# Information Representation Fairness in Long-Document Embeddings: The Peculiar Interaction of Positional and Language Bias

**Elias Schuhmacher*** **Andrianos Michail*** **Juri Opitz**
**Rico Sennrich** **Simon Clematide**
Department of Computational Linguistics
University of Zurich
<firstname>.<lastname>@uzh.ch

## Abstract

To be discoverable in an embedding-based search process, each part of a document should be reflected in its embedding representation. To quantify any potential reflection biases, we introduce a permutation-based evaluation framework. With this, we observe that state-of-the-art embedding models exhibit systematic positional and language biases when documents are longer and consist of multiple segments. Specifically, early segments and segments in higher-resource languages like English are over-represented, while later segments and segments in lower-resource languages are marginalized. In our further analysis, we find that the positional bias stems from front-loaded attention distributions in pooling-token embeddings, where early tokens receive more attention. To mitigate this issue, we introduce an inference-time attention calibration method that redistributes attention more evenly across document positions, increasing discoverabiltiy of later segments. Our evaluation framework and attention calibration is available at github.com/impresso/fair-sentence-transformers

## 1 Introduction

Text embedding models serve as the backbone of search engines and retrieval modules in Retrieval Augmented Generation (RAG) and agentic systems. These models map documents and queries into a shared vector space where cosine similarity determines discoverability—any distortion during embedding directly affects which information can be retrieved. Previous work has shown that models exhibit a pronounced *positional bias*—prioritizing information based on where it appears (e.g., Zhu et al., 2024; Zeng et al., 2025; Fayyaz et al., 2025; Ognawala and Cureton-Griffiths, 2025; Liu et al., 2024; Mohtashami and Jaggi, 2023; Lee et al., 2025; Coelho et al., 2024; Modarressi et al., 2025).

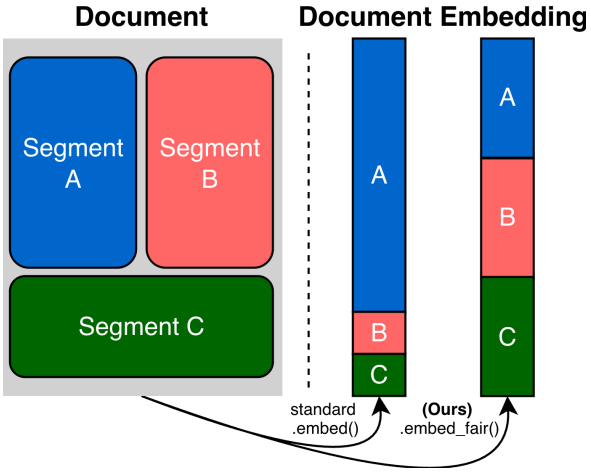



**Figure 1:** We demonstrate that standard document encoders generate skewed embeddings that underrepresent later segments when processing multi-segment documents. We show that an inference-time attention calibration method yields embeddings that are fairer regardless of the relative order of the segment.

Longer documents often place distinct information in different sections—legal documents put definitions in appendices, technical manuals include troubleshooting after introductory content. When embedding models prioritize early content, key information in later positions becomes invisible to semantic search. Consider newspaper pages with multiple independent articles, sometimes in different languages. If the model prioritizes early content, later articles become undiscoverable. Structural conventions and placement determine which information is matched, often over relevance.

Figure 1 illustrates this problem: due to positional bias, the representation of a multi-segment document is dominated by the first segment, making information from later segments invisible.

We formalize this as a threat to *Information Representation Fairness*: when a document composed of multiple segments is embedded, each segment should contribute equally to the resulting vector

*Equal contribution.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 4996–5028
July 2-7, 2026 

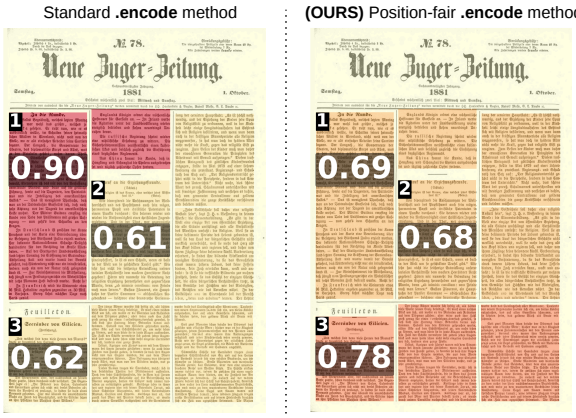



**Figure 2:** A case study that visualizes how positional bias can lead to unfair content representation (left) and shows the effect of applying our proposed mitigation (right). Concretely, the image shows the similarities between the page embedding and the standalone embeddings of three articles. Standard encoding (left) over-represents the first article. Attention calibration (right) distributes similarity more evenly across articles.

representation, regardless of its position or language. Our analysis focuses on a general search scenario in which a user searches for document-level information using one of its segments. In this scenario, a fair document embedding should be equally similar to all of its constituent segments, regardless of segment position or source language. Our setting most closely resembles page-level newspaper embeddings, where distinct articles are concatenated in arbitrary order—making discoverability largely determined by whichever article appears first. Figure 2 illustrates this issue on a real document from an historic newspaper, showing the practical impact of our calibration method.

As first steps toward fairer long-document embedding models, our contributions are:

- An evaluation framework that uses a comparable corpus to quantify positional biases in long-context embedding models.
- A mixed-language extension that reveals source language bias as an additional dimension of unfairness.
- An analysis of the attention patterns of embedding models that use the <s>-token as the pooling token, identifying attention-based causes of positional bias.
- An inference-time attention calibration method that mitigates the positional bias and produces fairer embeddings.

By providing this methodology to detect, quantify, and explain prevalent biases, we advance the understanding of long-document embeddings, a pertinent interpretability challenge (Opitz et al., 2025).

## 2 Related Work

### Positional Bias in Long-Context Models

Early evidence of long-context limitations of transformer-based models was provided by Mohtashami and Jaggi (2023), who introduced a passkey retrieval challenge to test decoder language models and showed that these models often fail to retrieve the correct key when it is buried deep in the input. Similarly, Liu et al. (2024) observe a "lost in the middle" phenomenon stating that decoder models struggle to use information placed in the middle of their input context as opposed to information placed at the start or end. Lee et al. (2025) showed that embeddings of encoder models are more strongly influenced by perturbations at the start of a text than by identical perturbations in the middle or at the end. Zeng et al. (2025) empirically demonstrated that, when performing question answering retrieval, embedding models exhibit higher performance degradation when the related content is placed later within the context.

### Attention Calibration in Decoder-Based Language Models

Hsieh et al. (2024) proposed a method to counteract the U-shaped positional bias in decoder-based language models, attributing the "lost in the middle" phenomenon (Liu et al., 2024) to a corresponding U-shaped pattern in decoder self-attention. This method estimates positional bias by performing one inference per position while shifting a fixed dummy segment across all (e.g., 20) positions, thereby obtaining a baseline attention distribution that reflects positional bias. The estimated bias is then subtracted from the observed segment attention to produce calibrated attention, which is used during a "positionally fair" inference.

## 3 Methods

We introduce the notation of *segments* and *documents* used in this work and define the two concepts: *positional fairness* and *information retention*.

### Notation

Let $\Sigma$ be a finite vocabulary (alphabet) of tokens, and let $\Sigma^*$ denote its Kleene closure–the set of all

finite token sequences over $\Sigma$. We define a *segment* $s \in \Sigma^*$: a logically coherent unit of text of arbitrary length. Each segment is treated as an indivisible unit of information. Further, we define a *segment set* $S = \{s_1, \ldots, s_n\}$: an unordered collection of $n$ segments, with $s_i \neq s_h$ for $i \neq h$. Finally, we define a *document* $D \in \Sigma^*$: a concatenation of multiple segments. Given a segment set $S$, we construct $n!$ distinct documents by permuting segment order. We employ this permutation setup to eliminate the segment content as a confounding factor in our analysis. Given a segment set $S$ of size $n \geq 1$ and a specific ordering $j$ out of all $n!$ distinct orderings, document $D_j$ is the concatenation (with whitespace) of the $n$ segments in $S$ following this ordering. We denote the segment at position $i$ within document $D_j$ as $s_i^{D_j}$, and use 1-indexed positions.

We consider *text embedding models* with $\mathscr{L}$ transformer layers, following the functional form: $F_\theta : \Sigma^* \to \mathbb{R}^d$.

We write $\mathbf{e}^{iso}_{s_i^{D_j}} = F_\theta(s_i^{D_j})$ to denote the embedding of segment $i$ from document $D_j$. Note that this is the embedding of the segment *isolated* on its own as standalone text, i.e., without contextual information from other segments in $D_j$. Similarly, $\mathbf{e}_{D_j} = F_\theta(D_j)$ the embedding of a document $D_j$.

For mean-pooled models, we write $\mathbf{e}^{ctx}_{s_i^{D_j}}$ to denote the *contextualized* segment embedding (late chunking; (Günther et al., 2025)) of the segment at position $i$ within document $D_j$:

$$\mathbf{e}^{ctx}_{s_i^{D_j}} = \frac{1}{|P(s_i^{D_j})|} \sum_{p \in P(s_i^{D_j})} \mathbf{h}_p^{(\mathscr{L})},$$

with $\mathbf{h}_p^{(\mathscr{L})}$ the contextual token representation at the final layer $\mathscr{L}$ of embedding model $F_\theta$ for the token at position index $p$, $P(s_i^{D_j})$ the set of token index positions belonging to segment $s_i^{D_j}$ within document $D_j$, and $|P(s_i^{D_j})|$ the number of tokens in segment $s_i^{D_j}$. As segment-wise mean pooling requires meaningful token-level representations, this operation is only applicable for embedding models preserving token-level semantics during fine-tuning, such as mean-pooled models.

### Construction of Segments and Documents

We construct a multilingual comparable corpus based on topic-aligned Wikipedia articles in six languages. We include high-resource (English, Chinese, German, Italian) and lower-resource (Korean, Hindi) languages. To improve cross-lingual semantic alignment, we apply heuristic length-based filtering, retaining articles whose lengths are within ±70% of the English article length, measured in XLM-R subtokens. On average, articles are shorter than English by 31% ($zh$), 25% ($de$), 30% ($it$), 37% ($ko$), and 29% ($hi$). A single sample is shown in Figure 9 in the Appendix.

Each segment $s$ is a Wikipedia article between 1,000 and 2,000 tokens. Each document $D$ consists of $n$ randomly drawn segments, with a maximum document length of 8,192 tokens. We set $n \in \{3, 4, 5, 6\}$. Documents follow a language configuration $\mathbf{L} = (\mathcal{L}_1, \ldots, \mathcal{L}_n)$, where $\mathcal{L}_i \in \{en, zh, de, it, ko, hi\}$ specifies the language at position $i$. We construct **monolingual documents** with $\mathbf{L} = (\mathcal{L}, \ldots, \mathcal{L})$, and **mixed-language documents** with $\mathbf{L} = (\mathcal{L}_{lead}, \mathcal{L}_{later}, \ldots, \mathcal{L}_{later})$ and $\mathcal{L}_{lead} \neq \mathcal{L}_{later}$. Each combination of $(n, \mathbf{L})$ defines an experiment instance. Owing to the factorial number of segment permutations, we construct 1,002 documents for $n$=3, 1,008 for $n$=4, 1,080 for $n$=5, and 10,080 for $n$=6.

### Examined Embedding Models

We analyze the information representation fairness of two state-of-the-art multilingual encoder-based embedding models that employ different pooling strategies to produce document embeddings. **mGTE**[1] represents a document using the **start-of-sequence token** <s> embedding (Zhang et al., 2024). **jina-v3**[2] uses **mean pooling** over contextualized token representations from a LoRA adapter (Sturua et al., 2024). We use the text-matching LoRA adapter.

Both models (i) support long contexts of up to 8,192 tokens, (ii) cover all six languages considered in this study, and (iii) perform well on the Multilingual MTEB benchmark for long-context ($\geq 8,192$ tokens) (Enevoldsen et al., 2025).

### Positional Fairness

Let $Q_S$ denote a data-generating process of $n$-sized segment sets $S$, each containing $n$ Wikipedia articles. We write $\mathrm{supp}(Q_S)$ for the set of segment sets that occur with positive probability under $Q_S$. For a fixed $S$, let $J \mid S \sim \mathrm{Unif}\{1, \ldots, n!\}$ be a specific ordering sampled uniformly from all possible $n!$ permutations, and let $D_J$ be the correspond-

[1] Alibaba-NLP/gte-multilingual-base
[2] jinaai/jina-embeddings-v3

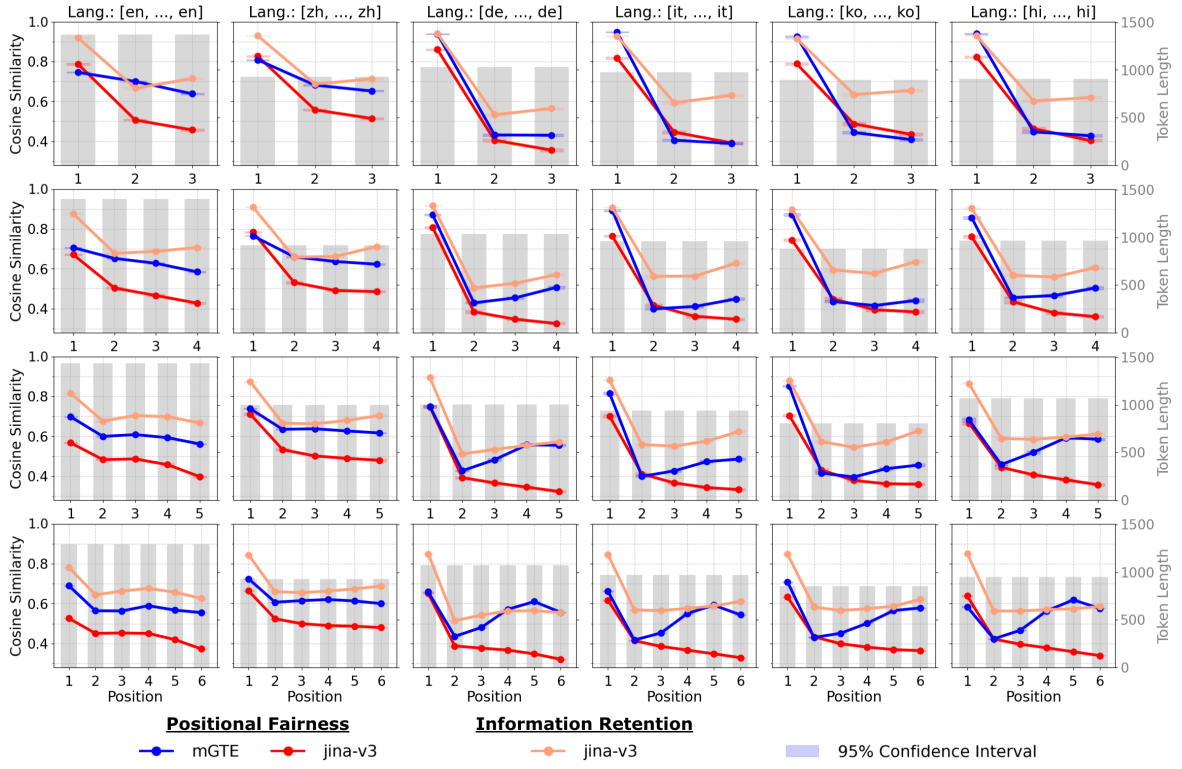



**Figure 3:** Monolingual experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show (i) average representation in the global document embedding (mGTE and jina-v3), and (ii) average information retention (jina-v3) per segment position. (Gray bars) show average token length per segment position.

ing document. Let $\sigma_{i,J} = \cos\big(\mathbf{e}_{D_J}, \mathbf{e}^{iso}_{s_i \in D_J}\big)$ be the cosine similarity between the document embedding and the standalone embedding of the segment placed at position $i$ in $D_J$.

An embedding model is *fair* if $\mathbb{E}[\,\sigma_{1,J} \mid S\,] = \cdots = \mathbb{E}[\,\sigma_{n,J} \mid S\,] \quad \forall\, S \in \mathrm{supp}(Q_S)$, i.e., the document embedding represents the semantic content of all constituent segments equally well *regardless of their position* in the document.

We fit an ordinary least squares regression (OLS; Gelman et al. (2021)) with categorical indicators for positions to estimate deviations from positional fairness. Let $\mathcal{S}$ be the finite collection of segment sets in our dataset. For each $S \in \mathcal{S}$, for each permutation $j \in \{1, \ldots, n!\}$ of that set, and for each position $i \in \{1, \ldots, n\}$, we observe the similarity $\sigma_{i,j}$. We pool all observations across $S \in \mathcal{S}$ and estimate $\sigma_{i,j} = \beta_0 + \sum_{p=2}^{n} \beta_p \, \mathbb{I}\{i = p\} + \varepsilon_{i,j,S}$, where $\beta_p$ captures the difference of position $p$ relative to the baseline (similarity at position 1). Standard errors are clustered at the *segment-set* level, i.e., $\varepsilon_{i,j,S}$ may be arbitrarily correlated within a given $S$ but is assumed uncorrelated across different segment sets. The null hypothesis of *positional fairness* is $H_0^{(p)} : \beta_p = 0 \quad (p = 2, \ldots, n)$.

### Information Retention

Let $\tau_{i,J} = \cos\big(\mathbf{e}^{iso}_{s_i \in D_J}, \mathbf{e}^{ctx}_{s_i \in D_J}\big)$ be the cosine similarity between the standalone embedding and the *contextualized* embedding of the same segment if placed at position $i$ in $D_J$. This measure quantifies how strongly the semantic information of a segment changes when contextualized inside a longer, multi-segment document.

An embedding model exhibits *no position-dependent information retention* if $\mathbb{E}[\,\tau_{1,J} \mid S\,] = \cdots = \mathbb{E}[\,\tau_{n,J} \mid S\,] \quad \forall\, S \in \mathrm{supp}(Q_S)$, i.e., the semantics of any segment is equally well preserved during contextualization *regardless of its position* within a document. We estimate $\tau_{i,j} = \beta_0^{(\tau)} + \sum_{p=2}^{n} \beta_p^{(\tau)} \, \mathbb{I}\{i = p\} + \varepsilon^{(\tau)}_{i,j,S}$, with the null hypothesis $H_0^{(p)} : \beta_p^{(\tau)} = 0 \quad (p = 2, \ldots, n)$.

## 4 Analysis

### 4.1 Monolingual Documents

#### Positional Fairness

Figure 3 depicts the segment-representation profiles of the $n{\times}\mathbf{L}$ (4×6) combinations of monolingual experiment instances (shown in blue (mGTE) and dark red (jina-v3)). We observe a consis-

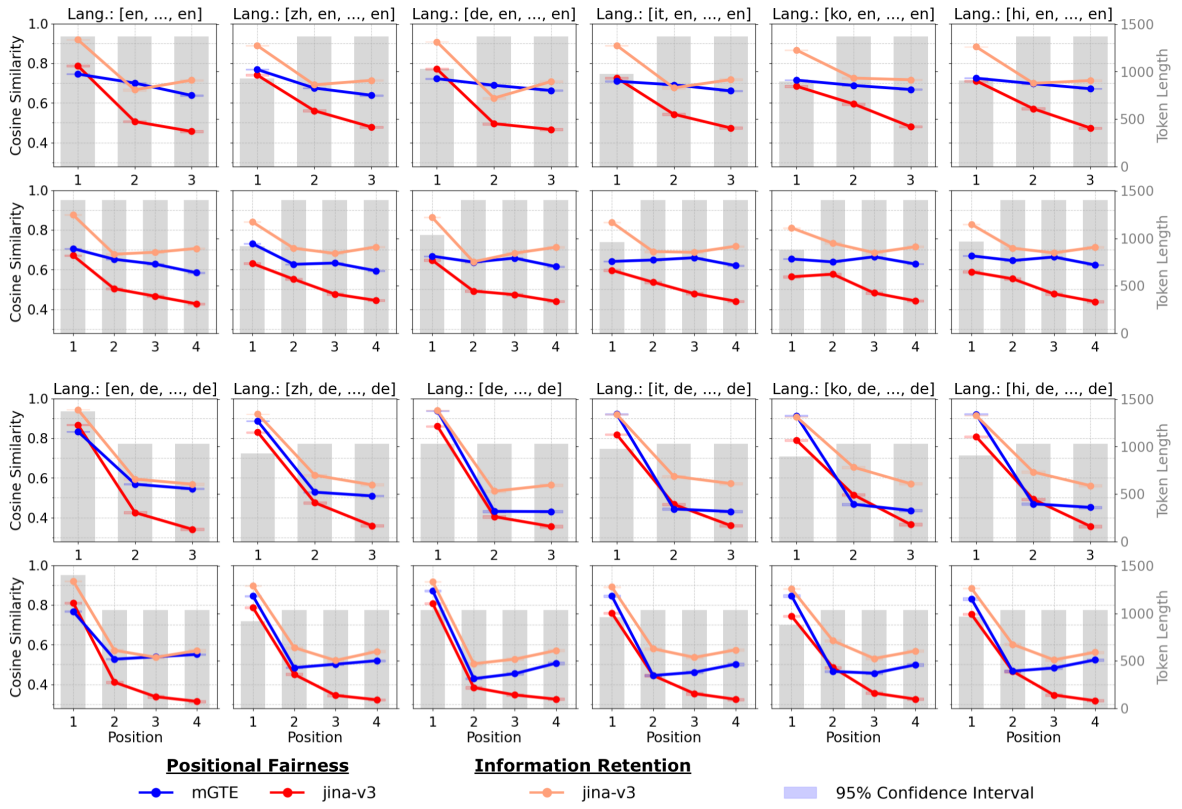



**Figure 4:** Preferential treatment of English segments: later segments in $\mathcal{L}_{later} = en$ (top 2 rows) are better represented in the global document embedding (mGTE and jina-v3) and exhibit higher information retention (jina-v3) than later segments in $\mathcal{L}_{later} = de$ (bottom 2 rows).

tent positional bias: any segment—if placed at the first position of a document—is captured substantially better in the global document embedding than other segments. This results in L-shaped segment-representation profiles. The strength of this positional bias varies by language: the difference in the document representation between the first-positioned segment and any later segment is markedly smaller for English and Chinese.

The OLS estimates strongly reject the null hypotheses $H_0^{(p)} : \beta_p = 0$ for most $p$ in all experiment instances. The largest negative coefficients typically occur at $p$=2 and $p$=3, indicating that the semantic content of second- and third-positioned segments is most underrepresented in the global document representation. Detailed OLS estimates are shown in Tables 1 and 2 in the Appendix. The magnitude of the positional effects attenuates for later positions.

### Information Retention

Figure 3 shows the segment-retention profiles of the 4×6 monolingual experiment instances (shown in light red (jina-v3)). We identify a clear pattern: first-positioned segments retain most of their semantic content, whereas later-positioned segments show progressively greater divergence between their standalone and contextualized information. This results in L-shaped segment-retention profiles. The magnitude of this positional effect varies by language: English and Chinese segments show greater resilience to semantic distortion. We strongly reject the null hypotheses $H_0^{(p)} : \beta_p^{(\tau)} = 0$ for all $p$ in all experiment instances. Detailed OLS estimates are shown in Table 11 in the Appendix.

## 4.2 Mixed-Language Documents

### Positional Fairness

For mixed-language documents, positional bias persists but is significantly influenced by language-specific effects. Most notably, English (and, for mGTE, also Chinese) segments are well represented in the global document embedding, regardless of the positional placement in the document. If placed at later positions, this language preference counteracts positional bias. Figure 4 illustrates this preferential treatment of English segments compared to German segments: experiment instances with $\mathcal{L}_{later} = en$ (top 2 rows) exhibit

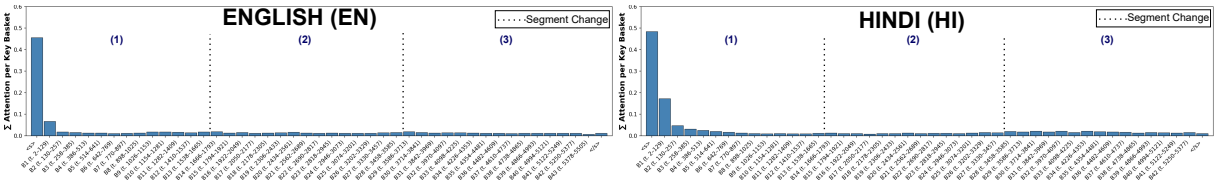



**Figure 5:** Front-loaded self-attention distribution of the <s>-query token over key baskets (basket size $\mathfrak{B}$=128) in English (left) and Hindi (right) documents ($n$=3). Average of the last six transformer layers.

flatter segment-representation profiles (shown in blue (mGTE) and dark red (jina-v3)) compared to $\mathcal{L}_{later} = de$ (bottom 2 rows). Detailed results of all language configurations are shown in Figures 10 to 13 in the Appendix. The OLS estimates confirm these results, detailed in Tables 3 to 10 in the Appendix.

### Information Retention

Positional effects remain in mixed-language documents, but similar to positional fairness, they exhibit language-specific effects. Specifically, English segments show greater resilience to interference from surrounding multilingual segments compared to other languages. Figure 4 shows these language-specific effects: experiment instances with $\mathcal{L}_{later} = en$ (top 2 rows) exhibit flatter information-retention profiles (shown in light red (jina-v3)) than the corresponding instances with $\mathcal{L}_{later} = de$ (bottom 2 rows). In the Appendix, detailed results of all language configurations are shown in Figures 10 to 13 and OLS estimates confirming these results are detailed in Tables 12 to 15.

### Analysis Overview

As to positional fairness, independent of the embedding model, source languages, or document lengths, we observe a systematic **first-position bias**: the first segment is represented most strongly in the global document embedding. In mixed-language documents, position effects persist but are counteracted by **language preferences**: segments in English (and, for mGTE, Chinese) are represented well in the global document embedding, regardless of their position in the document.

Similarly, information retention is highest at the **first-positioned segment**. The magnitude of this effect **depends on language**. In monolingual documents, English shows smaller positional effects than German, Italian, Korean, and Hindi. In mixed-language documents, the first-position bias persists, but **English** segments retain much of their original representation, regardless of their position.

### 4.3 Self-Attention Distribution

To examine potential origins of the positional bias, we analyze the self-attention distribution of the mGTE embedding model. We hypothesize a *front-loaded* distribution over long inputs: early tokens receive more attention mass than later tokens, thereby exerting greater influence on intermediate representations and, ultimately, on the global document embedding. As the mGTE model uses <s>-pooling, we examine how the <s>-token query allocates attention mass over the sequence. We aggregate destinations (keys) into fixed-size, contiguous baskets of size 128 and show the <s>- and </s>-token separately to avoid distorting the first and last basket by special tokens. If positional bias in global document representations is indeed driven by front-loaded self-attention distributions, we would expect L-shaped attention profiles.

Figure 5 shows that among content-bearing baskets, the first basket (tokens 2–129) receives the most attention from the <s>-token. This aligns with the overrepresentation of the first-positioned segment in global document embeddings. Furthermore, slight mid/late-sequence increases in the attention profiles align with the U-shaped segment-representation profiles observed previously (cf. English in Figure 14a and Hindi in Figure 14b in the Appendix). For some experiment instances, we observe partial agreement between self-attention profiles and the similarity of each segment's representation to the document representation; hence, we conclude that pooling-token self-attention can help explain positional bias patterns in document embeddings, but does not capture all nuances.

## 5 Inference-Time Attention Calibration

Motivated by the (partial) alignment between attention allocation and positional bias, we propose to re-weight attention scores during the forward pass. With this, we aim to counteract the front-loaded distribution, thereby allowing information from later tokens to be better represented in the

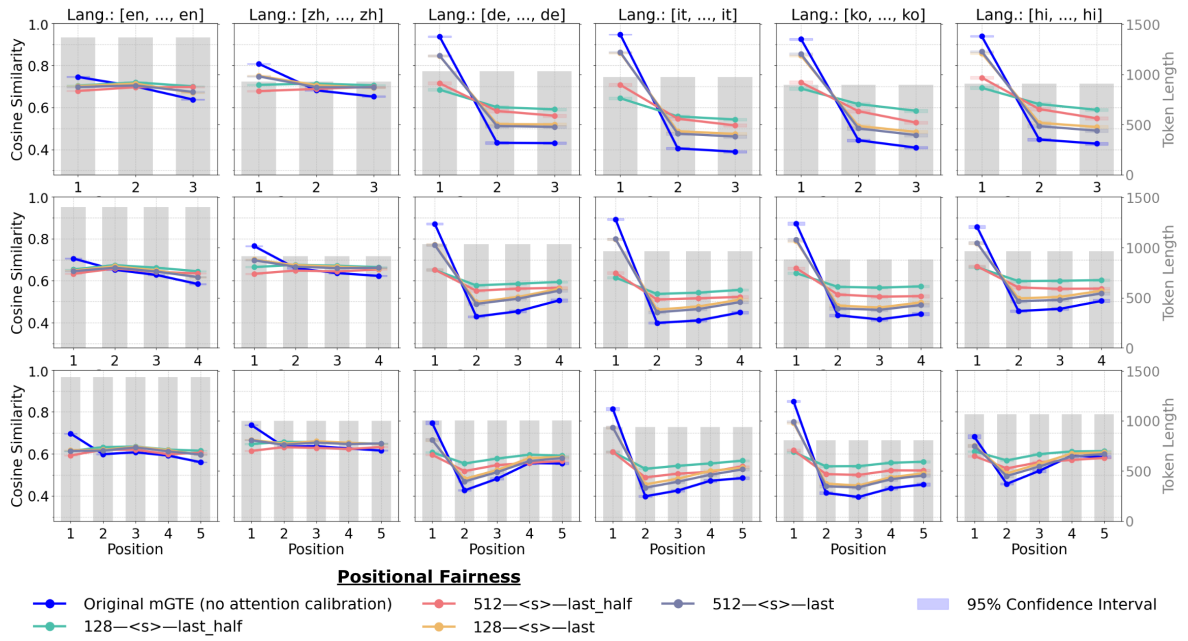



**Figure 6:** Monolingual document experiments using attention calibration on mGTE. Notation: *128—<s>—last_half*: Calibrated embeddings with hyperparameters $\mathfrak{B}$=128, $\mathfrak{L}^C$=$\{7, \ldots, 12\}$

global document embedding. This attention calibration operates entirely at inference time, i.e., it does not require any additional training.

## Approach

Inspired by attention calibration for question answering in decoder models (Hsieh et al., 2024), we propose a different calibration technique for embedding models: inference-time attention mass equalization across positional baskets. Unlike Hsieh et al. (2024), who require one inference per position, our method requires only two forward passes to obtain a calibrated embedding. We partition key positions into fixed-size, contiguous baskets and enforce a uniform attention mass for each basket, while preserving the relative distribution within each basket. This spreads the global attention budget of the <s>-token over the full key sequence—later baskets receive as much total attention as earlier ones—while keeping fine-grained, within-basket patterns unchanged. Figure 7 shows our approach on a conceptual level.

Our approach calibrates the <s>-query row and takes two hyperparameters: (i) **basket size** ($\mathfrak{B}$): partitions the key sequence into $\mathfrak{B}$-sized baskets[3]; (ii) **calibrated layer set** ($\mathfrak{L}^C$): subset of transformer layers $\mathfrak{L}^C \subseteq \{1, \ldots, \mathscr{L}\}$ where calibration is applied independently for each head.

[3]We put the <s>-pooling token in its own basket, yielding a total of $\lceil \frac{L-1}{\mathfrak{B}} \rceil + 1$ baskets for an $L$-sized sequence.

## Results: Monolingual Documents

We apply attention calibration to mGTE and report results for the following hyperparameters: $\mathfrak{B} \in \{128, 512\}$ and $\mathfrak{L}^C \in \{\{7, \ldots, 12\}, \{12\}\}$.

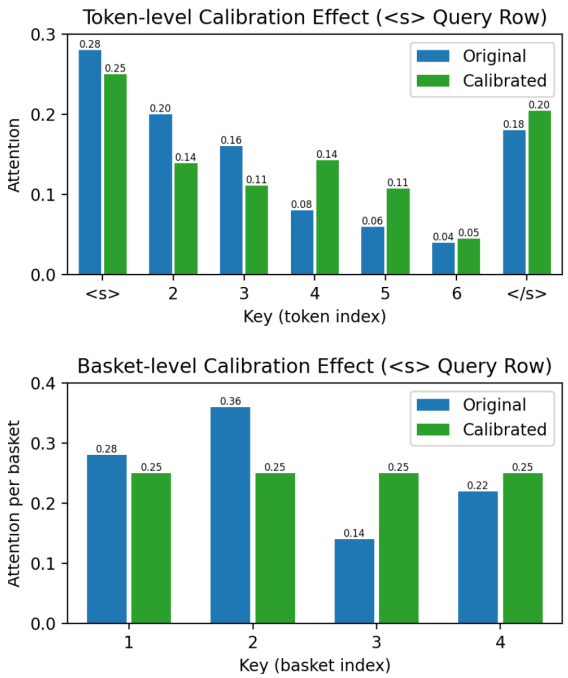



**Figure 7:** Conceptual illustration of the attention calibration. Calibration is applied to the <s>-query token. The <s>-key token is treated individually (basket index 1 contains only the <s>-key token). Exemplarily, the remaining key sequence is grouped into baskets of size 2 (e.g., basket index 2 contains token indices 2 and 3).

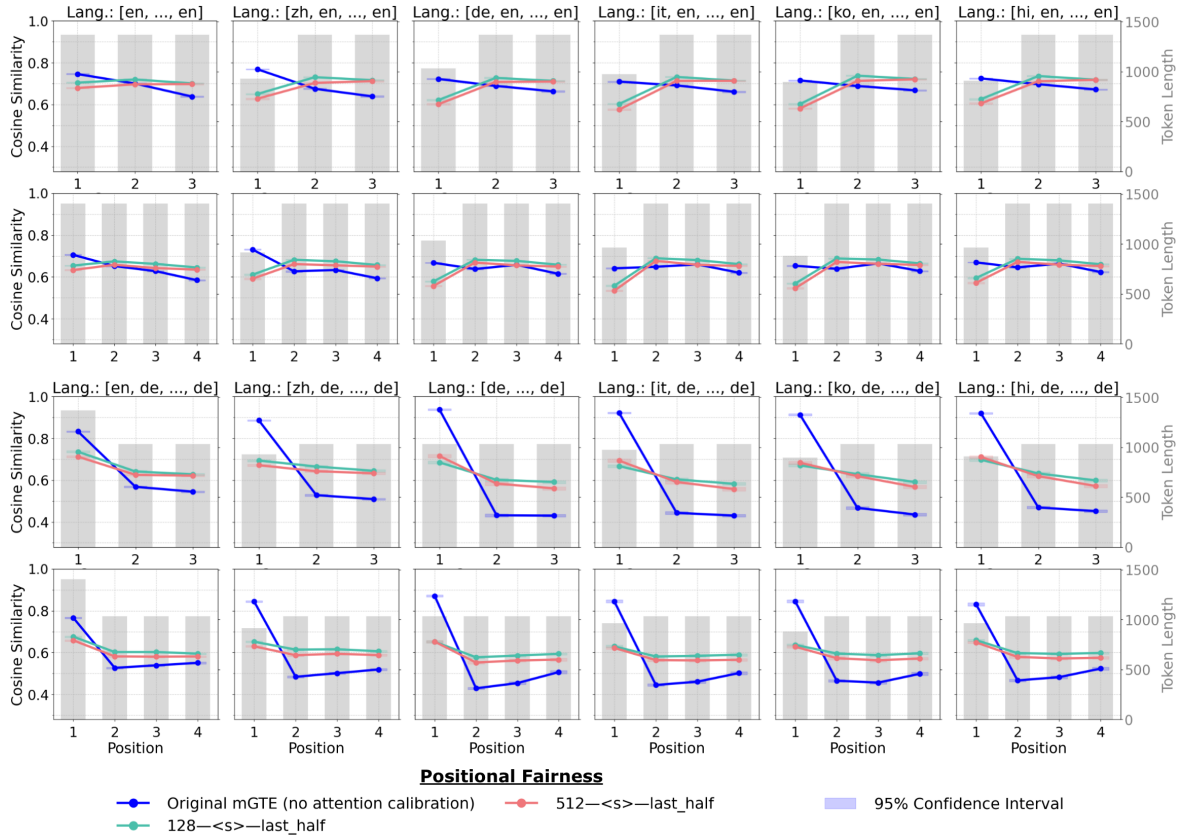



**Figure 8:** Mixed-language document experiments using attention calibration on mGTE. Notation: *128—<s>—last_half*: Calibrated embeddings with hyperparameters $\mathfrak{B}$=128, $\mathfrak{L}^C$=$\{7, \ldots, 12\}$

Figure 6 shows a clear reduction of positional bias in all experiment instances when calibrated attention is used. This is achieved through two complementary adjustments: (i) the representation of the first-positioned segment is decreased relative to the uncalibrated baseline, counteracting its overrepresentation; (ii) the representations of later-positioned segments are increased, showing that their semantic content is better represented in the global document embedding after calibration. The OLS estimates confirm the substantial reduction of positional bias using calibrated embeddings, detailed in Table 16 in the Appendix.

### Results: Mixed-Language Documents

We examine experiment instances with $\mathcal{L}_{later} \in \{en, zh, de\}$, and use the following hyperparameters: $\mathfrak{B} \in \{128, 512\}$ and $\mathfrak{L}^C$=$\{7, \ldots, 12\}$.

Figure 8 shows experiment instances with $\mathcal{L}_{later} = en$ (top 2 rows) and corresponding instances with $\mathcal{L}_{later} = de$ (bottom 2 rows). Only for $\mathcal{L}_{later} = en$ (and, in weaker form, for $\mathcal{L}_{later} = zh$ in Figure 18 in the Appendix), we observe an **inverted L-shape**. We interpret this as further evidence for the **English/Chinese language preferences** of the model: While in the uncalibrated model (blue), the first-position bias partly offsets the language penalty of a non-English segment (e.g., German), the calibrated model reduces this positional uplift, leading to later English/Chinese segments being more strongly represented than the first-positioned segment. Detailed results of all language configurations are shown in Figures 16 to 18 in the Appendix. The OLS estimates confirm these results in Tables 17 to 19 in the Appendix.

Additionally, attention calibration reduces secondary positional effects, such as U-shaped tails. While for uncalibrated embeddings, we occasionally observe U-shaped segment-representation profiles for $\mathcal{L}_{later} = de$, attention calibration erases this curvature ($\beta_2 \approx \beta_3 \approx \beta_4 \approx \beta_5$; cf. Table 18 in the Appendix).

### Evaluating the Loss of Semantics

Semantics in the embedding space do not collapse. In fact, for later-positioned segments, the similarity between the document embedding and the standalone segment embedding is often equal to or higher under calibration than without, indicating that calibration preserves the original semantic con-

tent of the embeddings. To validate this finding, we conducted a control experiment for the two most aggressive calibration settings, calibrating only the document embeddings $\mathbf{e}_{D_j}$ while keeping the standalone segment embeddings $\mathbf{e}^{iso}_{s_i^{D_j}}$ unchanged. As shown in Figure 15 in the Appendix, the unaltered embeddings exhibit near-identical or increased similarity with the calibrated document embeddings compared to the uncalibrated baseline. This indicates that calibration increases the meaningful information of later-positioned segments in the global document representation, as semantic degradation would have produced decreased similarities instead.

## 6 Conclusions

We investigate how position and language affect Information representation fairness when embedded texts are composed of multiple, thematically independent segments (e.g., newspapers). To this end, we provide (i) a diagnostic framework to quantify position–language interactions and positional fairness in long-document embeddings; (ii) clear empirical evidence that state-of-the-art encoder-based embedding models suffer from a first-position bias and model-dependent language preferences which both decrease fairness as to how long documents are embedded and how segments are contextualized; (iii) empirical evidence for front-loaded self-attention distributions that align with the observed first-position bias in global document embeddings; (iv) a training-free attention calibration method that substantially reduces positional biases by distributing representational capacity more evenly across a document. Together, they offer practitioners the tools to diagnose and reduce representational inequalities in page-level applications while offering a foundation for future work towards fairer multilingual long-context representations.

## 7 Limitations

First, our notion of information representation fairness relies on cosine similarity between document embeddings and standalone segment embeddings as a proxy for how strongly a segment is reflected in the global representation. While this operationalization enables a tractable and model-agnostic evaluation, it does not directly measure intrinsic contribution of segments to the embedding. In particular, cosine similarity is sensitive to properties of the embedding space and scaling, and thus to some extent may conflate representational alignment with true information contribution. Consequently, our fairness metric should be interpreted as a relative measure under a fixed probing representation, rather than a definitive characterization of segment-level influence.

Second, our analysis is restricted to *encoder-based* long-context embedding models. While these models are widely used in embedding-based retrieval systems, decoder-based and hybrid architectures may exhibit different positional and language bias patterns. Extending the proposed evaluation framework to other embedding paradigms, including instruction-tuned or generative embedding models, remains an open direction.

Third, our evaluation relies on a multilingual comparable corpus constructed from Wikipedia articles. Although this design allows us to control semantic content across languages and systematically isolate positional effects, Wikipedia articles differ from real-world long documents such as reports, or legal texts in structure, discourse style, and noise. As a result, the absolute magnitude of the observed biases may differ in applied settings, even if the underlying mechanisms persist.

Fourth, we focus on representation-level effects and do not evaluate downstream task performance. Our notion of information representation fairness is defined geometrically—via cosine similarity and discoverability in embedding space—independent of any specific retrieval or ranking pipeline. While this abstraction enables task-agnostic diagnosis, future work should connect these findings to end-task outcomes in realistic retrieval systems. Early indications suggest the method is effective in downstream retrieval scenarios, but thorough evaluation remains for future work.

Fifth, our proposed attention calibration method is evaluated primarily on a single embedding model (mGTE) that uses pooling-token representations. While the method is conceptually model-agnostic and operates entirely at inference time, its effectiveness for other architectures and pooling strategies calls for further empirical validation by future work.

Finally, our study focuses on positional and language biases but does not consider other potential sources of representational inequality, such as topic frequency or domain mismatch. While prior work has examined how differences between naturally written text and LLM-rewritten text in various tones can affect embeddings (Cao, 2025). Understanding

how such biases interact with positional and language effects in long-document embeddings is an important avenue for future research.

## Acknowledgments

The authors received funding through the project *Impresso – Media Monitoring of the Past II Beyond Borders: Connecting Historical Newspapers and Radio*. Impresso is a research project funded by the Swiss National Science Foundation (SNSF 213585) and the Luxembourg National Research Fund (17498891).

## References

Hongliu Cao. 2025. Writing style matters: An examination of bias and fairness in information retrieval systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 336–344, New York, NY, USA. Association for Computing Machinery.

João Coelho, Bruno Martins, Joao Magalhaes, Jamie Callan, and Chenyan Xiong. 2024. Dwell in the Beginning: How Language Models Embed Long Documents for Dense Retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–377, Bangkok, Thailand. Association for Computational Linguistics.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. MMTEB: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

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

A. Gelman, J. Hill, and A. Vehtari. 2021. *Regression and Other Stories*. Analytical Methods for Social Research. Cambridge University Press.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. 2025. Late chunking: Contextual chunk embeddings using long-context embedding models. *Preprint*, arXiv:2409.04701.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Reagan J. Lee, Samarth Goel, and Kannan Ramchandran. 2025. Quantifying positional biases in text embedding models. *Preprint*, arXiv:2412.15241.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173. Place: Cambridge, MA Publisher: MIT Press.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 54567–54585. Curran Associates, Inc.

Saahil Ognawala and Alex Cureton-Griffiths. 2025. Long-Context Embedding Models are Blind Beyond 4K Tokens.

Juri Opitz, Lucas Moeller, Andrianos Michail, Sebastian Padó, and Simon Clematide. 2025. Interpretable text embeddings and text similarity explanation: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22303–22319, Suzhou, China. Association for Computational Linguistics.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual Embeddings With Task LoRA. *arXiv preprint*. ArXiv:2409.10173 [cs].

Ziyang Zeng, Dun Zhang, Jiacheng Li, Zoupanxiang, Yudong Zhou, and Yuqing Yang. 2025. An empirical study of position bias in modern information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5069–5081, Suzhou, China. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

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

# A Appendix

**Sample of Wikipedia Comparable Corpus**

| | |
|---|---|
| **English** | Buckingham Palace () is a royal residence in London and the administrative headquarters of the monarch of the United Kingdom. Located in the City of Westminster, the palace is often at the centre of state occasions and royal hospitality. It has been a focal point for the British people at times of national rejoicing and mourning. [...] |
| **Chinese** | 白金汉宫（、））是一座位於英國倫敦的皇家寢宮，也是英國君主的行政總部。<br>是英國君主位於倫敦的主要寢宮及辦公處。宮殿坐落在大倫敦西敏市，<br>是國家慶典和王室歡迎禮舉行場地之一，也是一處重要的旅遊景點。在英國歷史上的歡慶或危機時刻，<br>白金漢宮也是一處重要的集會場所。 [...] |
| **German** | Der Buckingham Palace () ist die offizielle Residenz des britischen Monarchen in London. Das Gebäude im Stadtbezirk City of Westminster dient auch offiziellen Staatsanlässen. So werden dort ausländische Staatsoberhäupter bei ihrem Besuch in Großbritannien empfangen. Daneben ist er ein wichtiger Anziehungspunkt für Touristen. [...] |
| **Italian** | Buckingham Palace, situato nella Città di Westminster a Londra, è la residenza ufficiale del sovrano del Regno Unito, attualmente Carlo III. L'espressione "Buckingham Palace" o semplicemente "The Palace" è diventata comune per esprimere tutto ciò che riguarda gli ambienti della corte e della famiglia reale. [...] |
| **Korean** | 버킹엄 궁전()은 영국 런던에 있는 궁전이다. 1703년 버킹엄 공작 존 셰필드의 저택으로 세워진 것을 1761년에 조지 3세에게 양도되어 지금의 모습으로 증개축을 한 후 사저로 이용되다가 1837년 빅토리아 여왕의 즉위식 때에 궁전으로 격상되어 이후 역대 군주들이 상주하였다. 영국 군주의 공식적인 사무실 및 주거지로 쓰이고 있기 때문에 현재 영국 왕실의 대명사이기도 하다. [...] |
| **Hindi** | बकिंघम पैलेस(, ब्रिटिश उच्चारण:बखिंग्हॅम् प़ॅलेस्) ब्रिटिश राजशाही का लंदन स्थित आधिकारिक निवास है। वेस्टमिंस्टर शहर में स्थित यह राजमहल राजकीय आयोजनों और शाही आतिथ्य का केंद्र है। यह ब्रिटेन वासियों के लिये राष्ट्रीय हर्षोन्माद और संकट के समय चर्चा का विषय रहा है। मूलतः बकिंघम हाउस के रूप में जाना [...] |

**Figure 9:** Sample of the multilingual Wikipedia comparable corpus across six languages. Exemplarily, the sample with ID 2059 (Wikipedia article about *Buckingham Palace*) is shown. Every sample consists of the given article in all six languages. Articles are truncated for illustration purposes.

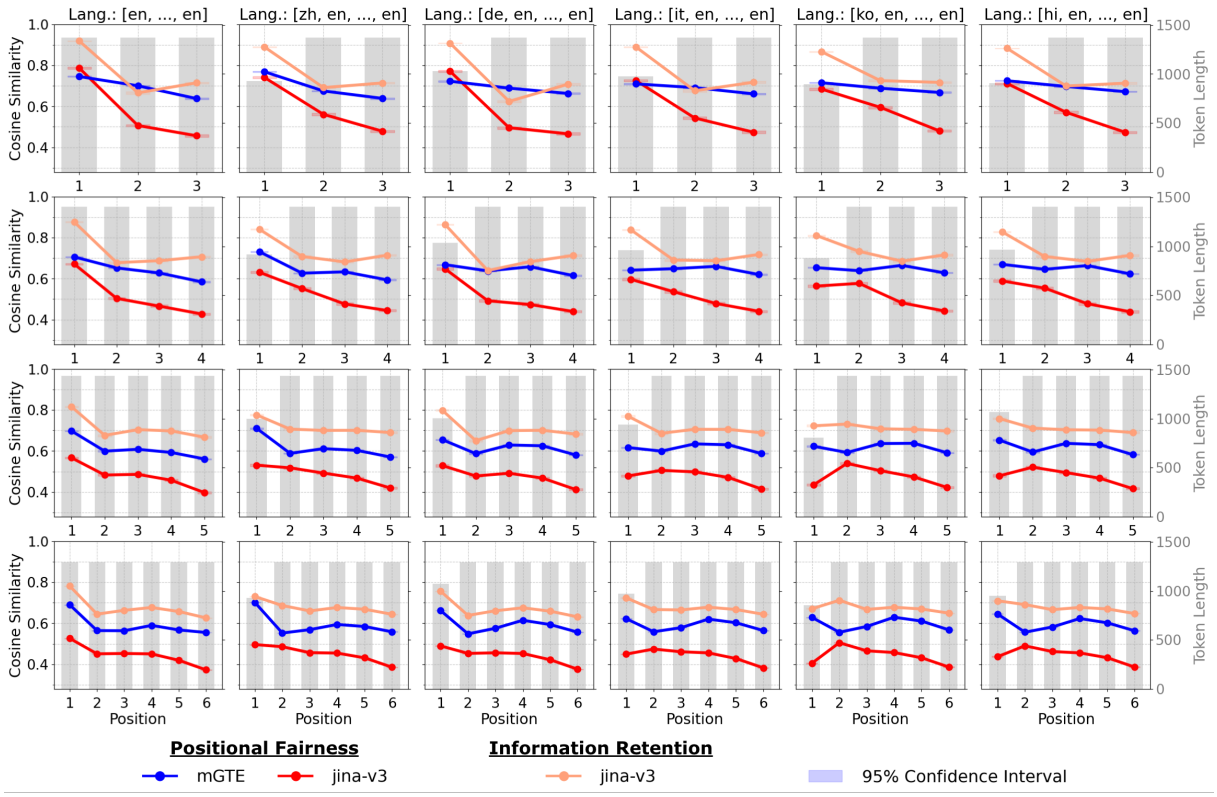



**Figure 10:** Subplots show different English ($\mathcal{L}_{later} = en$) mixed-language experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show (i) average representation in the global document embedding (mGTE and jina-v3), and (ii) average information retention (jina-v3) per segment position. Right y-axes show average token length per segment position (gray bars).

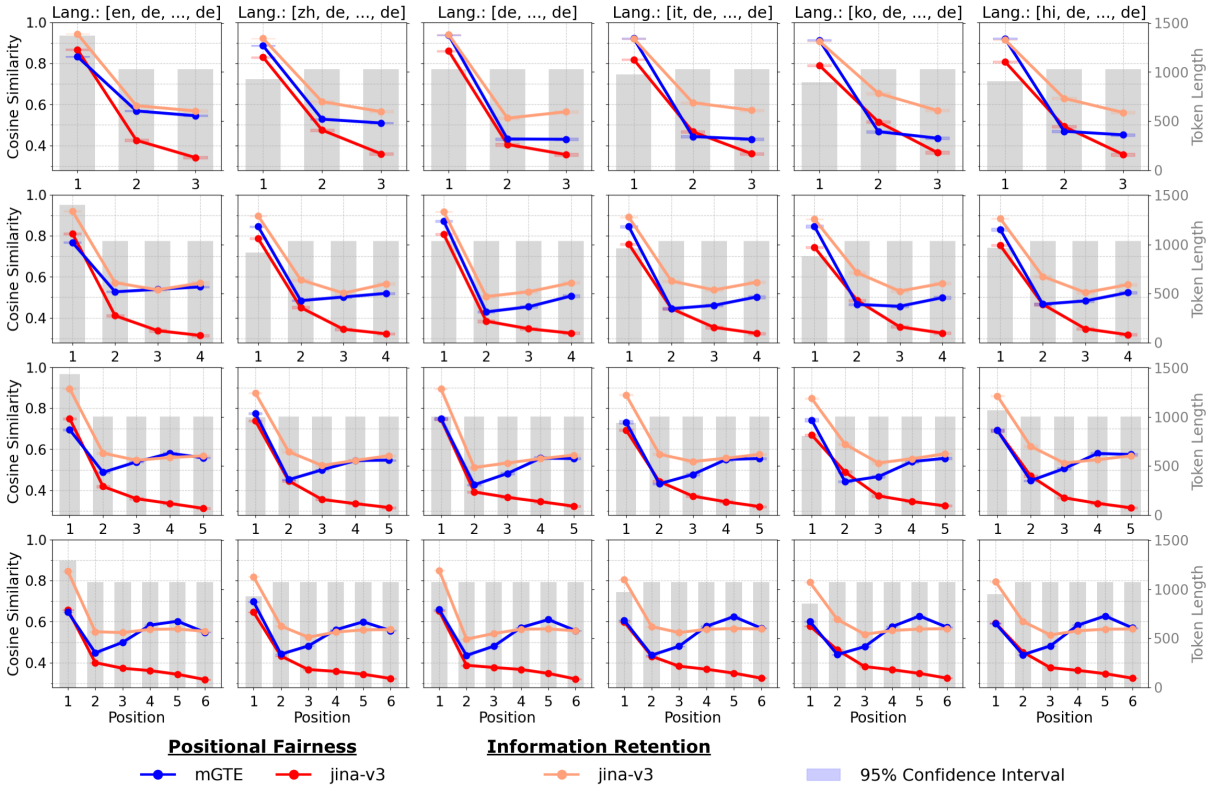



**Figure 11:** Subplots show different German ($\mathcal{L}_{later} = de$) mixed-language experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show (i) average representation in the global document embedding (mGTE and jina-v3), and (ii) average information retention (jina-v3) per segment position. Right y-axes show average token length per segment position (gray bars).

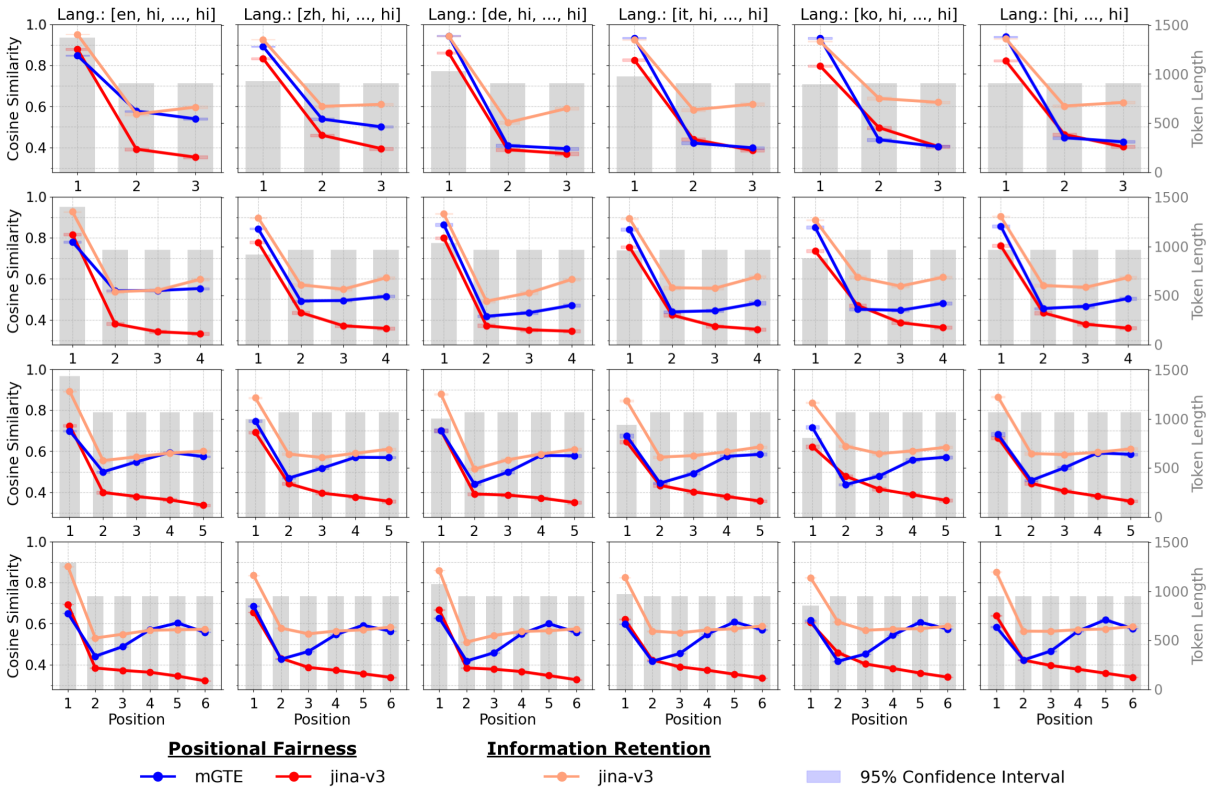



**Figure 12:** Subplots show different Hindi ($\mathcal{L}_{later} = hi$) mixed-language experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show (i) average representation in the global document embedding (mGTE and jina-v3), and (ii) average information retention (jina-v3) per segment position. Right y-axes show average token length per segment position (gray bars).

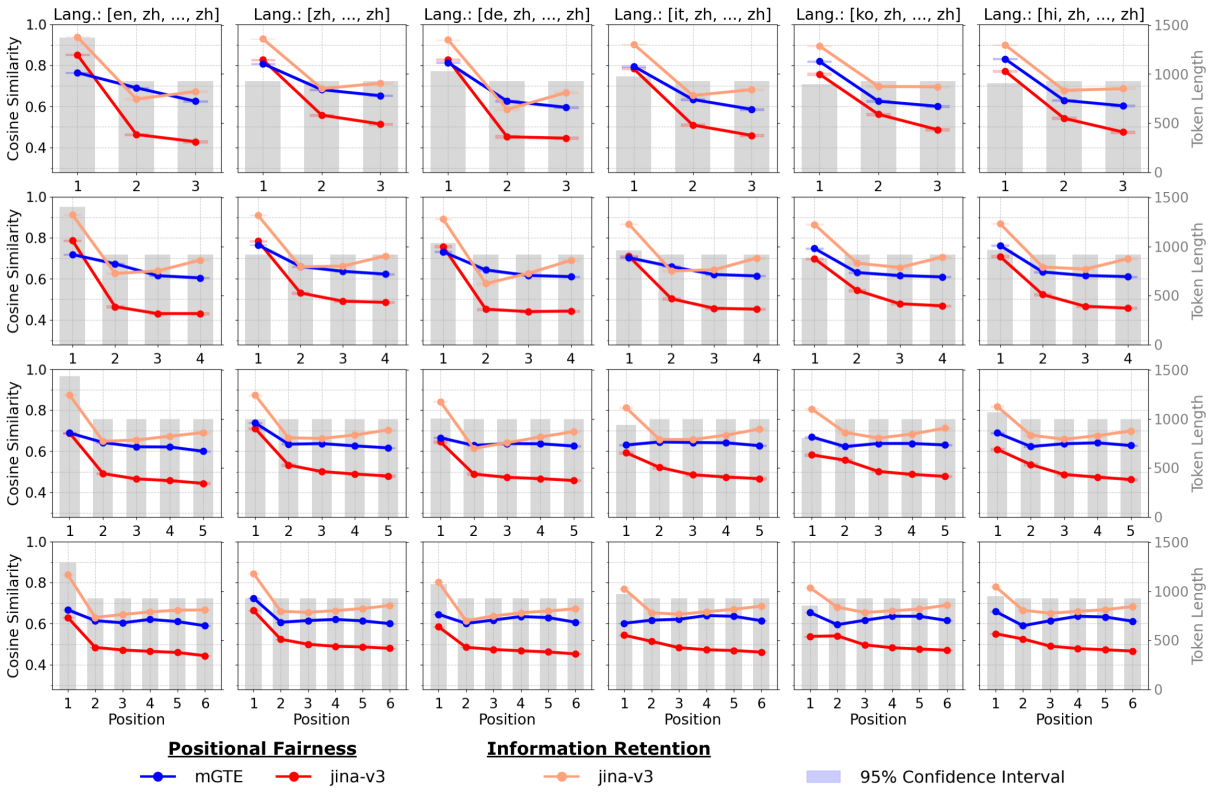



**Figure 13:** Subplots show different Chinese ($\mathcal{L}_{later} = zh$) mixed-language experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show (i) average representation in the global document embedding (mGTE and jina-v3), and (ii) average information retention (jina-v3) per segment position. Right y-axes show average token length per segment position (gray bars).

**OLS Coefficients Positional Fairness (monolingual, mGTE)**

| $n$ | | en | zh | de | it | ko | hi |
|---|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.75 | 0.81 | 0.94 | 0.95 | 0.93 | 0.94 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.05 | -0.13 | **-0.51** | -0.54 | -0.48 | -0.49 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.11 | -0.16 | **-0.51** | -0.56 | -0.52 | -0.51 |
| 4 | $\beta_0$ $(Intercept)$ | 0.71 | 0.77 | 0.87 | 0.89 | 0.87 | 0.86 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.05 | -0.11 | -0.44 | -0.49 | -0.44 | -0.40 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.08 | -0.13 | -0.42 | -0.48 | -0.46 | -0.39 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.12 | -0.14 | -0.36 | -0.44 | -0.43 | -0.35 |
| 5 | $\beta_0$ $(Intercept)$ | 0.70 | 0.74 | 0.75 | 0.82 | 0.85 | 0.68 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.10 | -0.10 | -0.32 | -0.42 | -0.44 | -0.22 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.09 | -0.10 | -0.27 | -0.39 | -0.46 | -0.16$^{\square}$ |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.10 | -0.11 | -0.19$^{\square}$ | -0.34 | -0.41 | -0.09$^{\bullet}$ |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.14 | -0.12 | -0.19$^{\diamond}$ | -0.33 | -0.40 | -0.10$^{\bullet}$ |
| 6 | $\beta_0$ $(Intercept)$ | 0.69 | 0.72 | 0.66 | 0.66 | 0.71 | 0.58 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.13 | -0.12 | **-0.22** | -0.24 | -0.28 | -0.16 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.13 | -0.11 | **-0.18** | -0.21 | -0.26 | -0.12$^{\square}$ |
| | $\beta_4$ $(4\ vs.\ 1)$ | **-0.10** | -0.10 | -0.09$^{\diamond}$ | -0.11$^{\square}$ | -0.21 | -0.02$^{\bullet}$ |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.12 | -0.11 | -0.05$^{\bullet}$ | -0.07$^{\bullet}$ | -0.14$^{\square}$ | 0.03$^{\bullet}$ |
| | $\beta_6$ $(6\ vs.\ 1)$ | **-0.14** | -0.12 | -0.10 | -0.12$^{\diamond}$ | -0.13$^{\diamond}$ | -0.01$^{\bullet}$ |

**Table 1:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the monolingual document setting, using mGTE embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\diamond}\mathfrak{p} < 0.01$, $^{\square}\mathfrak{p} < 0.05$, $^{\bullet}$not significant.

**OLS Coefficients Positional Fairness (monolingual, jina-v3)**

| $n$ | | en | zh | de | it | ko | hi |
|---|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.79 | 0.83 | 0.86 | 0.82 | 0.79 | 0.82 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.28 | -0.27 | -0.46 | -0.37 | -0.30 | -0.36 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.33 | -0.31 | -0.50 | -0.43 | -0.35 | -0.42 |
| 4 | $\beta_0$ $(Intercept)$ | 0.67 | 0.78 | 0.81 | 0.76 | 0.74 | 0.76 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.17 | -0.25 | -0.42 | -0.35 | -0.30 | -0.33 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.20 | -0.29 | -0.46 | -0.40 | -0.35 | -0.38 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.24 | -0.30 | -0.48 | -0.42 | -0.36 | -0.40 |
| 5 | $\beta_0$ $(Intercept)$ | 0.57 | 0.71 | 0.75 | 0.70 | 0.70 | 0.67 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.08 | -0.18 | -0.36 | -0.29 | -0.27 | -0.22 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.08 | -0.21 | -0.38 | -0.33 | -0.33 | -0.26 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.11 | -0.22 | -0.40 | -0.36 | -0.34 | -0.28 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.17 | -0.23 | -0.42 | -0.37 | -0.34 | -0.31 |
| 6 | $\beta_0$ $(Intercept)$ | 0.53 | 0.66 | 0.65 | 0.62 | 0.63 | 0.64 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.08 | -0.14 | -0.26 | -0.20 | -0.20 | -0.22 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.07 | -0.16 | -0.27 | -0.23 | -0.23 | -0.24 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.08 | -0.17 | -0.28 | -0.25 | -0.25 | -0.26 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.11 | -0.18 | -0.30 | -0.27 | -0.26 | -0.28 |
| | $\beta_6$ $(6\ vs.\ 1)$ | -0.15 | -0.18 | -0.33 | -0.29 | -0.27 | -0.30 |

**Table 2:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the monolingual document setting, using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$.

**OLS Coefficients Positional Fairness**
**(mixed-language English, mGTE)**

| $n$ | | [zh, en, ..., en] | [de, en, ..., en] | [it, en, ..., en] | [ko, en, ..., en] | [hi, en, ..., en] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.77 | 0.72 | 0.71 | 0.72 | 0.72 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.09 | -0.03 | -0.02 | -0.03 | -0.03 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.13 | -0.06 | -0.05 | -0.05 | -0.05 |
| 4 | $\beta_0$ $(Intercept)$ | 0.73 | 0.67 | 0.64 | 0.65 | 0.67 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.10 | -0.03 | 0.01$^{\bullet}$ | -0.01$^{\diamond}$ | -0.02 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.10 | -0.01$^{\square}$ | 0.02 | 0.01$^{\square}$ | -0.01$^{\bullet}$ |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.14 | -0.05 | -0.02 | -0.03 | -0.05 |
| 5 | $\beta_0$ $(Intercept)$ | 0.71 | 0.65 | 0.62 | 0.62 | 0.65 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.12 | -0.07 | -0.02$^{\bullet}$ | -0.03$^{\diamond}$ | -0.06 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.10 | -0.03 | 0.02$^{\bullet}$ | 0.01$^{\bullet}$ | -0.02$^{\bullet}$ |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.11 | -0.03 | 0.01$^{\bullet}$ | 0.01$^{\bullet}$ | -0.02$^{\square}$ |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.14 | -0.07 | -0.03$^{\diamond}$ | -0.03 | -0.07 |
| 6 | $\beta_0$ $(Intercept)$ | 0.70 | 0.66 | 0.62 | 0.63 | 0.64 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.15 | -0.11 | -0.06 | -0.07 | -0.09 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.13 | -0.09 | -0.04 | -0.04 | -0.06 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.11 | -0.05 | -0.00$^{\bullet}$ | 0.00$^{\bullet}$ | -0.02 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.12 | -0.07 | -0.02$^{\diamond}$ | -0.02$^{\diamond}$ | -0.04 |
| | $\beta_6$ $(6\ vs.\ 1)$ | -0.14 | -0.11 | -0.06 | -0.06 | -0.08 |

**Table 3:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**English**; $\mathcal{L}_{later} = en$), using mGTE embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\diamond}\mathfrak{p} < 0.01$, $^{\square}\mathfrak{p} < 0.05$, $^{\bullet}$not significant.

**OLS Coefficients Positional Fairness (mixed-language German, mGTE)**

| $n$ | | [en, de, ..., de] | [zh, de, ..., de] | [it, de, ..., de] | [ko, de, ..., de] | [hi, de, ..., de] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.83 | 0.89 | 0.92 | 0.91 | 0.92 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.26 | -0.36 | -0.48 | -0.45 | -0.45 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.29 | -0.38 | -0.49 | -0.48 | -0.47 |
| 4 | $\beta_0$ $(Intercept)$ | 0.77 | 0.84 | 0.84 | 0.85 | 0.83 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.24 | -0.36 | -0.40 | -0.38 | -0.36 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.23 | -0.34 | -0.38 | -0.39 | -0.35 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.22 | -0.33 | -0.34 | -0.35 | -0.31 |
| 5 | $\beta_0$ $(Intercept)$ | 0.69 | 0.78 | 0.73 | 0.74 | 0.69 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.21 | -0.32 | -0.30 | -0.30 | -0.24 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.16 | -0.28 | -0.25 | -0.28 | -0.19 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.11 | -0.23 | -0.18 | -0.20 | -0.11$^{\diamond}$ |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.14 | -0.23 | -0.18 | -0.19 | -0.12 |
| 6 | $\beta_0$ $(Intercept)$ | 0.65 | 0.70 | 0.61 | 0.60 | 0.59 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.20 | -0.26 | -0.17 | -0.16 | -0.15 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.15 | -0.21 | -0.13 | -0.12 | -0.11 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.06 | -0.14 | -0.03$^{\bullet}$ | -0.03$^{\bullet}$ | -0.01$^{\bullet}$ |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.04 | -0.10 | 0.02$^{\bullet}$ | 0.03$^{\bullet}$ | 0.04$^{\square}$ |
| | $\beta_6$ $(6\ vs.\ 1)$ | -0.10 | -0.14 | -0.04$^{\diamond}$ | -0.03$^{\square}$ | -0.02$^{\bullet}$ |

**Table 4:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**German**; $\mathcal{L}_{later} = de$), using mGTE embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\diamond}\mathfrak{p} < 0.01$, $^{\square}\mathfrak{p} < 0.05$, $^{\bullet}$not significant.

**OLS Coefficients Positional Fairness**
**(mixed-language Hindi, mGTE)**

| $n$ | | [en, hi, ..., hi] | [zh, hi, ..., hi] | [de, hi, ..., hi] | [it, hi, ..., hi] | [ko, hi, ..., hi] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0\ (Intercept)$ | 0.85 | 0.89 | 0.94 | 0.93 | 0.93 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.27 | -0.35 | -0.53 | -0.51 | -0.49 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.31 | -0.39 | -0.55 | -0.53 | -0.53 |
| 4 | $\beta_0\ (Intercept)$ | 0.78 | 0.84 | 0.86 | 0.84 | 0.85 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.24 | -0.35 | -0.45 | -0.40 | -0.40 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.24 | -0.35 | -0.43 | -0.40 | -0.40 |
| | $\beta_4\ (4\ vs.\ 1)$ | -0.23 | -0.33 | -0.39 | -0.36 | -0.37 |
| 5 | $\beta_0\ (Intercept)$ | 0.70 | 0.75 | 0.70 | 0.67 | 0.72 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.20 | -0.28 | -0.26 | -0.23 | -0.28 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.15 | -0.23 | -0.20 | -0.18 | -0.24 |
| | $\beta_4\ (4\ vs.\ 1)$ | -0.10 | -0.18 | -0.12$^{\diamond}$ | -0.10$^{\square}$ | -0.16 |
| | $\beta_5\ (5\ vs.\ 1)$ | -0.12 | -0.18 | -0.12 | -0.09$^{\square}$ | -0.15 |
| 6 | $\beta_0\ (Intercept)$ | 0.65 | 0.69 | 0.63 | 0.60 | 0.62 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.21 | -0.26 | -0.21 | -0.18 | -0.20 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.16 | -0.22 | -0.17 | -0.14 | -0.16 |
| | $\beta_4\ (4\ vs.\ 1)$ | -0.08 | -0.14 | -0.08$^{\diamond}$ | -0.05$^{\square}$ | -0.07$^{\diamond}$ |
| | $\beta_5\ (5\ vs.\ 1)$ | -0.05 | -0.09 | -0.03$^{\bullet}$ | 0.01$^{\bullet}$ | -0.01$^{\bullet}$ |
| | $\beta_6\ (6\ vs.\ 1)$ | -0.09 | -0.12 | -0.07 | -0.03$^{\bullet}$ | -0.04$^{\square}$ |

**Table 5:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**Hindi**; $\mathcal{L}_{later} = hi$), using mGTE embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\diamond}\mathfrak{p} < 0.01$, $^{\square}\mathfrak{p} < 0.05$, $^{\bullet}$not significant.

**OLS Coefficients Positional Fairness (mixed-language Chinese, mGTE)**

| $n$ | | [en, zh, ..., zh] | [de, zh, ..., zh] | [it, zh, ..., zh] | [ko, zh, ..., zh] | [hi, zh, ..., zh] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0\ (Intercept)$ | 0.76 | 0.81 | 0.80 | 0.82 | 0.83 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.07 | -0.19 | -0.16 | -0.19 | -0.20 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.14 | -0.22 | -0.21 | -0.22 | -0.23 |
| 4 | $\beta_0\ (Intercept)$ | 0.72 | 0.73 | 0.70 | 0.75 | 0.76 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.05 | -0.09 | -0.04 | -0.12 | -0.13 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.10 | -0.12 | -0.08 | -0.13 | -0.15 |
| | $\beta_4\ (4\ vs.\ 1)$ | -0.11 | -0.12 | -0.09 | -0.14 | -0.15 |
| 5 | $\beta_0\ (Intercept)$ | 0.69 | 0.67 | 0.63 | 0.67 | 0.69 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.05 | -0.04 | $0.01^{\bullet}$ | -0.05 | -0.07 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.07 | $-0.03^{\diamond}$ | $0.01^{\bullet}$ | $-0.03^{\square}$ | -0.05 |
| | $\beta_4\ (4\ vs.\ 1)$ | -0.07 | $-0.03^{\diamond}$ | $0.01^{\bullet}$ | $-0.03^{\square}$ | -0.05 |
| | $\beta_5\ (5\ vs.\ 1)$ | -0.09 | -0.04 | $-0.00^{\bullet}$ | $-0.04^{\diamond}$ | -0.06 |
| 6 | $\beta_0\ (Intercept)$ | 0.67 | 0.65 | 0.60 | 0.65 | 0.66 |
| | $\beta_2\ (2\ vs.\ 1)$ | -0.05 | -0.04 | $0.02^{\bullet}$ | -0.06 | -0.07 |
| | $\beta_3\ (3\ vs.\ 1)$ | -0.06 | -0.03 | $0.02^{\square}$ | -0.04 | -0.04 |
| | $\beta_4\ (4\ vs.\ 1)$ | -0.05 | $-0.01^{\bullet}$ | 0.04 | $-0.02^{\square}$ | $-0.02^{\diamond}$ |
| | $\beta_5\ (5\ vs.\ 1)$ | -0.06 | $-0.02^{\square}$ | 0.03 | $-0.02^{\square}$ | -0.03 |
| | $\beta_6\ (6\ vs.\ 1)$ | -0.08 | -0.04 | $0.01^{\bullet}$ | -0.04 | -0.05 |

**Table 6:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**Chinese**; $\mathcal{L}_{later} = zh$), using mGTE embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\diamond}\mathfrak{p} < 0.01$, $^{\square}\mathfrak{p} < 0.05$, $^{\bullet}$not significant.

**OLS Coefficients Positional Fairness (mixed-language English, jina-v3)**

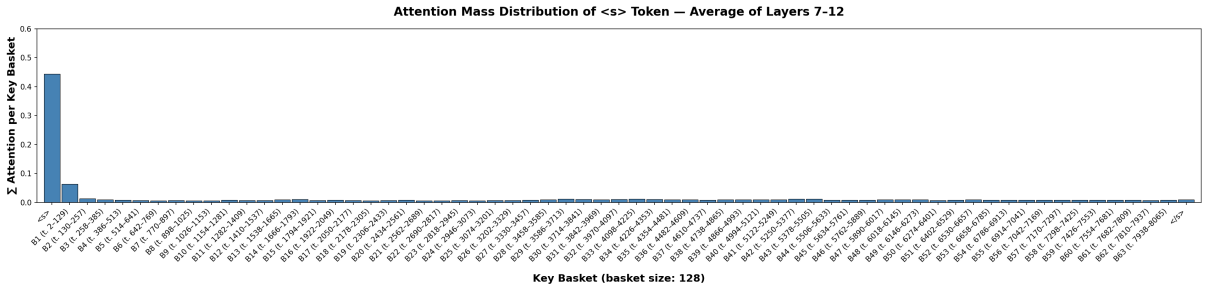
| $n$ | | [zh, en, ..., en] | [de, en, ..., en] | [it, en, ..., en] | [ko, en, ..., en] | [hi, en, ..., en] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.74 | 0.77 | 0.73 | 0.68 | 0.71 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.18 | -0.28 | -0.18 | -0.09 | -0.14 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.26 | -0.30 | -0.25 | -0.20 | -0.24 |
| 4 | $\beta_0$ $(Intercept)$ | 0.63 | 0.65 | 0.60 | 0.56 | 0.59 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.08 | -0.16 | -0.06 | $0.01^{\bullet}$ | -0.04 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.15 | -0.17 | -0.12 | -0.08 | -0.11 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.19 | -0.21 | -0.16 | -0.12 | -0.15 |
| 5 | $\beta_0$ $(Intercept)$ | 0.53 | 0.53 | 0.48 | 0.43 | 0.48 |
| | $\beta_2$ $(2\ vs.\ 1)$ | $-0.01^{\bullet}$ | $-0.05^{\diamond}$ | $0.03^{\bullet}$ | 0.10 | $0.04^{\square}$ |
| | $\beta_3$ $(3\ vs.\ 1)$ | $-0.04^{\diamond}$ | $-0.04^{\square}$ | $0.02^{\bullet}$ | 0.07 | $0.02^{\bullet}$ |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.06 | -0.06 | $-0.01^{\bullet}$ | $0.04^{\square}$ | $-0.01^{\bullet}$ |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.11 | -0.12 | -0.06 | $-0.01^{\bullet}$ | -0.06 |
| 6 | $\beta_0$ $(Intercept)$ | 0.50 | 0.49 | 0.45 | 0.40 | 0.44 |
| | $\beta_2$ $(2\ vs.\ 1)$ | $-0.01^{\bullet}$ | $-0.04^{\square}$ | $0.02^{\bullet}$ | 0.10 | 0.05 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.04 | $-0.03^{\diamond}$ | $0.01^{\bullet}$ | 0.06 | $0.02^{\square}$ |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.04 | $-0.04^{\diamond}$ | $0.01^{\bullet}$ | 0.05 | $0.02^{\bullet}$ |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.06 | -0.07 | $-0.02^{\bullet}$ | $0.03^{\square}$ | $-0.01^{\bullet}$ |
| | $\beta_6$ $(6\ vs.\ 1)$ | -0.11 | -0.11 | -0.07 | $-0.02^{\bullet}$ | -0.05 |

**Table 7:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**English**; $\mathcal{L}_{later} = en$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\diamond}\mathfrak{p} < 0.01$, $^{\square}\mathfrak{p} < 0.05$, $^{\bullet}$not significant.

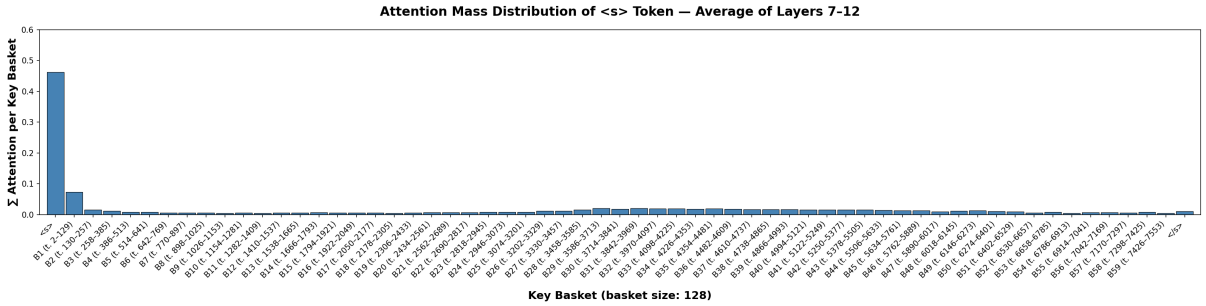
**OLS Coefficients Positional Fairness**
**(mixed-language German, jina-v3)**

| $n$ | | [en, de, …, de] | [zh, de, …, de] | [it, de, …, de] | [ko, de, …, de] | [hi, de, …, de] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.87 | 0.83 | 0.82 | 0.79 | 0.81 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.44 | -0.36 | -0.35 | -0.28 | -0.32 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.53 | -0.47 | -0.46 | -0.42 | -0.45 |
| 4 | $\beta_0$ $(Intercept)$ | 0.81 | 0.79 | 0.76 | 0.74 | 0.75 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.40 | -0.34 | -0.31 | -0.26 | -0.29 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.47 | -0.44 | -0.41 | -0.39 | -0.41 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.50 | -0.46 | -0.43 | -0.42 | -0.44 |
| 5 | $\beta_0$ $(Intercept)$ | 0.75 | 0.74 | 0.69 | 0.67 | 0.69 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.33 | -0.29 | -0.25 | -0.18 | -0.22 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.39 | -0.38 | -0.32 | -0.30 | -0.33 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.41 | -0.40 | -0.35 | -0.32 | -0.36 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.44 | -0.42 | -0.37 | -0.35 | -0.38 |
| 6 | $\beta_0$ $(Intercept)$ | 0.66 | 0.65 | 0.60 | 0.58 | 0.59 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.26 | -0.21 | -0.17 | -0.12 | -0.14 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.28 | -0.28 | -0.22 | -0.20 | -0.22 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.30 | -0.29 | -0.23 | -0.21 | -0.23 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.31 | -0.30 | -0.25 | -0.23 | -0.25 |
| | $\beta_6$ $(6\ vs.\ 1)$ | -0.34 | -0.32 | -0.27 | -0.25 | -0.27 |

**Table 8:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**German**; $\mathcal{L}_{later} = de$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$.




**(a)** English




**(b)** Hindi

**Figure 14:** Front-loaded self-attention distribution of the <s>-query token over key baskets (basket size $\mathfrak{B}$=128) in English (top) and Hindi (bottom) documents ($n$=5). For Hindi, we additionally observe slight mid/late-sequence increases in the attention distribution, leading to U-shaped attention profiles. Average of the last six transformer layers.

**OLS Coefficients Positional Fairness**
**(mixed-language Hindi, jina-v3)**

| $n$ | | [en, hi, ..., hi] | [zh, hi, ..., hi] | [de, hi, ..., hi] | [it, hi, ..., hi] | [ko, hi, ..., hi] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.88 | 0.83 | 0.86 | 0.83 | 0.80 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.49 | -0.37 | -0.47 | -0.39 | -0.30 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.53 | -0.44 | -0.49 | -0.44 | -0.39 |
| 4 | $\beta_0$ $(Intercept)$ | 0.82 | 0.78 | 0.80 | 0.75 | 0.74 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.44 | -0.34 | -0.43 | -0.33 | -0.27 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.47 | -0.41 | -0.45 | -0.39 | -0.35 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.48 | -0.42 | -0.45 | -0.40 | -0.37 |
| 5 | $\beta_0$ $(Intercept)$ | 0.72 | 0.69 | 0.70 | 0.65 | 0.62 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.33 | -0.25 | -0.31 | -0.21 | -0.14 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.34 | -0.30 | -0.31 | -0.24 | -0.21 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.36 | -0.31 | -0.33 | -0.27 | -0.23 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.39 | -0.34 | -0.35 | -0.29 | -0.26 |
| 6 | $\beta_0$ $(Intercept)$ | 0.69 | 0.66 | 0.67 | 0.62 | 0.61 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.31 | -0.22 | -0.28 | -0.20 | -0.15 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.32 | -0.27 | -0.29 | -0.23 | -0.20 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.33 | -0.28 | -0.30 | -0.25 | -0.22 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.35 | -0.30 | -0.32 | -0.27 | -0.25 |
| | $\beta_6$ $(6\ vs.\ 1)$ | -0.37 | -0.32 | -0.34 | -0.29 | -0.27 |

**Table 9:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**Hindi**; $\mathcal{L}_{later} = hi$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$.

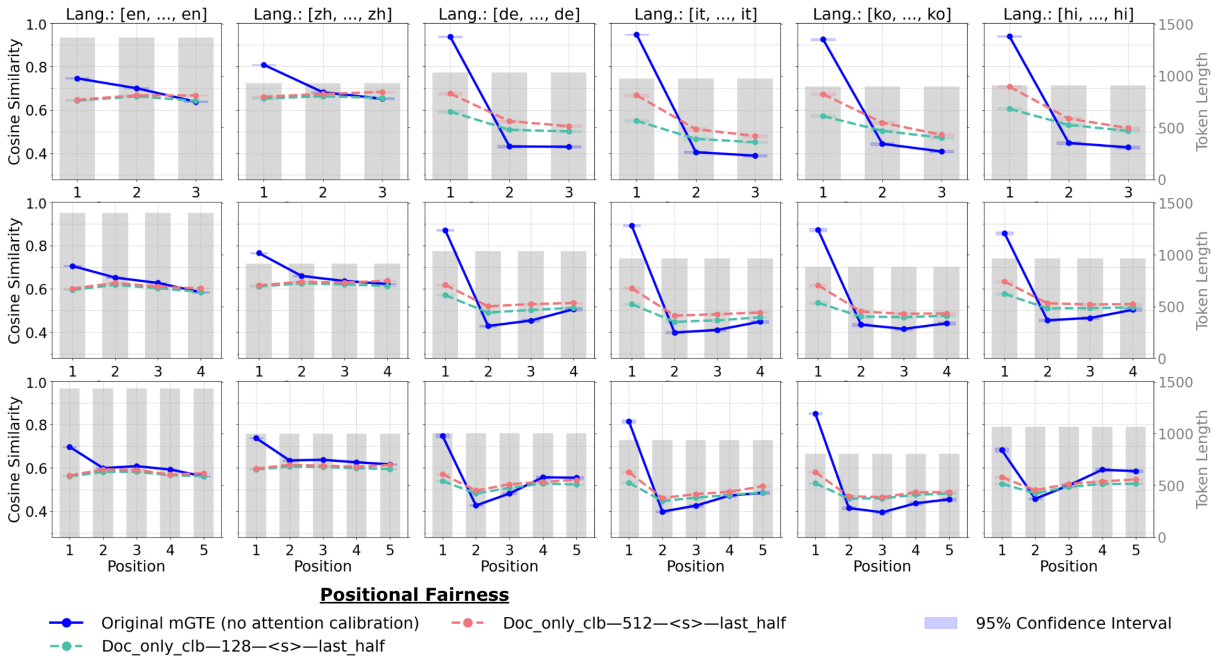



**Figure 15:** Control experiment to test for semantic fidelity of the attention calibration approach. Both dashed and solid lines use uncalibrated segment embeddings. Solid line uses uncalibrated document embeddings; dashed lines use calibrated document embeddings.

**OLS Coefficients Positional Fairness**
**(mixed-language Chinese, jina-v3)**

| $n$ | | [en, zh, ..., zh] | [de, zh, ..., zh] | [it, zh, ..., zh] | [ko, zh, ..., zh] | [hi, zh, ..., zh] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0$ $(Intercept)$ | 0.85 | 0.83 | 0.78 | 0.76 | 0.77 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.39 | -0.38 | -0.27 | -0.20 | -0.23 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.42 | -0.38 | -0.32 | -0.27 | -0.30 |
| 4 | $\beta_0$ $(Intercept)$ | 0.79 | 0.76 | 0.71 | 0.70 | 0.71 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.32 | -0.31 | -0.21 | -0.15 | -0.19 |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.36 | -0.32 | -0.25 | -0.22 | -0.24 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.36 | -0.31 | -0.26 | -0.23 | -0.25 |
| 5 | $\beta_0$ $(Intercept)$ | 0.69 | 0.65 | 0.59 | 0.58 | 0.61 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.20 | -0.16 | -0.07$^{\lozenge}$ | -0.03$^{\bullet}$ | -0.07$^{\lozenge}$ |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.22 | -0.17 | -0.11 | -0.08 | -0.12 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.23 | -0.18 | -0.12 | -0.09 | -0.13 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.24 | -0.19 | -0.13 | -0.10 | -0.15 |
| 6 | $\beta_0$ $(Intercept)$ | 0.63 | 0.58 | 0.54 | 0.54 | 0.55 |
| | $\beta_2$ $(2\ vs.\ 1)$ | -0.15 | -0.10 | -0.03$^{\bullet}$ | 0.00$^{\bullet}$ | -0.03$^{\bullet}$ |
| | $\beta_3$ $(3\ vs.\ 1)$ | -0.16 | -0.11 | -0.06 | -0.04$^{\lozenge}$ | -0.06 |
| | $\beta_4$ $(4\ vs.\ 1)$ | -0.16 | -0.12 | -0.07 | -0.06 | -0.07 |
| | $\beta_5$ $(5\ vs.\ 1)$ | -0.17 | -0.12 | -0.07 | -0.06 | -0.08 |
| | $\beta_6$ $(6\ vs.\ 1)$ | -0.19 | -0.13 | -0.08 | -0.07 | -0.08 |

**Table 10:** Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**Chinese**; $\mathcal{L}_{later} = zh$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\lozenge}\mathfrak{p} < 0.01$, $^{\bullet}$not significant.

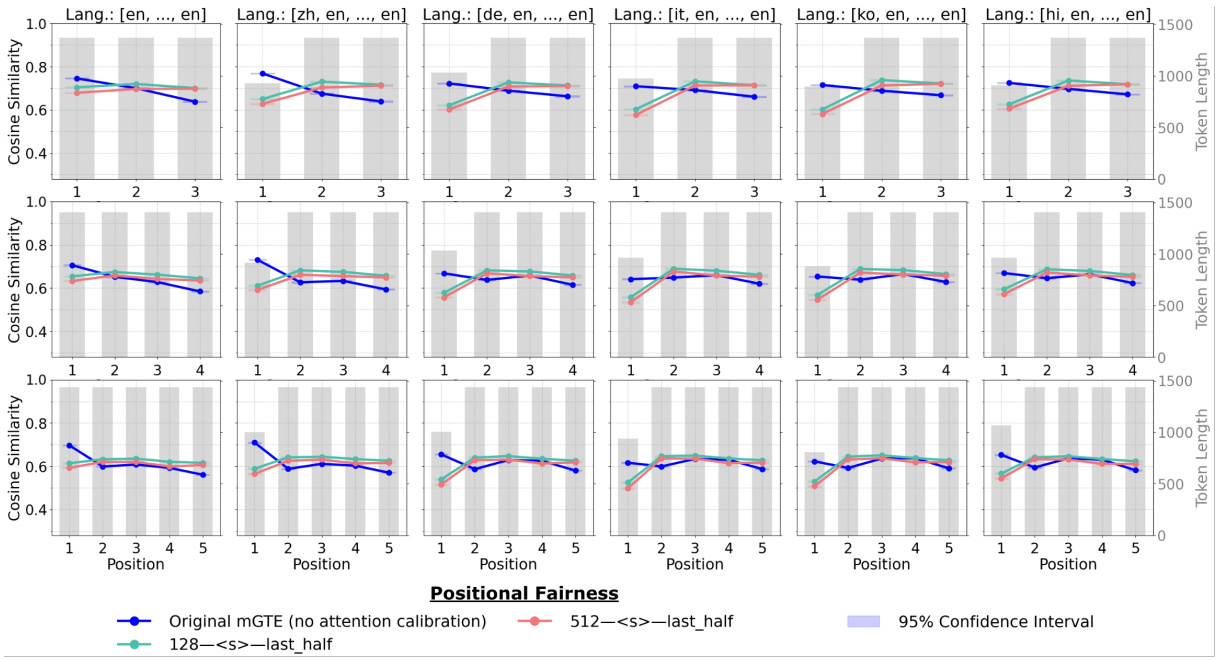



**Figure 16:** Comparison between attention-calibrated and uncalibrated mGTE embeddings. Subplots show different **English** ($\mathcal{L}_{later} = en$) mixed-language experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show average representation in the global document embedding per segment position. Right y-axes show average token length per segment position (gray bars). We show two differently parameterized attention calibrations: *128—<s>—last_half*: $\mathfrak{B}$=128, $\mathfrak{L}^C$={7, ..., 12}; *512—<s>—last_half*: $\mathfrak{B}$=512, $\mathfrak{L}^C$={7, ..., 12}.

**OLS Coefficients Information Retention (monolingual, jina-v3)**

| $n$ | | en | zh | de | it | ko | hi |
|---|---|---|---|---|---|---|---|
| 3 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.92 | 0.93 | 0.94 | 0.93 | 0.92 | 0.93 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.25 | -0.24 | **-0.41** | -0.34 | -0.28 | -0.33 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.20 | -0.22 | **-0.38** | -0.30 | -0.26 | -0.31 |
| 4 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.88 | 0.91 | 0.92 | 0.91 | 0.90 | 0.90 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.20 | -0.25 | -0.41 | -0.34 | -0.30 | -0.34 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.19 | -0.25 | -0.39 | -0.34 | -0.32 | -0.34 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.17 | -0.20 | -0.35 | -0.28 | -0.26 | -0.30 |
| 5 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.82 | 0.87 | 0.89 | 0.88 | 0.88 | 0.86 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.14 | -0.21 | -0.38 | -0.32 | -0.31 | -0.28 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.11 | -0.21 | -0.36 | -0.33 | -0.34 | -0.28 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.12 | -0.19 | -0.34 | -0.31 | -0.31 | -0.27 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.15 | -0.17 | -0.32 | -0.26 | -0.25 | -0.25 |
| 6 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.78 | 0.84 | 0.85 | 0.85 | 0.85 | 0.85 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.14 | -0.18 | **-0.33** | -0.28 | -0.27 | -0.29 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.12 | -0.19 | **-0.31** | -0.28 | -0.28 | -0.29 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.10 | -0.18 | **-0.29** | -0.27 | -0.27 | -0.28 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.12 | -0.17 | **-0.28** | -0.26 | -0.26 | -0.28 |
| | $\beta_6^{(\tau)}$ $(6\ vs.\ 1)$ | -0.16 | -0.16 | **-0.29** | -0.23 | -0.23 | -0.27 |

**Table 11:** Estimated OLS coefficients of the information retention analysis (similarity between a standalone segment embedding and its contextualized embedding within a long document) in the monolingual document setting, using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0^{(\tau)}$ captures the baseline information retention (similarity between standalone embedding of position 1 and contextualized embedding of position 1 within the document), $\beta_{p\geq 2}^{(\tau)}$ captures the difference between position $p$'s information retention (similarity between standalone embedding of position $p$ and contextualized embedding of position $p$ within the document) and the baseline information retention. Statistical significance: all values $\mathtt{p} < 0.001$.

**OLS Coefficients Information Retention (mixed-language English, jina-v3)**

| $n$ | | [zh, en, ..., en] | [de, en, ..., en] | [it, en, ..., en] | [ko, en, ..., en] | [hi, en, ..., en] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.89 | 0.91 | 0.89 | 0.87 | 0.88 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.20 | -0.28 | -0.21 | -0.14 | -0.18 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.18 | -0.20 | -0.17 | -0.15 | -0.17 |
| 4 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.84 | 0.86 | 0.84 | 0.81 | 0.83 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.13 | -0.22 | -0.15 | -0.08 | -0.12 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.16 | -0.18 | -0.15 | -0.12 | -0.14 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.13 | -0.15 | -0.12 | -0.09 | -0.12 |
| 5 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.78 | 0.80 | 0.77 | 0.72 | 0.76 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.07 | -0.15 | -0.08 | 0.01$^{\bullet}$ | -0.05$^{\diamond}$ |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.08 | -0.10 | -0.06 | -0.01$^{\bullet}$ | -0.05 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.07 | -0.10 | -0.06 | -0.02$^{\bullet}$ | -0.06 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.09 | -0.12 | -0.08 | -0.03$^{\bullet}$ | -0.07 |
| 6 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.73 | 0.76 | 0.72 | 0.67 | 0.71 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.04 | -0.12 | -0.06 | 0.04$^{\diamond}$ | -0.02$^{\bullet}$ |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.07 | -0.10 | -0.06 | -0.00$^{\bullet}$ | -0.04 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.05 | -0.08 | -0.05 | 0.01$^{\bullet}$ | -0.03 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.06 | -0.10 | -0.06 | -0.00$^{\bullet}$ | -0.04 |
| | $\beta_6^{(\tau)}$ $(6\ vs.\ 1)$ | -0.09 | -0.13 | -0.08 | -0.02$^{\square}$ | -0.06 |

**Table 12:** Estimated OLS coefficients of the information retention analysis (similarity between a standalone segment embedding and its contextualized embedding within a long document) in the mixed-language document setting (**English**; $\mathcal{L}_{later} = en$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0^{(\tau)}$ captures the baseline information retention (similarity between standalone embedding of position 1 and contextualized embedding of position 1 within the document), $\beta_{p\geq 2}^{(\tau)}$ captures the difference between position $p$'s information retention (similarity between standalone embedding of position $p$ and contextualized embedding of position $p$ within the document) and the baseline information retention. Statistical significance: all values $\mathfrak{p} < 0.001$; unless indicated otherwise: $^{\diamond}\mathfrak{p} < 0.01$, $^{\square}\mathfrak{p} < 0.05$, $^{\bullet}$not significant.

**OLS Coefficients Information Retention (mixed-language German, jina-v3)**

| $n$ | | [en, de, ..., de] | [zh, de, ..., de] | [it, de, ..., de] | [ko, de, ..., de] | [hi, de, ..., de] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.94 | 0.92 | 0.92 | 0.91 | 0.91 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.35 | -0.31 | -0.31 | -0.25 | -0.28 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.38 | -0.36 | -0.35 | -0.34 | -0.35 |
| 4 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.92 | 0.90 | 0.89 | 0.88 | 0.88 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.35 | -0.31 | -0.31 | -0.26 | -0.28 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.38 | -0.38 | -0.35 | -0.35 | -0.36 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.35 | -0.33 | -0.32 | -0.31 | -0.32 |
| 5 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.89 | 0.87 | 0.86 | 0.85 | 0.86 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.31 | -0.29 | -0.29 | -0.22 | -0.25 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.35 | -0.35 | -0.32 | -0.31 | -0.33 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.34 | -0.33 | -0.31 | -0.29 | -0.31 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.33 | -0.31 | -0.29 | -0.27 | -0.29 |
| 6 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.85 | 0.82 | 0.81 | 0.79 | 0.79 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.29 | -0.24 | -0.23 | -0.18 | -0.19 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.30 | -0.29 | -0.26 | -0.25 | -0.26 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.28 | -0.27 | -0.24 | -0.23 | -0.24 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.28 | -0.26 | -0.24 | -0.23 | -0.23 |
| | $\beta_6^{(\tau)}$ $(6\ vs.\ 1)$ | -0.29 | -0.26 | -0.24 | -0.23 | -0.23 |

**Table 13:** Estimated OLS coefficients of the information retention analysis (similarity between a standalone segment embedding and its contextualized embedding within a long document) in the mixed-language document setting (**German**; $\mathcal{L}_{later} = de$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0^{(\tau)}$ captures the baseline information retention (similarity between standalone embedding of position 1 and contextualized embedding of position 1 within the document), $\beta_{p\geq 2}^{(\tau)}$ captures the difference between position $p$'s information retention (similarity between standalone embedding of position $p$ and contextualized embedding of position $p$ within the document) and the baseline information retention. Statistical significance: all values $\mathfrak{p} < 0.001$.

**OLS Coefficients Information Retention (mixed-language Hindi, jina-v3)**

| $n$ | | [en, hi, ..., hi] | [zh, hi, ..., hi] | [de, hi, ..., hi] | [it, hi, ..., hi] | [ko, hi, ..., hi] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.95 | 0.93 | 0.94 | 0.93 | 0.92 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.39 | -0.33 | -0.42 | -0.34 | -0.28 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.35 | -0.32 | -0.35 | -0.31 | -0.30 |
| 4 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.93 | 0.90 | 0.92 | 0.89 | 0.89 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.39 | -0.33 | -0.43 | -0.34 | -0.28 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.38 | -0.35 | -0.38 | -0.34 | -0.32 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.33 | -0.29 | -0.32 | -0.28 | -0.28 |
| 5 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.89 | 0.86 | 0.88 | 0.85 | 0.84 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.34 | -0.27 | -0.37 | -0.27 | -0.21 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.32 | -0.29 | -0.32 | -0.27 | -0.25 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.30 | -0.27 | -0.29 | -0.25 | -0.23 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.29 | -0.25 | -0.27 | -0.22 | -0.22 |
| 6 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.88 | 0.83 | 0.86 | 0.82 | 0.82 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.35 | -0.26 | -0.35 | -0.26 | -0.22 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.33 | -0.28 | -0.32 | -0.27 | -0.26 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.31 | -0.27 | -0.30 | -0.26 | -0.25 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.31 | -0.27 | -0.29 | -0.25 | -0.25 |
| | $\beta_6^{(\tau)}$ $(6\ vs.\ 1)$ | -0.31 | -0.25 | -0.29 | -0.24 | -0.24 |

**Table 14:** Estimated OLS coefficients of the information retention analysis (similarity between a standalone segment embedding and its contextualized embedding within a long document) in the mixed-language document setting (**Hindi**; $\mathcal{L}_{later} = hi$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0^{(\tau)}$ captures the baseline information retention (similarity between standalone embedding of position 1 and contextualized embedding of position 1 within the document), $\beta_{p\geq 2}^{(\tau)}$ captures the difference between position $p$'s information retention (similarity between standalone embedding of position $p$ and contextualized embedding of position $p$ within the document) and the baseline information retention. Statistical significance: all values $\mathfrak{p} < 0.001$.

**OLS Coefficients Information Retention (mixed-language Chinese, jina-v3)**

| $n$ | | [en, zh, ..., zh] | [de, zh, ..., zh] | [it, zh, ..., zh] | [ko, zh, ..., zh] | [hi, zh, ..., zh] |
|---|---|---|---|---|---|---|
| 3 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.94 | 0.92 | 0.90 | 0.89 | 0.90 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.30 | -0.34 | -0.25 | -0.20 | -0.22 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.27 | -0.26 | -0.22 | -0.20 | -0.21 |
| 4 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.91 | 0.89 | 0.87 | 0.86 | 0.87 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.29 | -0.31 | -0.23 | -0.19 | -0.21 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.27 | -0.26 | -0.22 | -0.21 | -0.22 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.22 | -0.20 | -0.17 | -0.16 | -0.17 |
| 5 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.87 | 0.84 | 0.81 | 0.81 | 0.82 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.22 | -0.23 | -0.15 | -0.11 | -0.14 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.22 | -0.20 | -0.16 | -0.14 | -0.16 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.20 | -0.17 | -0.13 | -0.12 | -0.14 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.18 | -0.15 | -0.10 | -0.09 | -0.12 |
| 6 | $\beta_0^{(\tau)}$ $(Intercept)$ | 0.84 | 0.80 | 0.77 | 0.78 | 0.78 |
| | $\beta_2^{(\tau)}$ $(2\ vs.\ 1)$ | -0.21 | -0.19 | -0.12 | -0.10 | -0.12 |
| | $\beta_3^{(\tau)}$ $(3\ vs.\ 1)$ | -0.19 | -0.17 | -0.12 | -0.12 | -0.13 |
| | $\beta_4^{(\tau)}$ $(4\ vs.\ 1)$ | -0.18 | -0.15 | -0.11 | -0.11 | -0.12 |
| | $\beta_5^{(\tau)}$ $(5\ vs.\ 1)$ | -0.17 | -0.14 | -0.10 | -0.10 | -0.11 |
| | $\beta_6^{(\tau)}$ $(6\ vs.\ 1)$ | -0.17 | -0.13 | -0.09 | -0.09 | -0.10 |

**Table 15:** Estimated OLS coefficients of the information retention analysis (similarity between a standalone segment embedding and its contextualized embedding within a long document) in the mixed-language document setting (**Chinese**; $\mathcal{L}_{later} = zh$), using jina-v3 embeddings. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0^{(\tau)}$ captures the baseline information retention (similarity between standalone embedding of position 1 and contextualized embedding of position 1 within the document), $\beta_{p\geq 2}^{(\tau)}$ captures the difference between position $p$'s information retention (similarity between standalone embedding of position $p$ and contextualized embedding of position $p$ within the document) and the baseline information retention. Statistical significance: all values $\mathfrak{p} < 0.001$.

**OLS Coefficients Positional Fairness (monolingual, mGTE, calibration effect)**

| $n$ | | en | | zh | | de | | it | | ko | | hi | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* |
| 3 | $\beta_0$ | 0.70 | 0.75 | 0.71 | 0.81 | 0.68 | 0.94 | 0.64 | 0.95 | 0.69 | 0.93 | 0.69 | 0.94 |
| | $\beta_2$ | 0.02 | -0.05 | 0.01 | -0.13 | -0.08 | -0.51 | -0.09 | -0.54 | -0.07 | -0.48 | -0.08 | -0.49 |
| | $\beta_3$ | -0.00 | -0.11 | -0.00 | -0.16 | -0.09 | -0.51 | -0.10 | -0.56 | -0.11 | -0.52 | -0.10 | -0.51 |
| 4 | $\beta_0$ | 0.65 | 0.71 | 0.67 | 0.77 | 0.65 | 0.87 | 0.61 | 0.89 | 0.64 | 0.87 | 0.66 | 0.86 |
| | $\beta_2$ | 0.02 | -0.05 | 0.01 | -0.11 | -0.07 | -0.44 | -0.08 | -0.49 | -0.07 | -0.44 | -0.07 | -0.40 |
| | $\beta_3$ | 0.01 | -0.08 | 0.01 | -0.13 | -0.07 | -0.42 | -0.07 | -0.48 | -0.07 | -0.46 | -0.07 | -0.39 |
| | $\beta_4$ | -0.01 | -0.12 | -0.00 | -0.14 | -0.06 | -0.36 | -0.06 | -0.44 | -0.06 | -0.43 | -0.06 | -0.35 |
| 5 | $\beta_0$ | 0.61 | 0.70 | 0.65 | 0.74 | 0.61 | 0.75 | 0.61 | 0.82 | 0.61 | 0.85 | 0.61 | 0.68 |
| | $\beta_2$ | 0.02 | -0.10 | 0.01 | -0.10 | -0.05 | -0.32 | -0.08 | -0.42 | -0.07 | -0.44 | -0.04 | -0.22 |
| | $\beta_3$ | 0.02 | -0.09 | 0.01 | -0.10 | -0.03 | -0.27 | -0.07 | -0.39 | -0.07 | -0.46 | -0.01 | -0.16 |
| | $\beta_4$ | 0.01 | -0.10 | 0.01 | -0.11 | -0.01 | -0.19 | -0.06 | -0.34 | -0.05 | -0.41 | -0.00 | -0.09 |
| | $\beta_5$ | 0.00 | -0.14 | 0.00 | -0.12 | -0.02 | -0.19 | -0.04 | -0.33 | -0.05 | -0.40 | 0.00 | -0.10 |

**Table 16:** Comparison between attention-calibrated (c.) and uncalibrated (*uc.*) mGTE embeddings. Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the monolingual document setting. $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Calibrated embeddings use the following hyperparameters: $\mathfrak{B} = 128$, $\mathfrak{L}^C = \{7, \ldots, 12\}$.

**OLS Coefficients Positional Fairness (mixed-language English, mGTE, calibrated)**

| $n$ | | [zh, en, ..., en] | | [de, en, ..., en] | | [it, en, ..., en] | | [ko, en, ..., en] | | [hi, en, ..., en] | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* |
| 3 | $\beta_0$ | 0.65 | 0.77 | 0.62 | 0.72 | 0.60 | 0.71 | 0.60 | 0.72 | 0.63 | 0.72 |
| | $\beta_2$ | 0.08 | -0.09 | 0.11 | -0.03 | 0.13 | -0.02 | 0.14 | -0.03 | 0.11 | -0.03 |
| | $\beta_3$ | 0.07 | -0.13 | 0.09 | -0.06 | 0.11 | -0.05 | 0.12 | -0.05 | 0.09 | -0.05 |
| 4 | $\beta_0$ | 0.61 | 0.73 | 0.58 | 0.67 | 0.56 | 0.64 | 0.57 | 0.65 | 0.59 | 0.67 |
| | $\beta_2$ | 0.07 | -0.10 | 0.10 | -0.03 | 0.13 | -0.01 | 0.12 | -0.01 | 0.09 | -0.02 |
| | $\beta_3$ | 0.06 | -0.10 | 0.10 | -0.01 | 0.12 | 0.02 | 0.11 | 0.01 | 0.08 | -0.01 |
| | $\beta_4$ | 0.05 | -0.14 | 0.08 | -0.05 | 0.10 | -0.02 | 0.10 | -0.03 | 0.07 | -0.05 |
| 5 | $\beta_0$ | 0.59 | 0.71 | 0.54 | 0.65 | 0.53 | 0.62 | 0.53 | 0.62 | 0.57 | 0.65 |
| | $\beta_2$ | 0.05 | -0.12 | 0.10 | -0.07 | 0.12 | -0.02 | 0.12 | -0.03 | 0.08 | -0.06 |
| | $\beta_3$ | 0.06 | -0.10 | 0.11 | -0.03 | 0.12 | 0.02 | 0.12 | 0.01 | 0.08 | -0.02 |
| | $\beta_4$ | 0.04 | -0.11 | 0.10 | -0.03 | 0.11 | 0.01 | 0.11 | 0.01 | 0.07 | -0.02 |
| | $\beta_5$ | 0.04 | -0.14 | 0.09 | -0.07 | 0.10 | -0.03 | 0.10 | -0.03 | 0.06 | -0.07 |

**Table 17:** Comparison between attention-calibrated (c.) and uncalibrated (*uc.*) mGTE embeddings. Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**English**; $\mathcal{L}_{later} = en$). $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Calibrated embeddings use the following hyperparameters: $\mathfrak{B} = 128$, $\mathfrak{L}^C = \{7, \ldots, 12\}$.

**OLS Coefficients Positional Fairness**
**(mixed-language German, mGTE, calibrated)**

| $n$ | | [en, de, …, de] | | [zh, de, …, de] | | [it, de, …, de] | | [ko, de, …, de] | | [hi, de, …, de] | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* |
| 3 | $\beta_0$ | 0.74 | 0.83 | 0.69 | 0.89 | 0.67 | 0.92 | 0.67 | 0.91 | 0.70 | 0.92 |
| | $\beta_2$ | -0.09 | -0.26 | -0.03 | -0.36 | -0.06 | -0.48 | -0.04 | -0.45 | -0.07 | -0.45 |
| | $\beta_3$ | -0.11 | -0.29 | -0.05 | -0.38 | -0.08 | -0.49 | -0.08 | -0.48 | -0.10 | -0.47 |
| 4 | $\beta_0$ | 0.68 | 0.77 | 0.65 | 0.84 | 0.63 | 0.84 | 0.64 | 0.85 | 0.66 | 0.83 |
| | $\beta_2$ | -0.07 | -0.24 | -0.04 | -0.36 | -0.05 | -0.40 | -0.04 | -0.38 | -0.06 | -0.36 |
| | $\beta_3$ | -0.07 | -0.23 | -0.04 | -0.34 | -0.05 | -0.38 | -0.05 | -0.39 | -0.07 | -0.35 |
| | $\beta_4$ | -0.08 | -0.22 | -0.05 | -0.33 | -0.04 | -0.34 | -0.04 | -0.35 | -0.06 | -0.31 |
| 5 | $\beta_0$ | 0.64 | 0.69 | 0.63 | 0.78 | 0.59 | 0.73 | 0.60 | 0.74 | 0.62 | 0.69 |
| | $\beta_2$ | -0.07 | -0.21 | -0.05 | -0.32 | -0.04 | -0.30 | -0.03 | -0.30 | -0.06 | -0.24 |
| | $\beta_3$ | -0.05 | -0.16 | -0.04 | -0.28 | -0.02 | -0.25 | -0.02 | -0.28 | -0.04 | -0.19 |
| | $\beta_4$ | -0.05 | -0.11 | -0.03 | -0.23 | 0.00 | -0.18 | 0.00 | -0.20 | -0.02 | -0.11 |
| | $\beta_5$ | -0.06 | -0.14 | -0.04 | -0.23 | -0.00 | -0.18 | 0.00 | -0.19 | -0.03 | -0.12 |

**Table 18:** Comparison between attention-calibrated (c.) and uncalibrated (*uc.*) mGTE embeddings. Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**German**; $\mathcal{L}_{later} = de$). $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Calibrated embeddings use the following hyperparameters: $\mathfrak{B} = 128$, $\mathfrak{L}^C = \{7, \ldots, 12\}$.

**OLS Coefficients Positional Fairness**
**(mixed-language Chinese, mGTE, calibrated)**

| $n$ | | [en, zh, …, zh] | | [de, zh, …, zh] | | [it, zh, …, zh] | | [ko, zh, …, zh] | | [hi, zh, …, zh] | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* | c. | *uc.* |
| 3 | $\beta_0$ | 0.74 | 0.76 | 0.67 | 0.81 | 0.66 | 0.80 | 0.66 | 0.82 | 0.68 | 0.83 |
| | $\beta_2$ | -0.05 | -0.07 | 0.01 | -0.19 | 0.03 | -0.16 | 0.04 | -0.19 | 0.01 | -0.20 |
| | $\beta_3$ | -0.06 | -0.14 | -0.00 | -0.22 | 0.01 | -0.21 | 0.02 | -0.22 | -0.00 | -0.23 |
| 4 | $\beta_0$ | 0.69 | 0.72 | 0.62 | 0.73 | 0.60 | 0.70 | 0.62 | 0.75 | 0.65 | 0.76 |
| | $\beta_2$ | -0.03 | -0.05 | 0.05 | -0.09 | 0.07 | -0.04 | 0.05 | -0.12 | 0.02 | -0.13 |
| | $\beta_3$ | -0.04 | -0.10 | 0.04 | -0.12 | 0.07 | -0.08 | 0.04 | -0.13 | 0.01 | -0.15 |
| | $\beta_4$ | -0.05 | -0.11 | 0.03 | -0.12 | 0.06 | -0.09 | 0.04 | -0.14 | 0.01 | -0.15 |
| 5 | $\beta_0$ | 0.66 | 0.69 | 0.58 | 0.67 | 0.56 | 0.63 | 0.58 | 0.67 | 0.61 | 0.69 |
| | $\beta_2$ | -0.02 | -0.05 | 0.07 | -0.04 | 0.10 | 0.01 | 0.08 | -0.05 | 0.04 | -0.07 |
| | $\beta_3$ | -0.02 | -0.07 | 0.08 | -0.03 | 0.10 | 0.01 | 0.08 | -0.03 | 0.04 | -0.05 |
| | $\beta_4$ | -0.02 | -0.07 | 0.07 | -0.03 | 0.09 | 0.01 | 0.08 | -0.03 | 0.04 | -0.05 |
| | $\beta_5$ | -0.03 | -0.09 | 0.07 | -0.04 | 0.09 | -0.00 | 0.08 | -0.04 | 0.03 | -0.06 |

**Table 19:** Comparison between attention-calibrated (c.) and uncalibrated (*uc.*) mGTE embeddings. Estimated OLS coefficients of the positional fairness analysis (similarity between document embedding and standalone segment embeddings) in the mixed-language document setting (**Chinese**; $\mathcal{L}_{later} = zh$). $n$ denotes the number of segments per document. OLS regression uses positions as categorical variables; $\beta_0$ captures the baseline (similarity between document embedding and standalone embedding at position 1), $\beta_{p\geq 2}$ captures the difference between position $p$'s similarity (similarity between document embedding and standalone embedding at position $p$) and the baseline similarity. Calibrated embeddings use the following hyperparameters: $\mathfrak{B} = 128$, $\mathfrak{L}^C = \{7, \ldots, 12\}$.

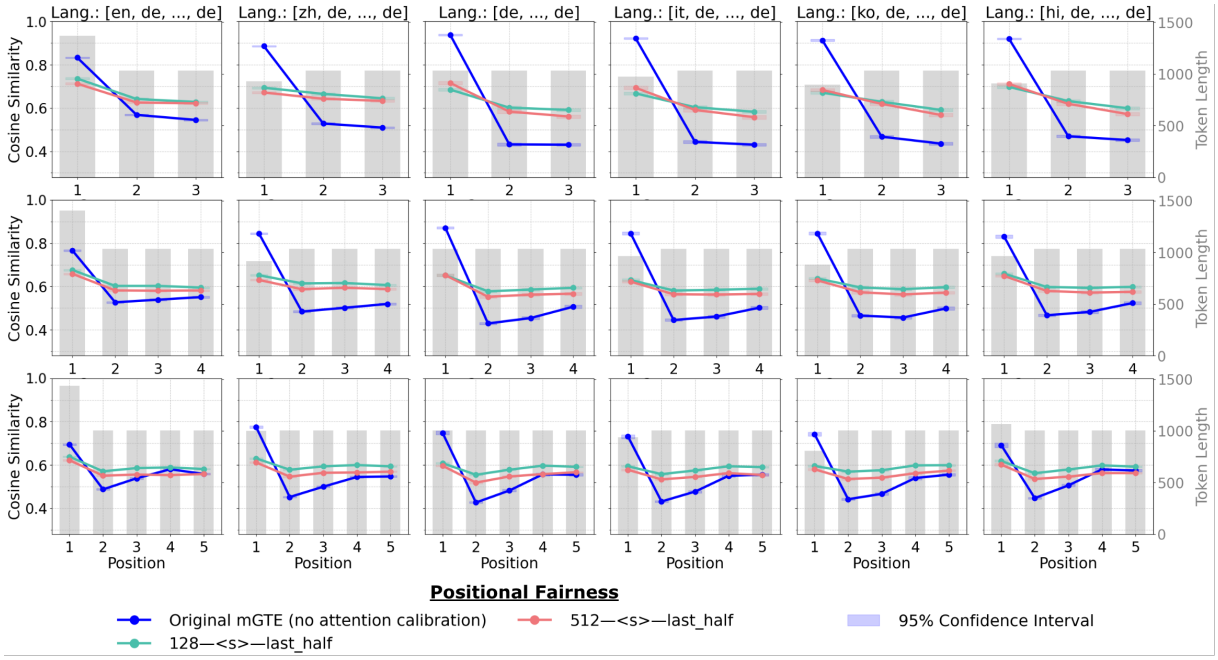



**Figure 17:** Comparison between attention-calibrated and uncalibrated mGTE embeddings. Subplots show different **German** ($\mathcal{L}_{later} = de$) mixed-language experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show average representation in the global document embedding per segment position. Right y-axes show average token length per segment position (gray bars). We show two differently parameterized attention calibrations: *128—<s>—last_half*: $\mathfrak{B}$=128, $\mathfrak{L}^C$={7, . . . , 12}; *512—<s>—last_half*: $\mathfrak{B}$=512, $\mathfrak{L}^C$={7, . . . , 12}.

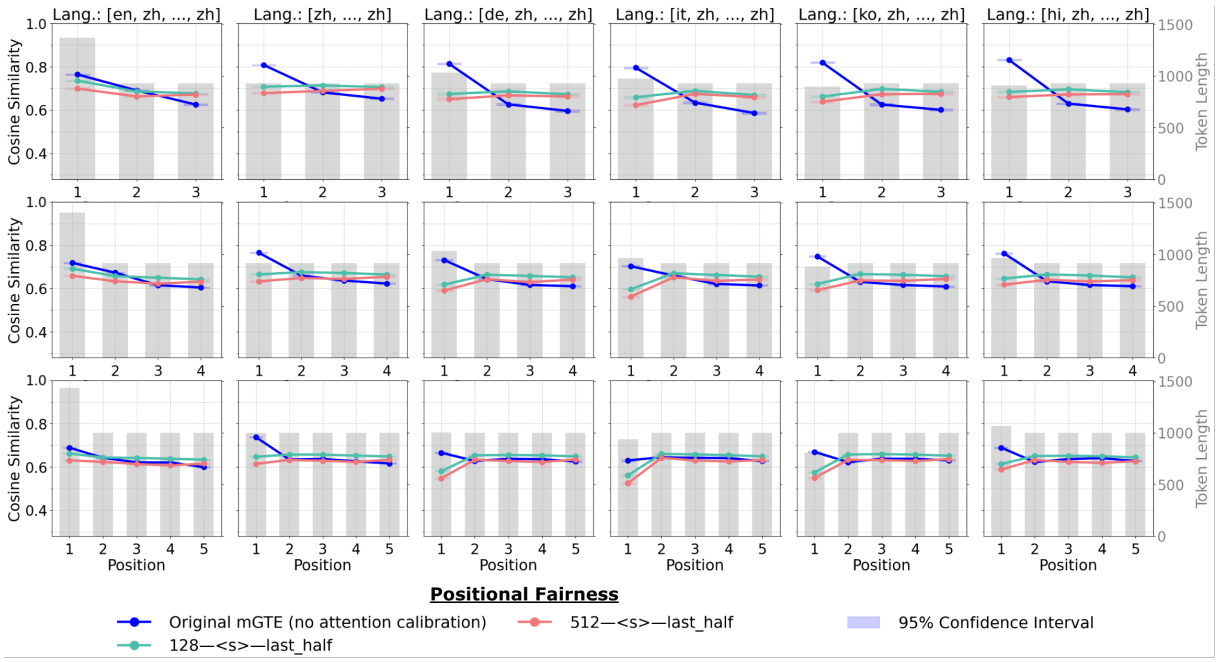



**Figure 18:** Comparison between attention-calibrated and uncalibrated mGTE embeddings. Subplots show different **Chinese** ($\mathcal{L}_{later} = zh$) mixed-language experiment instances $(n, \mathbf{L})$, where $n$ varies across rows, and $\mathbf{L}$ varies across columns. Left y-axes show average representation in the global document embedding per segment position. Right y-axes show average token length per segment position (gray bars). We show two differently parameterized attention calibrations: *128—<s>—last_half*: $\mathfrak{B}$=128, $\mathfrak{L}^C$={7, . . . , 12}; *512—<s>—last_half*: $\mathfrak{B}$=512, $\mathfrak{L}^C$={7, . . . , 12}.