stanton, K. Shivakumar, A. Siddhant, A. Andreassen, C. Araya, N. Sethi, R. Shivanna, S. Hand, A. Bapna, A. Khodaei, A. Miech, G. Tanzer, A. Swing, S. Thakoor, L. Aroyo, Z. Pan, Z. Nado, J. Sygnowski, S. Winkler, D. Yu, M. Saleh, L. Maggiore, Y. Bansal, X. Garcia, M. Kazemi, P. Patil, I. Dasgupta, I. Barr, M. Giang, T. Kagohara, I. Danihelka, A. Marathe, V. Feinberg, M. Elhawaty, N. Ghelani, D. Horgan, H. Miller, L. Walker, R. Tanburn, M. Tariq, D. Shrivastava, F. Xia, Q. Wang, C.-C. Chiu, Z. Ashwood, K. Baatarsukh, S. Samangooei, R. L. Kaufman, F. Alcober, A. Stjerngren, P. Komarek, K. Tsihlas, A. Boral, R. Comanescu, J. Chen, R. Liu, C. Welty, D. Bloxwich, C. Chen, Y. Sun, F. Feng, M. Mauger, X. Dotiwalla, V. Hellendoorn, M. Sharman, I. Zheng, K. Haridasan, G. Barth-Maron, C. Swanson, D. Rogozińska, A. Andreev, P. K. Rubenstein, R. Sang, D. Hurt, G. Elsayed, R. Wang, D. Lacey, A. Ilić, Y. Zhao, A. Iwanicki, A. Lince, A. Chen, C. Lyu, C. Lebsack, J. Griffith, M. Gaba, P. Sandhu, P. Chen, A. Koop, R. Rajwar, S. H. Yeganeh, S. Chang, R. Zhu, S. Radpour, E. Davoodi, V. I. Lei, Y. Xu, D. Toyama, C. Segal, M. Wicke, H. Lin, A. Bulanova, A. P. Badia, N. Rakićević, P. Sprechmann, A. Filos, S. Hou, V. Campos, N. Kassner, D. Sachan, M. Fortunato, C. Iwuanyanwu, V. Nikolaev, B. Lakshminarayanan, S. Jazayeri, M. Varadarajan, C. Tekur, D. Fritz, M. Khalman, D. Reitter, K. Dasgupta, S. Sarcar, T. Ornduff, J. Snaider, F. Huot, J. Jia, R. Kemp, N. Trdin, A. Vijayakumar, L. Kim, C. Angermueller, L. Lao, T. Liu, H. Zhang, D. Engel, S. Greene, A. White, J. Austin, L. Taylor, S. Ashraf, D. Liu, M. Georgaki, I. Cai, Y. Kulizhskaya, S. Goenka, B. Saeta, Y. Xu, C. Frank, D. de Cesare, B. Robenek, H. Richardson, M. Alnahlawi, C. Yew, P. Ponnapalli, M. Tagliasacchi, A. Korchemniy, Y. Kim, D. Li, B. Rosgen, K. Levin, J. Wiesner, P. Banzal, P. Srinivasan, H. Yu, Çağlar Ünlü, D. Reid, Z. Tung, D. Finchelstein, R. Kumar, A. Elisseeff, J. Huang, M. Zhang, R. Aguilar, M. Giménez, J. Xia, O. Dousse, W. Gierke, D. Yates, K. Jalan, L. Li, E. Latorre-Chimoto, D. D. Nguyen, K. Durden, P. Kallakuri, Y. Liu, M. Johnson, T. Tsai, A. Talbert, J. Liu, A. Neitz, C. Elkind, M. Selvi, M. Jasarevic, L. B. Soares, A. Cui, P. Wang, A. W. Wang, X. Ye, K. Kallarackal, L. Loher, H. Lam, J. Broder, D. Holtmann-Rice, N. Martin, B. Ramadhana, M. Shukla, S. Basu, A. Mohan, N. Fernando, N. Fiedel, K. Paterson, H. Li, A. Garg, J. Park, D. Choi, D. Wu, S. Singh, Z. Zhang, A. Globerson, L. Yu, J. Carpenter, F. de Chaumont Quitry, C. Radebaugh, C.-C. Lin, A. Tudor, P. Shroff, D. Garmon, D. Du, N. Vats, H. Lu, S. Iqbal, A. Yakubovich, N. Tripuraneni, J. Manyika, H. Qureshi, N. Hua, C. Ngani, M. A. Raad, H. Forbes, J. Stanway, M. Sundararajan, V. Ungureanu, C. Bishop, Y. Li, B. Venkatraman, B. Li, C. Thornton, S. Scellato, N. Gupta, Y. Wang, I. Tenney, X. Wu, A. Shenoy, G. Carvajal, D. G. Wright, B. Bariach, Z. Xiao, P. Hawkins, S. Dalmia, C. Farabet, P. Valenzuela, Q. Yuan, A. Agarwal, M. Chen, W. Kim, B. Hulse, N. Dukkipati, A. Paszke, A. Bolt, K. Choo, J. Beattie, J. Prendki, H. Vashisht, R. Santamaria-Fernandez, L. C. Cobo, J. Wilkiewicz, D. Madras, A. Elqursh, G. Uy, K. Ramirez, M. Harvey, T. Liechty, H. Zen, J. Seibert, C. H. Hu, A. Khorlin, M. Le, A. Aharoni, M. Li, L. Wang, S. Kumar, N. Casagrande, J. Hoover, D. E. Badawy, D. Soergel, D. Vnukov, M. Miecnikowski, J. Simsa, P. Kumar, T. Sellam, D. Vlasic, S. Daruki, N. Shabat, J. Zhang, G. Su, J. Zhang, J. Liu, Y. Sun, E. Palmer, A. Ghaffarkhah, X. Xiong, V. Cotruta, M. Fink, L. Dixon, A. Sreevatsa, A. Goedeckemeyer, A. Dimitriev, M. Jafari, R. Crocker, N. FitzGerald, A. Kumar, S. Ghemawat, I. Philips, F. Liu, Y. Liang, R. Sterneck, A. Repina, M. Wu, L. Knight, M. Georgiev, H. Lee, H. Askham,

A. Chakladar, A. Louis, C. Crous, H. Cate, D. Petrova, M. Quinn, D. Owusu-Afriyie, A. Singhal, N. Wei, S. Kim, D. Vincent, M. Nasr, C. A. Choquette-Choo, R. Tojo, S. Lu, D. de Las Casas, Y. Cheng, T. Bolukbasi, K. Lee, S. Fatehi, R. Ananthanarayanan, M. Patel, C. Kaed, J. Li, S. R. Belle, Z. Chen, J. Konzelmann, S. Põder, R. Garg, V. Koverkathu, A. Brown, C. Dyer, R. Liu, A. Nova, J. Xu, A. Walton, A. Parrish, M. Epstein, S. McCarthy, S. Petrov, D. Hassabis, K. Kavukcuoglu, J. Dean, and O. Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[27] R. C. Timmer, D. Liebowitz, S. Nepal, and S. Kanhere. Tsm: Measuring the enticement of honeyfiles with natural language processing, 2022.

[28] H. Tsukagoshi, R. Sasano, and K. Takeda. Comparison and combination of sentence embeddings derived from different supervision signals, 2022.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[30] H. Wachsmuth, S. Syed, and B. Stein. Retrieval of the best counterargument without prior topic knowledge. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1023. URL https://aclanthology.org/P18-1023.

[31] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022.

[32] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[33] K. H. Yichen Yang, Xiaosen Wang. Robust textual embedding against word-level adversarial attacks, 2022.

[34] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez. Raft: Adapting language model to domain specific rag, 2024.

[35] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification, 2016.

[36] S. L. Z. Y. O. R. B. C. X. W. Z. W. Zhenyu Zhang, Runjin Chen. Found in the middle: How language models use long contexts better via plug-and-play positional encoding, 2024.

[37] D. Zhu, L. Wang, N. Yang, Y. Song, W. Wu, F. Wei, and S. Li. Longembed: Extending embedding models for long context retrieval, 2024.

A Model details

Embed-English-v3.0 [20] has a content length of 512 tokens and an unknown number of parameters. The model is accessed via the Cohere API.

BGE-m3 [3] has a content length of 8912 tokens, is comprised of 568M parameters, and was trained using the APE positional encoding method.

E5-Large-V2 [31] has a content length of 512 tokens, is comprised of 335M parameters, and was trained using the APE positional encoding method.

Nomic-Embed-Text-v1.5 [18] has a content length of 8192 tokens, is comprised of 137M parameters, and was trained using the RoPE positional encoding method.

E5-RoPE-base [37] has a content length of 512 tokens, is comprised of 108M parameters, and was trained using the RoPE positional encoding method.

Jina-Embeddings-v2-Base[12] has a content length of 8192 tokens, is comprised of 137M parameters, and was trained using the ALiBi positional encoding method.

Mosaic-Bert-Base [19] has a content length of 1024 tokens, is comprised of 110M parameters, and was trained using the ALiBi positional encoding method.

B Dataset details

PubMed Publications [4]: We use PubMed publication abstracts to assess the impact of our ablations on scientific writing. Scientific texts are characterized by their structured presentation of information and specialized vocabulary. Understanding how embeddings capture this complexity can provide insights into their utility in academic and research applications. This dataset is comprised of 270,000 datapoints.

Paul Graham Essay Collection [10]: We analyze over 200 essays written by Paul Graham, varying from 400 to 70,000 words. Paul Graham's essays are known for their thoughtful, reflective style and coherent argument structure, making them ideal for studying how embeddings handle nuanced and complex idea development over long texts. This dataset is comprised of 215 datapoints.

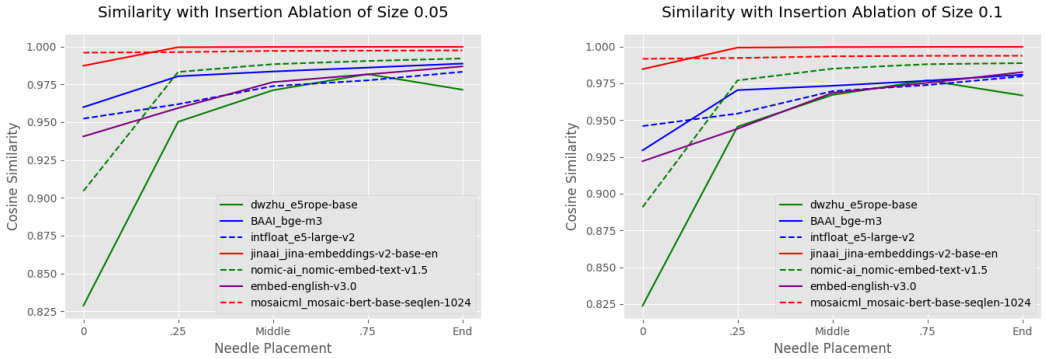
Amazon Reviews [35]: Drawn from MTEB's Amazon Polarity dataset, this helps us examine consumer review text. Reviews are direct and opinion-rich, offering a perspective on how embeddings process everyday language and sentiment, which is crucial for applications in consumer analytics. This dataset is comprised of 4 million datapoints.

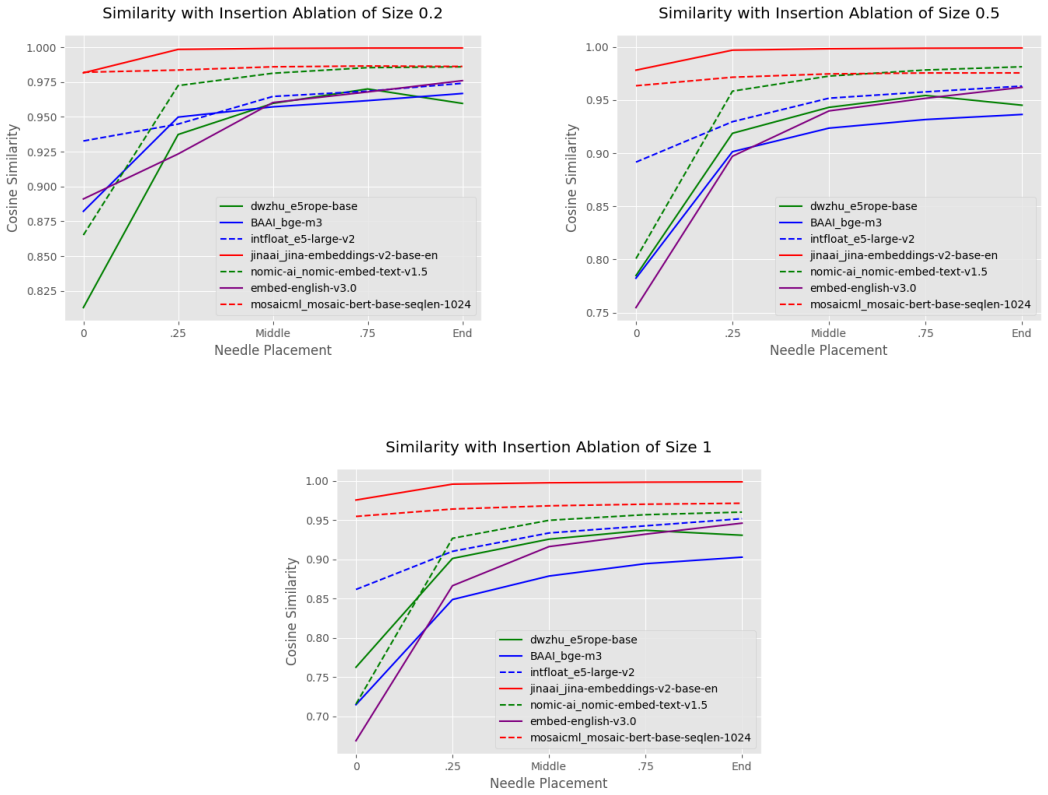
Argumentative Analysis [30]: From the BiER benchmark's Argumentative Analysis (ArguAna) dataset, we explore embeddings of formal persuasive writing. This dataset includes well constructed arguments that are ideal for testing how embeddings capture logical structure and the effectiveness of rhetoric. This dataset is comprised of 10,000 datapoints.

Reddit Posts [9]: More Informal and diverse writing styles can be found on Reddit. This dataset introduces grammar, style, and subject matter diversity into our tests, extending our findings to be more robust and adaptable to a wide range of writing styles. This dataset is comprised of 450,000 datapoints.

C Cosine similarities across insertion ablation sizes and datasets

The following are the results of running insertion and removal ablations of given sizes on input examples. These are the results of the average cosine similarity across all datasets.





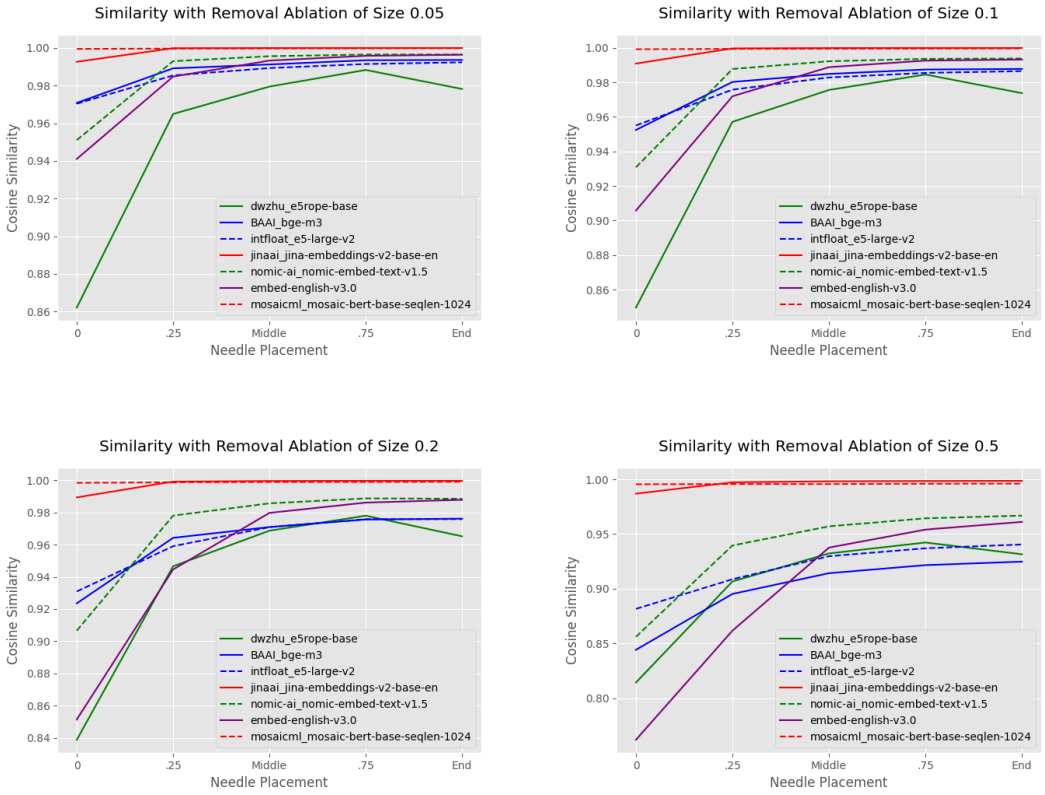
D Cosine similarities across deletion of ablation sizes and datasets

E Sentence Position Against Shuffled Text

Three sentence length range buckets (65-75, 75-85, 95-105) were omitted due to small sample size (n=6). Examples with less than 5 sentences each were omitted.

Table 2: ALiBi

Sentence Length Range	**Correlation**	**P-value**	**Number of Samples**
5-15	-0.120560	1.037594e-24	904
15-25	-0.083780	2.757708e-05	132
25-35	-0.015695	5.596307e-01	48
35-45	-0.037581	5.387906e-02	66
45-55	-0.008077	4.455038e-01	178
55-65	-0.019355	1.426657e-01	98



NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe the main sections in the paper and summarize them to fit into the abstract guidelines.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included limitations in each of our experimental setup sections.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

Table 3: APE

Sentence Length Range	**Correlation**	**P-value**	**Number of Samples**
5-15	-0.204936	1.196681e-88	904
15-25	-0.123513	1.420863e-18	132
25-35	-0.036560	2.585037e-02	48
35-45	-0.034370	7.942209e-04	66
45-55	-0.009526	5.458451e-02	178
55-65	-0.004620	4.229611e-01	98

Table 4: RoPE

Sentence Length Range	**Correlation**	**P-value**	**Number of Samples**
5-15	-0.201598	3.154022e-69	904
15-25	-0.098903	1.669302e-11	132
25-35	-0.044463	8.444829e-03	48
35-45	-0.021359	4.203218e-02	66
45-55	-0.009357	6.387572e-02	178
55-65	-0.008881	1.290475e-01	98

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not provide any theoretical result, and only provide an intuition for explaining our work.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we have provided extensive details on how we conducted our experiments, both in the experimental setup sections, as well as appendicies.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, we submit code to reproduce our findings.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we specify all training details relevant to the experiments such that they can be reproduced easily.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we provide proper evidence for reviewers to see the variance of the data within our analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we describe our compute used, as well as these differences between models that we tested.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the ethics guidelines and have taken all steps to ensure they are being followed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The authors discuss the effects of this work in the later parts of the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release data or models that have a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Code has been marked with the author, unless it was originally created by the authors themselves.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The author releases the code to be able to easily reproduce results.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The experiments do not use human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The authors do not use human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.