

---

# Quantifying Positional Biases in Text Embedding Models

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Embedding models are crucial for tasks in Information Retrieval (IR) and semantic  
2 similarity measurement, yet their handling of longer texts and associated positional  
3 biases remains underexplored. We explore the effect of content position and  
4 size within an embedding model’s input on its final embedding vector, finding a  
5 significant overweighting of initial content in an input. We employ two ablations  
6 to test this effect, inserting irrelevant text into a document and removing text from  
7 a document. We find that perturbations to the beginning of a document reduce its  
8 cosine similarity with the original document by 12.3% more than perturbations done  
9 to the end, with this trend holding across multiple models and datasets. Next, we  
10 attempt to reconstruct a document’s embedding vector from the embeddings of its  
11 sentences, achieving an 85.5%  $R^2$  using a simple linear regression to weight each  
12 sentence’s contribution to the document embedding. Using this finding, we can  
13 assign an importance weight to each sentence and find a -0.55% correlation between  
14 sentence starting index and importance score. We also measure a statistically  
15 significant difference between a sentence’s importance score and the expected  
16 importance of equal weighting amongst all sentences. To ensure our results are not  
17 the effect of dataset bias, we shuffle the sentences in each tested document before  
18 repeating our experiments and see similar results. Finally, we focus on the role of  
19 positional encodings and training methodology on this bias and introduce a data  
20 augmentation scheme we title Position-Aware Data Sampling (PADS) to remedy  
21 these issues. Fine-tuning an embedding model with only 20% of our data using  
22 PADS leads to a 49.6\$ reduction in the differing effect of perturbations done at  
23 the beginning and end of a document, suggesting PADS as an effective avenue  
24 for reducing positional bias in our models and improving the performance of both  
25 retrieval and document processing systems.

## 26 1 Introduction

27 Embedding models are increasingly used to encode text information in a way that aligns semantically  
28 with their intended applications [6, 36]. However, their effectiveness in long-context settings,  
29 particularly how they encode larger documents, remains less explored. Due to the typical limitations  
30 of models’ context windows, techniques like document chunking are employed to fit large documents  
31 into manageable segments [42]. Yet, research into optimal chunking strategies is still emerging,  
32 often leading to preliminary findings that may not provide the most effective results without costly,  
33 domain-specific adjustments [40].

34 This study investigates the influence of content position and size within an embedding model’s input on  
35 the resulting text embedding vector. Our findings indicate that embedding models disproportionately  
36 weigh the beginning of the text, often assigning greater importance to the first sentences of a multi-  
37 sentence or long-context input. To substantiate this observation, we conducted two types of ablation

38 studies: one involving the insertion of irrelevant text ("needles") into the document [11], and another  
39 involving the removal of varying chunks from the document. Our results show that inserting irrelevant  
40 text into the beginning of a text significantly reduces the cosine similarity between the altered and  
41 original document embeddings by up to 8.5% more than insertion in the middle and 12.3% more than  
42 insertion at the end. Removal experiments reinforce this trend, with the largest decreases in similarity  
43 occurring when text is removed from the beginning of a document.

44 We then employ a regression analysis method, finding that using a simple linear regression model to  
45 reconstruct a document embedding vector through the embedding vectors of its constituent sentences  
46 yields a 0.85  $R^2$  score averaged across all documents [31]. This result indicates that we can effectively  
47 back out the contribution of each sentence's embedding vector to the final document embedding vector  
48 by analyzing our regression's weights. We observe a significant decline in regression coefficients as  
49 the position of a sentence within its document increases to be further from the front, underscoring a  
50 systematic favoring towards the initial content of our input. To ensure our results are not the effect of  
51 dataset bias, we repeat all experiments with each document's sentences shuffled in a random order,  
52 and achieve similar results.

53 Next, we delve into potential reasons for this bias, focusing on the role of positional encodings and  
54 data treatment in the training process of embedding models. Most training techniques for embedding  
55 models use simple truncation to pre-process their data if it exceeds the model's context window  
56 [18, 38]. This can have confounding effects on real-world retrieval situations where early sections  
57 of a document may have a disproportionately high similarity despite key information being located  
58 elsewhere [1]. To address this issue we propose a new data augmentation scheme titled Position-  
59 Aware Data Sampling (PADS). PADS randomly samples a consecutive set of tokens within each  
60 document in place of simple truncation, varying the positioning and size of our sample to improve  
61 model robustness. We fine-tune BAAI's BGE-Small-en-v1.5 model on our dataset augmented with  
62 PADS and achieve a 49.6% improvement in closing the gap between model similarity scores for  
63 perturbations at the beginning vs. the end of a document, averaged across both insertion and removal  
64 tasks.

65 We conclude by discussing the implications of embedding models and potential biases they may hold,  
66 emphasizing the need for future research to study the output of embedding models and improve the  
67 processing of long documents.

## 68 2 Background

### 69 2.1 Bidirectional encoding in embedding models

70 Embedding models, particularly those utilizing transformer encoder architectures [34], employ layers  
71 of bidirectional self-attention blocks to process text [6]. These models are distinct from decoders  
72 in that they generate a fixed-length vector representing the entire input text. This is achieved by  
73 producing an output matrix  $L \times D$  (where  $L$  is the sequence length and  $D$  is the dimensionality of  
74 the embeddings), and then applying either mean or max pooling across the  $L$  dimension [25]. Such  
75 pooling operations are position-invariant, theoretically suggesting an unbiased treatment of input  
76 positions in terms of attention and representation [28].

77 The core operation in these models is the attention mechanism, which can be represented mathemati-  
78 cally as:

$$A = \text{softmax} \left( \frac{X^T X}{\sqrt{d}} \right) X^T$$

79 Here,  $X$  is the  $L \times D$  input matrix to the attention mechanism, and  $d$  is the scaling factor derived  
80 from the dimensionality  $D$  of the embeddings. Unlike generative models where a causal attention  
81 mask is used to zero out certain elements in our softmax operation, embedding models are fully  
82 bi-direction and do not employ an attention mask.

83 We use cosine similarity to compare the output embeddings from these models, especially to study  
84 the effects of textual modifications such as insertions or deletions. Cosine similarity measures the  
85 cosine of the angle between two vectors, thus providing a scale- and orientation-invariant metric to  
86 assess the similarity between two text representations [16].

$$\text{cosine}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

87 Due to the invariance of the architecture and similarity measurement we employ, the last systematic  
88 source of bias stems from learned positional embeddings used in our models and the models' training  
89 methodology, which are heavily connected.

## 90 **2.2 Dataset and human-level writing bias**

91 It's important to note that human writing often emphasizes key information at the beginning and  
92 end of documents, a technique that may introduce biases in datasets used for embedding studies.  
93 Such biases could be a reason for embeddings to skew towards these positions. To mitigate this, our  
94 study employs data augmentation and ablation techniques aimed at isolating and understanding these  
95 effects, thereby ensuring that our findings more accurately reflect model behavior rather than dataset  
96 peculiarities.

## 97 **2.3 Document chunking for information retrieval tasks**

98 In practical applications, documents often exceed the context length capabilities of embedding models,  
99 necessitating chunking strategies like naive, recursive, or semantic chunking [7, 8]. This process  
100 divides a document into smaller pieces that fit within a model's context window, then embeds each  
101 chunk separately for insertion into a vector database [13] and downstream use in Retrieval-Augmented  
102 Generation (RAG) [15] tasks. Understanding the impact of chunking on embedding quality and  
103 potential positional biases is essential for optimizing information retrieval strategies.

## 104 **2.4 Interpretability in High-Dimensional Semantic Spaces**

105 High-dimensional semantic spaces, where text embeddings reside, offer a compact yet expansive  
106 representation of language [5]. Recent advancements in embedding interpretability have demonstrated  
107 that certain dimensions in these spaces may correspond to specific linguistic or semantic features,  
108 such as sentiment or subject matter. However, understanding the contribution of individual sentences  
109 to the overall document embedding requires extending these concepts to more complex structures  
110 beyond single words or phrases. Research in this area has shown that vector operations, such as  
111 adding embeddings, can produce new vectors that represent the semantic meaning of their components  
112 [26]. This property of embedding models can be used in various advanced NLP applications such as  
113 analogy solving to semantic search.

# 114 **3 Effect of sentence-level positioning in embedding output**

115 We explore how the position and size of a sentence in a text influence a document's final embedding  
116 vector. Our methodology adapts the needle-in-a-haystack test [11], traditionally used for generative  
117 models in information retrieval [30], to evaluate embedding models.

## 118 **3.1 Experimental setup**

### 119 **3.1.1 Insertion of Irrelevant Text**

120 We investigate the impact of adding irrelevant or adversarial text ("needle") to a document. After  
121 inserting the needle, we generate a new embedding for the altered text and compare it to the original  
122 using cosine similarity. We vary the needle's length (5%, 10%, 25%, 50%, and 100% of the original  
123 text's token count) and position (beginning, middle, end) across 15 experimental conditions. We use  
124 an extended version of Lorem Ipsum placeholder text [32] that exceeds the length of our longest  
125 datapoint and is structured in paragraph format to achieve a needle with structural similarity to our  
126 data while avoiding a confounding effect on the embedding model.

### 127 3.1.2 Removal of Text

128 In a parallel experiment, we remove portions of text (10%, 25%, 50% of sentences, rounded up) from  
129 different positions (beginning, middle, end) in the document. The resulting text is then embedded,  
130 and its similarity to the original embedding is measured using cosine similarity.

## 131 3.2 Models

132 We test various open and closed-source models to demonstrate the consistency of our results across  
133 multiple popular embedding models.

134 **Closed source models** We test Cohere’s Embed-English-v3.0 [24] and OpenAI’s Text-Embedding-  
135 3-Small [21], which have context lengths of 512 and 8192 tokens, respectively. For texts exceeding  
136 these limits, we truncate from the beginning to fit the models’ context windows. Both models are  
137 accessed via their respective APIs.

138 **Open source models** Our experiments also involve open-source models, specifically BAAI’s BGE-  
139 m3 [3], Nomic AI’s Nomic-Embed-Text-v1.5 [20], and Jina AI’s Jina-Embeddings-v2-Base [12],  
140 selected for their performance on the MTEB leaderboard. These models have a maximum context  
141 length of 8192 tokens and 137M parameters. Similarly to the closed-source models, we front-truncate  
142 if our inputs exceed a model’s context window.

## 143 3.3 Datasets

144 To minimize dataset bias and validate our findings across diverse text types, we selected datasets  
145 representing a range of writing categorizations and lengths:

- 146 1. **PubMed Publications:** We use PubMed publication abstracts [4] to assess the impact of  
147 our ablations on scientific writing. Scientific texts are characterized by their structured  
148 presentation of information and specialized vocabulary. Understanding how embeddings  
149 capture this complexity can provide insights into their utility in academic and research  
150 applications.
- 151 2. **Paul Graham Essay Collection:** We analyze over 200 essays written by Paul Graham [10],  
152 varying from 400 to 70,000 words. Paul Graham’s essays are known for their thoughtful,  
153 reflective style and coherent argument structure, making them ideal for studying how  
154 embeddings handle nuanced and complex idea development over long texts.
- 155 3. **Amazon Reviews:** Drawn from MTEB’s Amazon Polarity dataset [41], this helps us  
156 examine consumer review text. Reviews are direct and opinion-rich, offering a perspective on  
157 how embeddings process everyday language and sentiment, which is crucial for applications  
158 in consumer analytics.
- 159 4. **Argumentative Analysis:** From the BiER benchmark’s Argumentative Analysis (ArguAna)  
160 dataset [35], we explore embeddings of formal persuasive writing. This dataset includes well-  
161 constructed arguments that are ideal for testing how embeddings capture logical structure  
162 and the effectiveness of rhetoric.
- 163 5. **Reddit Posts:** More Informal and diverse writing styles can be found on Reddit [9]. This  
164 dataset introduces grammar, style, and subject matter diversity into our tests, extending our  
165 findings to be more robust and adaptable to a wide range of writing styles.

## 166 3.4 Results and discussion

167 Our results indicate a pronounced drop in similarity when irrelevant text is inserted at the beginning  
168 of documents, with less impact seen when additions occur in the middle or end. Specifically, for the  
169 BGE-m3 model, we see that the addition of a needle that is equal to 5% of the total content (across  
170 the 6 datasets, an average of 1-2 sentences) results in the similarity to reduce to .98, compared to  
171 .995 for a needle placed at the end of the input. This is reflected across all datasets tested on, with the  
172 largest decrease within the Paul Graham Essay Collection of a similarity score of .85.

173 This trend intensifies with larger insertions, where inserting text equivalent to 50% of the document  
174 decreases similarity to 0.87 at the beginning versus 0.97 at the end, a 10.3% decrease. We find this

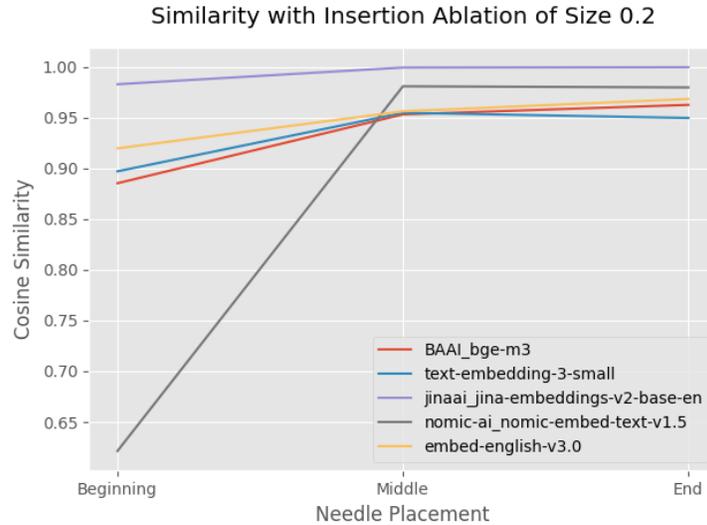


Figure 1: Cosine Similarity vs. Needle Size and Position

175 trend robust to model differences, as all 5 models tested have an average decrease of 7%. Notably,  
 176 even significant alterations where half of the text is irrelevant still retain a minimum similarity of  
 177 0.7, suggesting an unexpected robustness of the embeddings to extensive modifications. We leave  
 178 investigation of this behavior to future work.

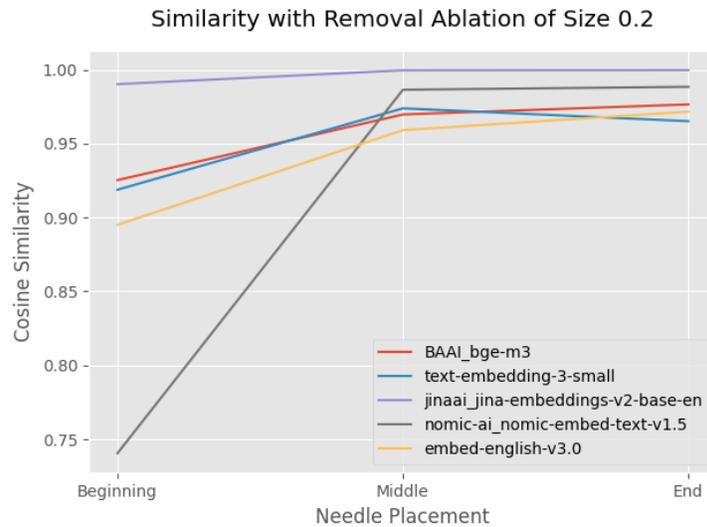


Figure 2: Cosine Similarity vs. Removal Size and Position

179 Similar trends are observed in the removal experiments, with the largest impacts on similarity  
 180 occurring when sentences are removed from the beginning. Removing half the sentences from the  
 181 beginning of the document leads to a median similarity 10.6% lower than removals from the end,  
 182 with no significant difference noted between middle and end removals in contrast to our findings  
 183 during the insertion experiments. Interestingly, even a 50% text removal from the middle maintains a  
 184 median 95% similarity, corroborating our findings during insertion, where we expect to but fail to  
 185 observe a large drop in similarity. The downstream effects of these results are left to future work.

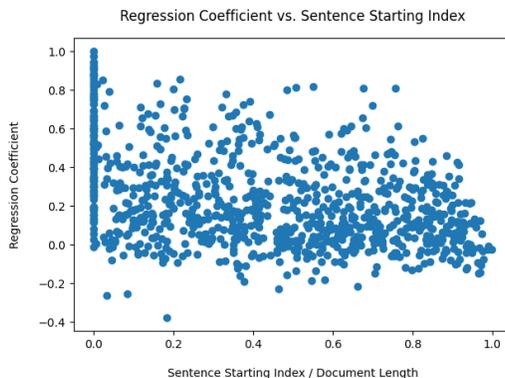


Figure 3: Regression Coefficients vs. Sentence Position

186 **4 Analysis of embedding decomposition**

187 Next, we explore the impact of sentence-level positioning on the final document embedding vector  
 188 through regression analysis, which offers a more direct method to quantify the contribution of  
 189 individual sentences to a document’s embedding representation.

190 **4.1 Reconstructing embedding vectors through linear combinations of constituents**

191 To start, we wanted to validate the assumption that the sentence embeddings of a larger document can  
 192 meaningfully be used as a proxy for the original document embedding [33].

193 To test this, we wanted to determine how much reconstruction loss we would incur from using an  
 194 optimal linear combination of sentence embeddings instead of the document embedding. Optimizing  
 195 for train  $R^2$ , we use Ordinary Least Squares (OLS) regression to reconstruct the document embedding  
 196 from its sentence embeddings, with the document embedding as our response and each sentence as a  
 197 predictive datapoint for our regression. Our model choice is notable for its simplicity and the direct  
 198 interpretability of its coefficients [29], though we acknowledge and check for potential issues posed  
 199 by OLS, such as multicollinearity. Our regressions use normalized embeddings (L2 norm of 1) to  
 200 ensure scale invariance [27].

201 First, we separate our data points into their component sentences by use of punctuation such as  
 202 periods, and new lines. We then use the embeddings of these sentences to run a regression against the  
 203 original document embedding. Using a regression on the sentence embeddings from 3 models across  
 204 all 6 datasets led to  $R^2$  ranging from 0.75 to 1 with an average of 0.876, indicating that approximately  
 205 87.6% of the variance in document embeddings can be accounted for by their component embeddings.  
 206 The MAE across our models and datapoints ranged between 0.001 and 0.01 with an average of  
 207 0.0069, suggesting minimal deviation in the reconstructed vectors.

208 **4.2 Analyzing regression coefficients as importance weights**

209 Given the high explanatory power of our regression models, the coefficients given to each sentence  
 210 (datapoint) in our regression are strong indicators to determine their relative importance to the total  
 211 document. To standardize our comparisons across documents, we standardized each coefficient vector  
 212 by its L2 norm. One potential issue with this approach is the presence of negative values coefficients,  
 213 but these tended to be rare and very low in magnitude when compared to positive coefficients in the  
 214 same vector.

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

218 There is a downward trend in coefficient values with increasing sentence position, suggesting a  
 219 positional bias where earlier sentences generally have a greater impact on the document’s overall

220 semantic representation. To quantify this observation, we plot regression coefficients against sentence  
221 positions over all the documents in our dataset. (Figure 3).

### 222 4.3 Embedding positional bias is robust to human-level writing bias

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

229 More specifically, we expect the weight assigned to the first sentence to follow a uniform weight of  
230  $\frac{1}{\text{num\_sentences}}$ . However, this analysis shows a strong negative correlation ( $r = -0.5$ ) and significant  
231 deviations from the expected uniform distribution ( $\alpha \ll 0.001$ ), confirming a systematic positional  
232 influence within document embeddings. These findings suggest that the embedding models may  
233 inherently prioritize the initial information presented in any text sequence, irrespective of its original  
234 position in the document.

## 235 5 Isolating the Role of Training Methodology in Model Biases

236 Embedding models commonly employ truncation strategies due to their limited context windows,  
237 directly impacting how documents are processed and understood. Prominent models such as OpenAI’s GPT-3 [2]  
238 and Google’s BERT [6] can only process up to 1,024 and 512 tokens, respectively. When documents exceed  
239 these limits, content at the end is often discarded, inherently prioritizing the beginning. As shown in the  
240 previous experiments, this systematic truncation is not merely a technical necessity but a fundamental design  
241 choice that influences model behavior, as the initial sections of documents—typically containing abstracts  
242 or executive summaries—are disproportionately represented.

The relative importance on embedding impact for a given point within a model context length can be  
mathematically described as

$$\text{imp}(t_i) = u(t_i) - \beta(t_i)$$

244 where  $t_i$  represents the number of non-padding tokens encountered at position  $i$ ,  $\text{imp}$  represents  
245 importance or the relative impact of position  $i$  on embedding output,  $u$  represents the total number of  
246 effective updates at  $t_i$ , and  $\beta$  represents the total number of decay opportunities.  $t_i \in [0, N]$  where  $N$   
247 is the number of training examples. An effective update is defined as a single model update based on  
248 a non-padding token in that position, whereas a decay opportunities is a model update being either  
249 empty or having a padding token in that position.

250 Following traditional truncation methods, positions earlier in the context window will be used more  
251 often than those at the end. We can model this as a monotonically decreasing function as the  
252 number of effective updates decrease as  $i$  increases. Due to this implicit bias, the relative importance  
253 ( $\text{imp}(t_1) \geq \text{imp}(t_2) \geq \dots \geq \text{imp}(t_N)$ ) of earlier positions on embedding output will always be  
254 greater than or equal to the positions later in context.

255 Although this monotonic impact on position can theoretically be removed by maintaining an equal  
256 number of effective updates throughout the context, it is unknown what the impacts on computational  
257 costs, and model performance would be. Completing pre-training with this bias in mind will require  
258 considerable research to full understand the impacts, leading us to believe that this bias will continue  
259 in future models.

### 260 5.1 Is it possible to remove positional bias in post-training?

261 Following our theory on bias learned through the pre-training process, we experiment with smaller,  
262 cost-effective fine-tuning methods to remove this bias [33]. We do this by fine-tune models to use  
263 data without the front-truncation, yet still holds similar semantic meaning to the initial data points.

264 We propose a new framework, Position-Aware Data Sampling (PADS), where subsets of data points  
265 are randomly sampled based on input position, to solve this positional bias. The method augments the

266 data by inputting training points that would normally be truncated, and randomly selecting subsets of  
 267 each data point based on position away from the beginning of the original input. For example, instead  
 268 of front-truncating 50% the length of a given example, we select uniformly a token position from 0 to  
 269  $n/2$ , where  $n$  is the token length of the data point.

270 In our fine-tuning experiments, we create positive pairs by sampling from each original twice. For  
 271 negative pairs, we sample once from both the original and another random data point in the dataset.  
 272 Using these pairs, we use contrastive loss to fine-tune the model towards our goal. We follow these  
 273 steps for three datasets and using this to fine-tune BAAI’s BGE-small-en-v1.5. The three datasets  
 274 included are the Paul Graham Essay Collection, PubMed Publications, and Amazon Reviews. We  
 275 sample a maximum of 20% from each dataset, selecting 50 examples for the Paul Graham dataset  
 276 and 225 for the other two datasets. Following the procedure above, we select 50% of each original  
 277 datapoint and create a positive and negative pair from each, resulting in an augmented dataset of 1000  
 278 examples. We use cosine similarity within our contrastive loss function, and then use this with the  
 Adam optimizer for three epochs.

Table 1: Average Cosine Similarity between Original and Ablated Inputs

Model	Beginning	Middle	End
Original	0.923	0.979	0.983
Finetuned	0.984	0.993	0.993
Percent Improvement	6.1%	1.4%	1.0%
Original (external datasets)	0.920	0.978	0.982
Finetuned (external datasets)	0.988	0.995	0.995
Percent Improvement	6.8%	1.7%	1.3%

279

280 With this new method, we have been able to effectively remove positional bias and improve similarity  
 281 metrics to levels similar to when ablations are put in positions different from the beginning. The  
 282 new model has been able to reduce bias by 6.9% with insertion needles, and 6.1% averaged between  
 283 insertion and removal ablations. This work suggests that models can learn to fix its early positional  
 284 bias by sampling the subset position of the input it is training on.

## 285 6 Future work

286 Future work incorporating our findings can focus on three distinct directions:

287 **Alternative Evaluation Metrics** Exploring alternative evaluation metrics beyond cosine similarity  
 288 is essential to assess the effectiveness of embedding models. Future research should consider metrics  
 289 such as Word Mover’s Distance (WMD) [14] for capturing semantic similarity, BERTScore [39] for  
 290 evaluating contextual alignment, and NDCG (Normalized Discounted Cumulative Gain) [37] for  
 291 ranking quality in information retrieval tasks. Additionally, task-specific metrics like classification  
 292 F1-score, BLEU [22] for translation quality, and ROUGE [17] for summarization accuracy can  
 293 provide deeper insights into model performance. These specific metrics can offer a more detailed  
 294 understanding of how well embeddings preserve semantics and perform across various downstream  
 295 applications.

296 **Model Architecture and Training Process Innovations** Given our findings, model creators can  
 297 employ innovative training techniques such as sentence shuffling or random truncation of long  
 298 texts during the embedding training process. These methods can help mitigate positional biases and  
 299 enhance model robustness. Since embedding models use contrastive loss [19] rather than classification  
 300 loss like generative models, careful consideration is needed to determine the best way to compare  
 301 these ablations with their original texts. This could involve designing new contrastive learning  
 302 objectives that account for the positional integrity of the input text. Additionally, incorporating  
 303 architectural modifications, such as advanced attention mechanisms or positional encodings [23], can  
 304 further reduce biases and improve the models’ ability to handle long-context inputs. Experimentation  
 305 with these innovations can lead to embedding models that are more resilient to variations in input  
 306 structure, thereby enhancing their performance across a wide range of downstream tasks.

## 307 **Improved Document Chunking and Impact on Downstream Information Retrieval Tasks**

308 Currently, document chunking does not typically take text structure into account. Better chunking  
309 strategies can focus on isolating important sentences in the text from less useful content and using  
310 these as breaking points. For longer paragraphs without a clear partition point, special attention  
311 will need to be given to the paragraph content to determine where to set a breaking point that leads  
312 to the most useful chunk encoding for downstream tasks. Enhanced contextualization techniques,  
313 such as dynamic chunking strategies, can be developed to preserve semantic coherence and context,  
314 improving the overall quality of embeddings.

315 Future work should focus on how these improved chunking techniques impact downstream informa-  
316 tion retrieval tasks. By aligning chunking strategies with the inherent biases observed in our study, we  
317 can create more effective embeddings for tasks such as search engine optimization, recommendation  
318 systems, and document summarization. Studying the biases in embedding models and how they  
319 influence downstream performance on information retrieval benchmarks is crucial. Evaluating various  
320 chunking strategies, including those discussed here, can reveal how different approaches affect the  
321 retrieval accuracy and relevance of results. This integrated approach will provide a deeper under-  
322 standing of how to optimize embeddings for real-world applications, ensuring that the enhancements  
323 in chunking directly contribute to better performance in information retrieval tasks.

## 324 **7 Conclusion**

325 Our study reveals a positional bias in embedding models, where sentences at the beginning of a  
326 document disproportionately influence the embedding output. This finding is consistent across various  
327 models with differing context sizes and diverse datasets, evident in both text insertion and removal  
328 experiments. We further support this finding with regression analysis to further quantify the impact  
329 of biases towards a given embedding. Then, potential theories on the source of this bias are discussed,  
330 mainly that this trend is intrinsic to the models' training methodologies of truncation rather than  
331 dataset peculiarities themselves.

332 Implicit bias within embeddings models hinder performance in many critical applications across  
333 information retrieval. One main avenue of societal impact from this is within document search in  
334 both cultural and business contexts. In document retrieval in sensitive political topics, reducing bias  
335 within these models improve the ability to maintain relevant information towards a topic. However,  
336 negative impacts of this work include the spread of the knowledge of this bias where a bad actor can  
337 use this knowledge to retrieve particular non-optimal results aligned with their adversarial goals.

338 These insights suggest the need for revised training strategies that mitigate positional biases to achieve  
339 balanced semantic representations. Although our initial experiments show an example of reducing  
340 this bias through fine-tuning, more research must be conducted to have robust techniques to remove  
341 the bias at hand.

## 342 **8 Limitations**

343 We have limited our claims to using 6 models with 6 datasets, but this can be extended to look  
344 at positional bias for more models and datasets to eliminate implicit bias from the experimental  
345 design. The fine-tuning method can be adopted to pre-training method to look at the full effects and  
346 performance impacts, outside the post-training context.

## 347 **References**

- 348 [1] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek. Seven failure points  
349 when engineering a retrieval augmented generation system, 2024.
- 350 [2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan,  
351 P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child,  
352 A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray,  
353 B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei.  
354 Language models are few-shot learners, 2020.

- 355 [3] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. Bge m3-embedding: Multi-lingual,  
356 multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.
- 357 [4] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian. A discourse-  
358 aware attention model for abstractive summarization of long documents, 2018.
- 359 [5] G. Dar, M. Geva, A. Gupta, and J. Berant. Analyzing transformers in embedding space, 2023.
- 360 [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional  
361 transformers for language understanding, 2019.
- 362 [7] W. Fei, X. Niu, P. Zhou, L. Hou, B. Bai, L. Deng, and W. Han. Extending context window of  
363 large language models via semantic compression, 2023.
- 364 [8] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang.  
365 Retrieval-augmented generation for large language models: A survey, 2024.
- 366 [9] G. Geigle, N. Reimers, A. Rücklé, and I. Gurevych. Tweac: Transformer with extendable qa  
367 agent classifiers, 2021.
- 368 [10] S. Goel. paul graham essays (revision 0c7155a), 2024. URL [https://huggingface.co/  
369 datasets/sgoel19/paul\\_graham\\_essays](https://huggingface.co/datasets/sgoel19/paul_graham_essays).
- 370 [11] N. M. Guerreiro, E. Voita, and A. F. T. Martins. Looking for a needle in a haystack: A  
371 comprehensive study of hallucinations in neural machine translation, 2023.
- 372 [12] M. Günther, J. Ong, I. Mohr, A. Abdessalem, T. Abel, M. K. Akram, S. Guzman, G. Mastrapas,  
373 S. Sturua, B. Wang, M. Werk, N. Wang, and H. Xiao. Jina embeddings 2: 8192-token general-  
374 purpose text embeddings for long documents, 2024.
- 375 [13] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus, 2017.
- 376 [14] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document  
377 distances. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference  
378 on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–  
379 966, Lille, France, 07–09 Jul 2015. PMLR. URL [https://proceedings.mlr.press/v37/  
380 kusnerb15.html](https://proceedings.mlr.press/v37/kusnerb15.html).
- 381 [15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau  
382 Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-  
383 intensive nlp tasks, 2021.
- 384 [16] X. Li and J. Li. Angle-optimized text embeddings, 2024.
- 385 [17] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the  
386 Workshop on Text Summarization Branches Out (WAS 2004)*, pages 74–81, July 2004.
- 387 [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and  
388 V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- 389 [19] A. Mnih and Y. W. Teh. A fast and simple algorithm for training neural probabilistic language  
390 models, 2012.
- 391 [20] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar. Nomic embed: Training a repro-  
392 ducible long context text embedder, 2024.
- 393 [21] OpenAI. New embedding models and api updates, jan 2024. URL [https://openai.com/  
394 index/new-embedding-models-and-api-updates/](https://openai.com/index/new-embedding-models-and-api-updates/). Accessed: 2024-05-18.
- 395 [22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation  
396 of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for  
397 Computational Linguistics (ACL-2002)*, pages 311–318, 2002. URL [http://www.aclweb.  
398 org/anthology/P02-1040.pdf](http://www.aclweb.org/anthology/P02-1040.pdf).

- 399 [23] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables  
400 input length extrapolation, 2022.
- 401 [24] N. Reimers. Introducing embed v3, nov 2023. URL [https://cohere.com/blog/  
402 introducing-embed-v3](https://cohere.com/blog/introducing-embed-v3). Accessed: 2024-05-18.
- 403 [25] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks,  
404 2019.
- 405 [26] L. K. Senel, I. Utlu, V. Yucesoy, A. Koc, and T. Cukur. Semantic structure and interpretability  
406 of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*,  
407 26(10):1769–1779, Oct. 2018. ISSN 2329-9304. doi: 10.1109/taslp.2018.2837384. URL  
408 <http://dx.doi.org/10.1109/TASLP.2018.2837384>.
- 409 [27] H. Steck, C. Ekanadham, and N. Kallus. Is cosine-similarity of embeddings really about  
410 similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, WWW ’24. ACM,  
411 May 2024. doi: 10.1145/3589335.3651526. URL [http://dx.doi.org/10.1145/3589335.  
412 3651526](http://dx.doi.org/10.1145/3589335.3651526).
- 413 [28] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu. Roformer: Enhanced transformer with  
414 rotary position embedding, 2023.
- 415 [29] T. Słoczyński. Interpreting ols estimands when treatment effects are heterogeneous: Smaller  
416 groups get larger weights, 2020.
- 417 [30] G. Team, M. Reid, N. Savinov, D. Teplyashin, Dmitry, Lepikhin, T. Lillicrap, J. baptiste Alayrac,  
418 R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. Dai,  
419 K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy,  
420 J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins,  
421 E. Rutherford, E. Moreira, K. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski,  
422 Z. Abbas, N. Schucher, A. Anand, R. Ives, J. Keeling, K. Lenc, S. Haykal, S. Shakeri, P. Shyam,  
423 A. Chowdhery, R. Ring, S. Spencer, E. Sezener, L. Vilnis, O. Chang, N. Morioka, G. Tucker,  
424 C. Zheng, O. Woodman, N. Attaluri, T. Kocisky, E. Eltyshev, X. Chen, T. Chung, V. Selo,  
425 S. Brahma, P. Georgiev, A. Slone, Z. Zhu, J. Lottes, S. Qiao, B. Caine, S. Riedel, A. Tomala,  
426 M. Chadwick, J. Love, P. Choy, S. Mittal, N. Houlsby, Y. Tang, M. Lamm, L. Bai, Q. Zhang,  
427 L. He, Y. Cheng, P. Humphreys, Y. Li, S. Brin, A. Cassirer, Y. Miao, L. Zilka, T. Tobin, K. Xu,  
428 L. Prolea, D. Sohn, A. Magni, L. A. Hendricks, I. Gao, S. Ontanon, O. Bunyan, N. Byrd,  
429 A. Sharma, B. Zhang, M. Pinto, R. Sinha, H. Mehta, D. Jia, S. Caelles, A. Webson, A. Morris,  
430 B. Roelofs, Y. Ding, R. Strudel, X. Xiong, M. Ritter, M. Dehghani, R. Chaabouni, A. Karmakar,  
431 G. Lai, F. Mentzer, B. Xu, Y. Li, Y. Zhang, T. L. Paine, A. Goldin, B. Neyshabur, K. Baumli,  
432 A. Levskaya, M. Laskin, W. Jia, J. W. Rae, K. Xiao, A. He, S. Giordano, L. Yagati, J.-B.  
433 Lespiau, P. Natsev, S. Ganapathy, F. Liu, D. Martins, N. Chen, Y. Xu, M. Barnes, R. May,  
434 A. Vezer, J. Oh, K. Franko, S. Bridgers, R. Zhao, B. Wu, B. Mustafa, S. Sechrist, E. Parisotto,  
435 T. S. Pillai, C. Larkin, C. Gu, C. Sorokin, M. Krikun, A. Guseynov, J. Landon, R. Datta,  
436 A. Pritzel, P. Thacker, F. Yang, K. Hui, A. Hauth, C.-K. Yeh, D. Barker, J. Mao-Jones, S. Austin,  
437 H. Sheahan, P. Schuh, J. Svensson, R. Jain, V. Ramasesh, A. Briukhov, D.-W. Chung, T. von  
438 Glehn, C. Butterfield, P. Jhakra, M. Wiethoff, J. Frye, J. Grimstad, B. Changpinyo, C. L. Lan,  
439 A. Bortsova, Y. Wu, P. Voigtlaender, T. Sainath, S. Gu, C. Smith, W. Hawkins, K. Cao, J. Besley,  
440 S. Srinivasan, M. Omernick, C. Gaffney, G. Surita, R. Burnell, B. Damoc, J. Ahn, A. Brock,  
441 M. Pajarskas, A. Petrushkina, S. Noury, L. Blanco, K. Swersky, A. Ahuja, T. Avrahami, V. Misra,  
442 R. de Liedekerke, M. Iinuma, A. Polozov, S. York, G. van den Driessche, P. Michel, J. Chiu,  
443 R. Blevins, Z. Gleicher, A. Recasens, A. Rrustemi, E. Gribovskaya, A. Roy, W. Gworek,  
444 S. M. R. Arnold, L. Lee, J. Lee-Thorp, M. Maggioni, E. Piqueras, K. Badola, S. Vikram,  
445 L. Gonzalez, A. Baddepudi, E. Senter, J. Devlin, J. Qin, M. Azzam, M. Trebacz, M. Polacek,  
446 K. Krishnakumar, S. yiin Chang, M. Tung, I. Penchev, R. Joshi, K. Olszewska, C. Muir,  
447 M. Wirth, A. J. Hartman, J. Newlan, S. Kashem, V. Bolina, E. Dabir, J. van Amersfoort,  
448 Z. Ahmed, J. Cobon-Kerr, A. Kamath, A. M. Hrafinkelsson, L. Hou, I. Mackinnon, A. Frechette,  
449 E. Noland, X. Si, E. Taropa, D. Li, P. Crone, A. Gulati, S. Cevey, J. Adler, A. Ma, D. Silver,  
450 S. Tokumine, R. Powell, S. Lee, K. Vodrahalli, S. Hassan, D. Mincu, A. Yang, N. Levine,  
451 J. Brennan, M. Wang, S. Hodkinson, J. Zhao, J. Lipschultz, A. Pope, M. B. Chang, C. Li,  
452 L. E. Shafey, M. Paganini, S. Douglas, B. Bohnet, F. Pardo, S. Odoom, M. Rosca, C. N.

453 dos Santos, K. Soparkar, A. Guez, T. Hudson, S. Hansen, C. Asawaroengchai, R. Addanki,  
454 T. Yu, W. Stokowiec, M. Khan, J. Gilmer, J. Lee, C. G. Bostock, K. Rong, J. Caton, P. Pejman,  
455 F. Pavetic, G. Brown, V. Sharma, M. Lučić, R. Samuel, J. Djolonga, A. Mandhane, L. L. Sjöstrand,  
456 E. Buchatskaya, E. White, N. Clay, J. Jiang, H. Lim, R. Hemsley, Z. Cankara, J. Labanowski,  
457 N. D. Cao, D. Steiner, S. H. Hashemi, J. Austin, A. Gergely, T. Blyth, J. Stanton, K. Shivakumar,  
458 A. Siddhant, A. Andreassen, C. Araya, N. Sethi, R. Shivanna, S. Hand, A. Bapna, A. Khodaei,  
459 A. Miech, G. Tanzer, A. Swing, S. Thakoor, L. Aroyo, Z. Pan, Z. Nado, J. Sygnowski, S. Winkler,  
460 D. Yu, M. Saleh, L. Maggiore, Y. Bansal, X. Garcia, M. Kazemi, P. Patil, I. Dasgupta, I. Barr,  
461 M. Giang, T. Kagohara, I. Danihelka, A. Marathe, V. Feinberg, M. Elhawaty, N. Ghelani,  
462 D. Horgan, H. Miller, L. Walker, R. Tanburn, M. Tariq, D. Shrivastava, F. Xia, Q. Wang, C.-C.  
463 Chiu, Z. Ashwood, K. Baatarsukh, S. Samangoeei, R. L. Kaufman, F. Alcober, A. Stjerngren,  
464 P. Komarek, K. Tsihlias, A. Boral, R. Comanescu, J. Chen, R. Liu, C. Welty, D. Bloxwich,  
465 C. Chen, Y. Sun, F. Feng, M. Mauger, X. Dotiwalla, V. Hellendoorn, M. Sharman, I. Zheng,  
466 K. Haridasan, G. Barth-Maron, C. Swanson, D. Rogozińska, A. Andreev, P. K. Rubenstein,  
467 R. Sang, D. Hurt, G. Elsayed, R. Wang, D. Lacey, A. Ilić, Y. Zhao, A. Iwanicki, A. Lince,  
468 A. Chen, C. Lyu, C. Lebsack, J. Griffith, M. Gaba, P. Sandhu, P. Chen, A. Koop, R. Rajwar,  
469 S. H. Yeganeh, S. Chang, R. Zhu, S. Radpour, E. Davoodi, V. I. Lei, Y. Xu, D. Toyama,  
470 C. Segal, M. Wicke, H. Lin, A. Bulanova, A. P. Badia, N. Rakićević, P. Sprechmann, A. Filos,  
471 S. Hou, V. Campos, N. Kassner, D. Sachan, M. Fortunato, C. Iwuanyanwu, V. Nikolaev,  
472 B. Lakshminarayanan, S. Jazayeri, M. Varadarajan, C. Tekur, D. Fritz, M. Khalman, D. Reitter,  
473 K. Dasgupta, S. Sarcar, T. Ornduff, J. Snaider, F. Huot, J. Jia, R. Kemp, N. Trdin, A. Vijayakumar,  
474 L. Kim, C. Angermueller, L. Lao, T. Liu, H. Zhang, D. Engel, S. Greene, A. White, J. Austin,  
475 L. Taylor, S. Ashraf, D. Liu, M. Georgaki, I. Cai, Y. Kulizhskaya, S. Goenka, B. Saeta, Y. Xu,  
476 C. Frank, D. de Cesare, B. Robenek, H. Richardson, M. Alnahlawi, C. Yew, P. Ponnappalli,  
477 M. Tagliasacchi, A. Korchemniy, Y. Kim, D. Li, B. Rosgen, K. Levin, J. Wiesner, P. Banzal,  
478 P. Srinivasan, H. Yu, Çağlar Ünlü, D. Reid, Z. Tung, D. Finchelstein, R. Kumar, A. Elisseff,  
479 J. Huang, M. Zhang, R. Aguilar, M. Giménez, J. Xia, O. Dousse, W. Gierke, D. Yates, K. Jalan,  
480 L. Li, E. Latorre-Chimoto, D. D. Nguyen, K. Durden, P. Kallakuri, Y. Liu, M. Johnson, T. Tsai,  
481 A. Talbert, J. Liu, A. Neitz, C. Elkind, M. Selvi, M. Jasarevic, L. B. Soares, A. Cui, P. Wang,  
482 A. W. Wang, X. Ye, K. Kallarakal, L. Loher, H. Lam, J. Broder, D. Holtmann-Rice, N. Martin,  
483 B. Ramadhana, M. Shukla, S. Basu, A. Mohan, N. Fernando, N. Fiedel, K. Paterson, H. Li,  
484 A. Garg, J. Park, D. Choi, D. Wu, S. Singh, Z. Zhang, A. Globerson, L. Yu, J. Carpenter,  
485 F. de Chaumont Quitry, C. Radebaugh, C.-C. Lin, A. Tudor, P. Shroff, D. Garmon, D. Du,  
486 N. Vats, H. Lu, S. Iqbal, A. Yakubovich, N. Tripuraneni, J. Manyika, H. Qureshi, N. Hua,  
487 C. Ngani, M. A. Raad, H. Forbes, J. Stanway, M. Sundararajan, V. Ungureanu, C. Bishop,  
488 Y. Li, B. Venkatraman, B. Li, C. Thornton, S. Scellato, N. Gupta, Y. Wang, I. Tenney, X. Wu,  
489 A. Shenoy, G. Carvajal, D. G. Wright, B. Bariach, Z. Xiao, P. Hawkins, S. Dalmia, C. Farabet,  
490 P. Valenzuela, Q. Yuan, A. Agarwal, M. Chen, W. Kim, B. Hulse, N. Dukkupati, A. Paszke,  
491 A. Bolt, K. Choo, J. Beattie, J. Prendki, H. Vashisht, R. Santamaria-Fernandez, L. C. Cobo,  
492 J. Wilkiewicz, D. Madras, A. Elqursh, G. Uy, K. Ramirez, M. Harvey, T. Liechty, H. Zen,  
493 J. Seibert, C. H. Hu, A. Khorlin, M. Le, A. Aharoni, M. Li, L. Wang, S. Kumar, N. Casagrande,  
494 J. Hoover, D. E. Badawy, D. Soergel, D. Vnukov, M. Miecnikowski, J. Sims, P. Kumar,  
495 T. Sellam, D. Vlasic, S. Daruki, N. Shabat, J. Zhang, G. Su, J. Zhang, J. Liu, Y. Sun, E. Palmer,  
496 A. Ghaffarkhah, X. Xiong, V. Cotruta, M. Fink, L. Dixon, A. Sreevatsa, A. Goedeckemeyer,  
497 A. Dimitriev, M. Jafari, R. Crocker, N. FitzGerald, A. Kumar, S. Ghemawat, I. Philips, F. Liu,  
498 Y. Liang, R. Sterneck, A. Repina, M. Wu, L. Knight, M. Georgiev, H. Lee, H. Askham,  
499 A. Chakladar, A. Louis, C. Crous, H. Cate, D. Petrova, M. Quinn, D. Owusu-Afriyie, A. Singhal,  
500 N. Wei, S. Kim, D. Vincent, M. Nasr, C. A. Choquette-Choo, R. Tojo, S. Lu, D. de Las Casas,  
501 Y. Cheng, T. Bolukbasi, K. Lee, S. Fatehi, R. Ananthanarayanan, M. Patel, C. Kaed, J. Li,  
502 S. R. Belle, Z. Chen, J. Konzelmann, S. Pöder, R. Garg, V. Koverkathu, A. Brown, C. Dyer,  
503 R. Liu, A. Nova, J. Xu, A. Walton, A. Parrish, M. Epstein, S. McCarthy, S. Petrov, D. Hassabis,  
504 K. Kavukcuoglu, J. Dean, and O. Vinyals. Gemini 1.5: Unlocking multimodal understanding  
505 across millions of tokens of context, 2024.

506 [31] G. Tennenholtz, Y. Chow, C.-W. Hsu, J. Jeong, L. Shani, A. Tulepbergenov, D. Ramachandran,  
507 M. Mladenov, and C. Boutilier. Demystifying embedding spaces using large language models,  
508 2024.

509 [32] R. C. Timmer, D. Liebowitz, S. Nepal, and S. Kanhere. Tsm: Measuring the enticement of  
510 honeyfiles with natural language processing, 2022.

511 [33] H. Tsukagoshi, R. Sasano, and K. Takeda. Comparison and combination of sentence embeddings  
512 derived from different supervision signals, 2022.

513 [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and  
514 I. Polosukhin. Attention is all you need, 2023.

515 [35] H. Wachsmuth, S. Syed, and B. Stein. Retrieval of the best counterargument without prior  
516 topic knowledge. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual  
517 Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages  
518 241–251, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:  
519 10.18653/v1/P18-1023. URL <https://aclanthology.org/P18-1023>.

520 [36] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Improving text embeddings  
521 with large language models, 2024.

522 [37] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, and W. Chen. A theoretical analysis of ndcg type  
523 ranking measures, 2013.

524 [38] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff. C-pack: Packaged resources to advance general  
525 chinese embedding, 2023.

526 [39] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text  
527 generation with bert, 2020.

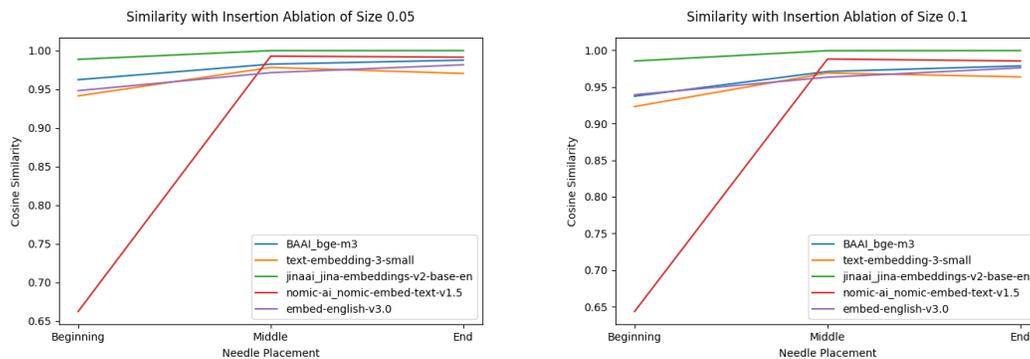
528 [40] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez. Raft: Adapting  
529 language model to domain specific rag, 2024.

530 [41] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification,  
531 2016.

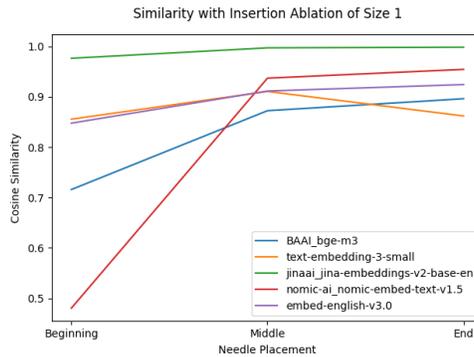
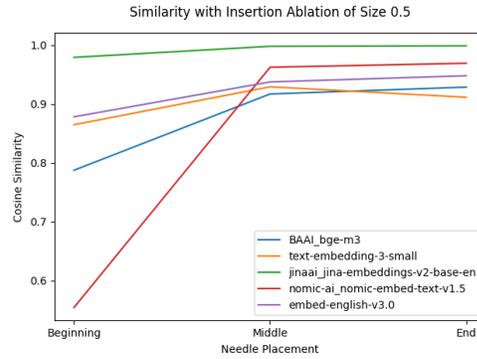
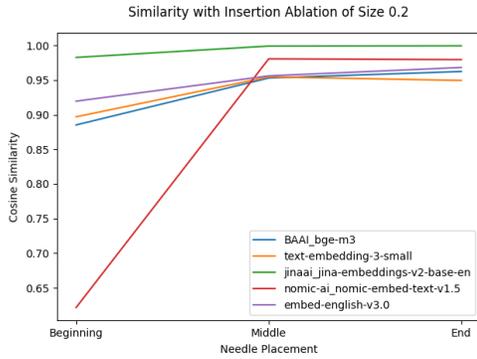
532 [42] D. Zhu, L. Wang, N. Yang, Y. Song, W. Wu, F. Wei, and S. Li. Longembed: Extending  
533 embedding models for long context retrieval, 2024.

## 534 A Cosine similarities across insertion ablation sizes and datasets

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



## 537 B Cosine similarities across deletion of ablation sizes and datasets



538 **NeurIPS Paper Checklist**

539 **1. Claims**

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

542 Answer: **[Yes]**

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

545 Guidelines:

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

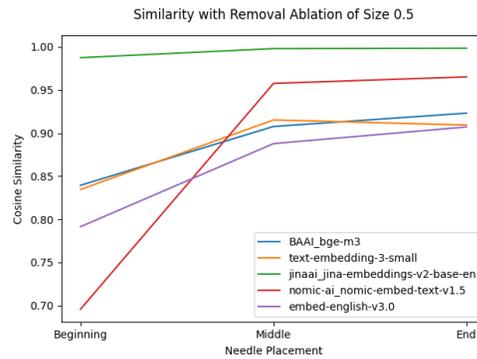
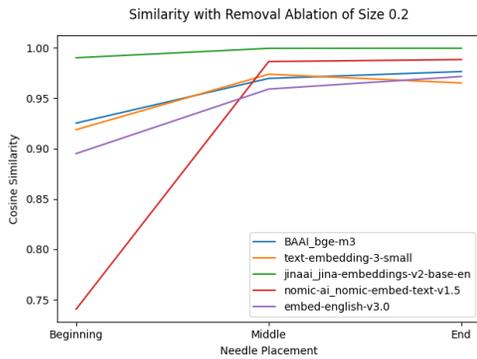
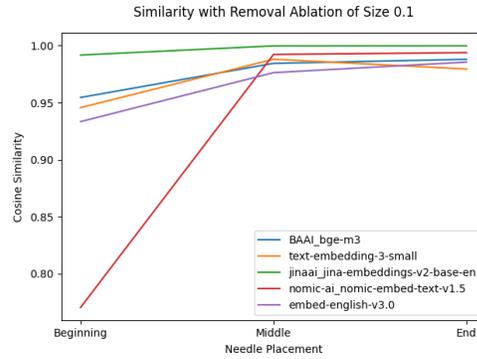
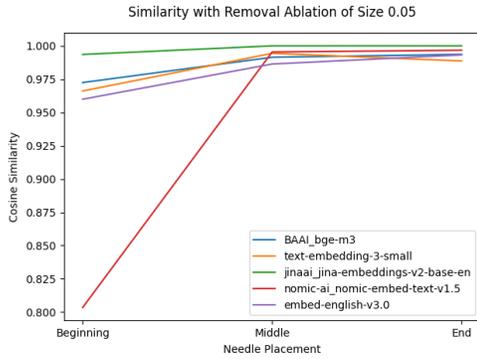
555 **2. Limitations**

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

557 Answer: **[Yes]**

558 Justification: We have a separate limitations section describing the limitation of every major  
559 finding within our paper.

560 Guidelines:



561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587

- 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.
- 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

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

590 Answer: [NA]

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

593 Guidelines:

- 594 • The answer NA means that the paper does not include theoretical results.
- 595 • All the theorems, formulas, and proofs in the paper should be numbered and cross-  
596 referenced.
- 597 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 598 • The proofs can either appear in the main paper or the supplemental material, but if  
599 they appear in the supplemental material, the authors are encouraged to provide a short  
600 proof sketch to provide intuition.
- 601 • Inversely, any informal proof provided in the core of the paper should be complemented  
602 by formal proofs provided in appendix or supplemental material.
- 603 • Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 604 4. Experimental Result Reproducibility

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

608 Answer: [Yes]

609 Justification: Yes, we posted supplementary code to reproduce our work.

610 Guidelines:

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

642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691

## 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, hyper-parameters, 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 error bars and 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.

- 692 • The factors of variability that the error bars are capturing should be clearly stated (for  
693 example, train/test split, initialization, random drawing of some parameter, or overall  
694 run with given experimental conditions).
- 695 • The method for calculating the error bars should be explained (closed form formula,  
696 call to a library function, bootstrap, etc.)
- 697 • The assumptions made should be given (e.g., Normally distributed errors).
- 698 • It should be clear whether the error bar is the standard deviation or the standard error  
699 of the mean.
- 700 • It is OK to report 1-sigma error bars, but one should state it. The authors should  
701 preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis  
702 of Normality of errors is not verified.
- 703 • For asymmetric distributions, the authors should be careful not to show in tables or  
704 figures symmetric error bars that would yield results that are out of range (e.g. negative  
705 error rates).
- 706 • If error bars are reported in tables or plots, The authors should explain in the text how  
707 they were calculated and reference the corresponding figures or tables in the text.

## 708 8. Experiments Compute Resources

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

712 Answer: [Yes]

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

715 Guidelines:

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

## 724 9. Code Of Ethics

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

727 Answer: [Yes]

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

730 Guidelines:

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

## 736 10. Broader Impacts

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

739 Answer: [Yes]

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

741 Guidelines:

- 742 • The answer NA means that there is no societal impact of the work performed.

- 743 • If the authors answer NA or No, they should explain why their work has no societal  
744 impact or why the paper does not address societal impact.
- 745 • Examples of negative societal impacts include potential malicious or unintended uses  
746 (e.g., disinformation, generating fake profiles, surveillance), fairness considerations  
747 (e.g., deployment of technologies that could make decisions that unfairly impact specific  
748 groups), privacy considerations, and security considerations.
- 749 • The conference expects that many papers will be foundational research and not tied  
750 to particular applications, let alone deployments. However, if there is a direct path to  
751 any negative applications, the authors should point it out. For example, it is legitimate  
752 to point out that an improvement in the quality of generative models could be used to  
753 generate deepfakes for disinformation. On the other hand, it is not needed to point out  
754 that a generic algorithm for optimizing neural networks could enable people to train  
755 models that generate Deepfakes faster.
- 756 • The authors should consider possible harms that could arise when the technology is  
757 being used as intended and functioning correctly, harms that could arise when the  
758 technology is being used as intended but gives incorrect results, and harms following  
759 from (intentional or unintentional) misuse of the technology.
- 760 • If there are negative societal impacts, the authors could also discuss possible mitigation  
761 strategies (e.g., gated release of models, providing defenses in addition to attacks,  
762 mechanisms for monitoring misuse, mechanisms to monitor how a system learns from  
763 feedback over time, improving the efficiency and accessibility of ML).

#### 764 11. Safeguards

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

768 Answer: [NA]

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

770 Guidelines:

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

#### 781 12. Licenses for existing assets

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

785 Answer: [Yes]

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

788 Guidelines:

- 789 • The answer NA means that the paper does not use existing assets.
- 790 • The authors should cite the original paper that produced the code package or dataset.
- 791 • The authors should state which version of the asset is used and, if possible, include a  
792 URL.
- 793 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 794 • For scraped data from a particular source (e.g., website), the copyright and terms of  
795 service of that source should be provided.

- 796
- 797
- 798
- 799
- 800
- 801
- 802
- 803
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://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.

804 **13. New Assets**

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

807 Answer: [\[Yes\]](#)

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

809 Guidelines:

- 810
- 811
- 812
- 813
- 814
- 815
- 816
- 817
- 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.

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

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

822 Answer: [\[NA\]](#)

823 Justification: The experiments do not use human subjects.

824 Guidelines:

- 825
- 826
- 827
- 828
- 829
- 830
- 831
- 832
- 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.

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

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

839 Answer: [\[NA\]](#)

840 Justification: The authors do not use human subjects.

841 Guidelines:

- 842
- 843
- 844
- 845
- 846
- 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.

847  
848  
849  
850  
851

- 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.