PoSum-Bench: Benchmarking Position Bias in LLM-based Conversational Summarization

Xu SUN[1,2], Lionel Delphin-Poulat[1], Christèle Tarnec[1], Anastasia Shimorina[1]

[1]Orange Research, [2]Université Paris Cité (France)
{firstname.lastname}@orange.com
Correspondence: xu.sun@orange.com

Abstract

Large language models (LLMs) are increasingly used for zero-shot conversation summarization, but often exhibit positional bias—tending to overemphasize content from the beginning or end of a conversation while neglecting the middle. To address this issue, we introduce PoSum-Bench, a comprehensive benchmark for evaluating positional bias in conversational summarization, featuring diverse English and French conversational datasets spanning formal meetings, casual conversations, and customer service interactions. We propose a novel semantic similarity-based sentence-level metric to quantify the direction and magnitude of positional bias in model-generated summaries, enabling systematic and reference-free evaluation across conversation positions, languages, and conversational contexts. Our benchmark and methodology thus provide the first systematic framework for reference-free evaluation of positional bias in conversational summarization, laying the groundwork for developing more balanced and unbiased summarization models[1].

1 Introduction

The rapid advancement of large language models (LLMs) has enabled zero-shot abstractive summarization of complex inputs such as dialogues and meetings across diverse domains (Zhang et al., 2024; Goyal et al., 2023). However, growing reliance on LLM-based summarization has raised concerns regarding biases in content selection and emphasis. Among these, *positional bias*—the tendency of models to favor information from specific parts of the input (e.g., early or late context) while neglecting other content—remains critically underexplored (Chhabra et al., 2024; Liu et al., 2024; Grenander et al., 2019; Zhu et al., 2021b). Such biases can undermine the completeness and fidelity of conversation summaries, where key information may appear throughout the discourse.

Existing methods for evaluating positional bias, such as n-gram mapping (Kim et al., 2019; Zhao et al., 2022), primarily rely on surface lexical matching and often fail to capture deeper semantic relationships. Furthermore, prior work typically treats positional bias as a monolithic phenomenon or focuses narrowly on lead bias (Ravaut et al., 2024; Liu et al., 2024), limiting the granularity and generalizability of analysis.

In this work, we introduce **PoSum-Bench** (**Po**sitional Bias and **Sum**marisation **Bench**mark), a comprehensive benchmark and evaluation framework for analyzing positional bias in conversational summarization. Our key contributions are:

1. **PoSum-Bench:** the first large-scale benchmark explicitly designed to assess positional bias across English and French conversational datasets, covering formal meetings, casual dialogues, and customer service interactions.

2. **Semantic-based evaluation methodology:** a novel framework that quantifies both the direction and magnitude of positional bias at the sentence level by aligning summary sentences with source utterances based on embedding-based semantic similarity, thereby capturing paraphrased or reworded content beyond surface overlap.

3. **Empirical analysis:** extensive experiments reveal that positional bias varies with conversation length, with stronger biases in shorter dialogues.

By providing a unified metric and diverse evaluation suite, PoSum-Bench enables systematic, fine-grained, and cross-linguistic analysis of positional bias in LLM-based conversational summarization.

[1] The PoSum-Bench is available at https://huggingface.co/datasets/Orange/POSUM_BENCH.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7985–8009
November 4-9, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Our work lies at the intersection of three research areas: conversational summarization, positional bias in text processing, and metrics for positional bias evaluation.

Conversational Summarization and Zero-Shot LLMs Conversational summarization aims to distill the essential information from dialogues into concise passages, enabling users to grasp key points without reviewing the often complex, multi-speaker context. While early summarization research focused primarily on news and documents, conversation summarization introduces unique challenges due to its semi-structured, speaker-shifting nature and lack of clear discourse organization (Chen and Yang, 2020; Feng et al., 2022; Rennard et al., 2023). The emergence of LLMs has made effective zero-shot summarization across domains possible (Zhang et al., 2024), with open-source models increasingly approaching the performance of proprietary counterparts (Bai et al., 2023; DeepSeek-AI et al., 2025). However, the biases these models introduce in handling conversational structures, particularly regarding positional information, remain underexplored.

Positional Bias in Summarization Positional bias has been widely studied in news summarization, where *lead bias*—the preference for early sentences—arises from journalistic writing conventions (Kedzie et al., 2018; Liu and Lapata, 2019; Xing et al., 2021; Zhu et al., 2021b). In contrast, dialogues and meetings distribute important information more evenly, rendering positional heuristics problematic. Recent findings highlight a "U-shaped" attention pattern in LLMs processing long inputs, where content at the beginning and end is favored over the middle (Liu et al., 2024; Ravaut et al., 2024). This pattern is particularly detrimental for conversational summarization, where key information often occurs mid-discussion. Unlike prior work that isolates lead or recency bias (Zhu et al., 2021b), our framework provides a unified quantification of both phenomena.

Metrics for Positional Bias Measurement Traditional evaluation metrics like ROUGE (Lin, 2004) were primarily developed for general summary quality evaluation rather than positional bias measurement, focusing on n-gram overlap with reference summaries and offering little insight into how generated summaries utilize different parts of the source. Most existing methods are reference-based, limiting applicability in low-resource settings. Recent semantic metrics, such as BERTScore (Zhang et al., 2020), similarly focus on summary quality assessment rather than positional analysis, and while they improve meaning similarity measurement, they still center on references rather than source utilization.

Recent studies have explored positional bias in summarization through various lexical mapping approaches. Ravaut et al. (2024) adopted a reference-free approach, computing the relative position of summary bigrams within source documents by dividing texts into 20 equal bins and measuring the distribution of matched bigrams across these bins. Similarly, Zhu et al. (2021a) investigated positional bias by tracking non-stop summary words in news interview transcripts, dividing positions into 100 bins and showing the important information laying both at the beginning and at the end. Wu et al. (2023) also applied a comparable binning strategy to examine the distribution of summary words in transcripts, demonstrating that meeting transcripts typically concentrate key information at extremities. While these approaches provide valuable insights into content positioning, they rely primarily on exact n-gram matching, potentially missing semantically equivalent but lexically distinct content. Furthermore, these methods primarily serve as analytical tools rather than formalized metrics for quantifying positional bias.

Our method advances this line of work by (1) aligning summary sentences to source utterances based on semantic similarity rather than lexical matching, (2) formalizing positional alignment into an interpretable bias score, and (3) adapting evaluation for multi-turn, multi-speaker conversations across languages. Crucially, our approach is entirely reference-free, directly comparing generated summaries to their source conversations without requiring human-written references. This enables robust positional bias evaluation even in domains and languages where annotated summaries are unavailable.

By combining a comprehensive benchmark with a robust, semantic-driven evaluation methodology, our work provides the first systematic framework for analyzing positional bias in conversational summarization and dialogue types, addressing limitations of prior approaches.

3 Benchmark Construction

We define a multi-turn conversation as a sequence of textual utterances:

$$C = \{c_1, c_2, \ldots, c_n\}, \tag{1}$$

where each c_i represents the textual content of the i-th turn in the conversation. The objective is to generate a summary:

$$S = \{s_1, s_2, \ldots, s_m\}, \tag{2}$$

where each s_j denotes a sentence, that succinctly captures the salient semantics and key information conveyed throughout the conversation.

In the zero-shot setting, a pre-trained language model, parameterized by θ, is employed to generate the summary directly, conditioned on a given prompt p. The summarization task can be formally expressed as:

$$S = f(C; \theta, p) = \text{argmax}_{S'} P(S' \mid C, p; \theta), \tag{3}$$

where $f(C; \theta, p)$ denotes the function implemented by the language model, which maps the input conversation C and prompt p to an output summary S. The term $P(S' \mid C, p; \theta)$ represents the conditional probability of a candidate summary S' given the conversation and the prompt, and the argmax operation selects the summary that maximizes this probability.

To rigorously investigate positional bias in conversation summarization, we construct a bilingual conversational dataset in English and French, spanning multiple domains. As illustrated in Figure 1, the overall workflow consists of five stages: (1) **Data Collection**: Gathering real-world conversational data from diverse sources, including dialogue corpora and meeting transcripts; (2) **Data Preprocessing**: Cleaning, normalizing, and formatting the raw texts to ensure consistency and quality; (3) **Dataset Preparation**: Curating the final dataset according to predefined filtering criteria and performing statistical analyses; (4) **Summary Generation**: Generating summaries using large language models with varying architectures and scales; (5) **Positional Bias Evaluation**: Quantifying positional bias in the generated summaries through a sentence-level semantic similarity-based computational method.

3.1 Data Collection

This section details the six English and one French conversational corpora that constitute the PoSum-Bench dataset.

3.1.1 English Conversational Corpora

ICSI (Janin et al., 2003) contains 59 multi-turn academic meetings recorded at the International Computer Science Institute, featuring natural discussions among students and researchers in a professional setting.

QMSUM (Zhong et al., 2021) consists of 1,808 query-summary pairs from 232 meetings. We selected meetings corresponding to the queries "Summarize the whole meeting" and "Summarize the meeting," ensuring no overlap with the ICSI dataset to maintain diversity.

DialogueSUM (Chen et al., 2021) provides 13,460 multi-turn dialogues collected from diverse real-world sources, capturing a wide range of speaking styles, roles, and interaction patterns.

MeetingBank (Hu et al., 2023) includes 1,366 public parliamentary committee meetings from U.S. cities, characterized by formal structures, clear speaker roles, and established turn-taking protocols.

SummEdits (Laban et al., 2023) offers conversational data from the "Sales Call" domain, generated via ChatGPT-3.5 templates and rigorously human-verified, featuring structured interactions between sales representatives and customers.

TweetSum (He et al., 2020) contains two-party dialogues from 12 major events, enriching the dataset with shorter, platform-specific conversations marked by clear speaker tags.

3.1.2 French Conversational Corpora

DECODA (Bechet et al., 2012) comprises 1,514 anonymized call-center dialogues from the Paris public transport authority (RATP), totaling approximately 74 hours of manually transcribed and annotated data, thereby enhancing the linguistic diversity of the benchmark.

3.2 Data Preprocessing

To ensure consistent quality and format across heterogeneous conversational datasets, we developed a unified preprocessing pipeline comprising two main steps: (1) Turn Segmentation: For corpora lacking clear turn boundaries (e.g., ICSI), we applied rule-based segmentation based on punctuation markers (e.g., question marks, periods, exclamation points) to delineate dialogue turns; (2)

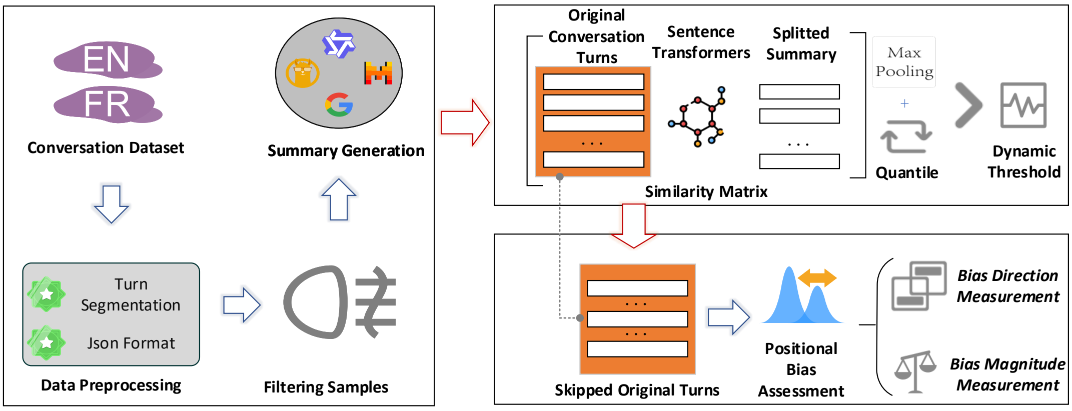

Figure 1: Pipeline for PoSum-Benchmark dataset construction and positional bias evaluation.

Standardized Formatting: All datasets were converted into a consistent JSON format, detailed in Appendix J.

3.3 Dataset Preparation

Dataset	Nb. of Instances	Avg. Words	Avg. Turns
ICSI	59	8,916	166
MeetingBank	500	10,370	310
DialogueSUM	500	133	10
QMSUM	214	8,645	46
SummEdits	500	404	36
TweetSum	500	103	5
DECODA (FR)	500	397	53
Total	2,773	EN: 1,505,323 FR: 198,500	EN: 43,893 FR: 26,500

Table 1: Summary of conversational datasets in our benchmark.

To ensure a comprehensive and balanced evaluation corpus with adequate representation from each data source, we implemented a capped sampling approach by selecting up to 500 instances per dataset. For datasets with more than 500 instances, we randomly sampled 500 examples to avoid overrepresentation, while for smaller datasets (e.g., ICSI and QMSUM), we included all available instances. This resulted in a total of 2,773 instances (2,273 in English and 500 in French), ensuring diversity in conversation types and lengths.

Table 1 summarizes the key statistics of the final dataset. It covers various conversation types—formal meetings (ICSI, MeetingBank), general dialogues (DialogueSUM), query-driven discussions (QMSUM), online interactions (SummEdits, TweetSum), and French customer service exchanges (DECODA)—spanning a wide range of lengths, from short dialogues (103 words on average in TweetSum) to extended meetings (10,370 words on average in MeetingBank), offering robust evaluation across summarization scenarios.

For further analysis, conversations were grouped by length. K-means clustering was applied to four-dimensional token count vectors (one per model tokenizer), resulting in three categories: short (667 tokens), medium (10,564 tokens), and long (20,549 tokens) conversations. This model-aware clustering accounts for tokenizer differences, enabling a more detailed positional bias analysis.

3.4 Summary Generation

To produce high-quality summaries for evaluation, we designed a controlled summary generation pipeline encompassing model selection, prompt formulation, and a multi-tiered quality control strategy.

Model Selection We utilized a diverse suite of LLMs, incorporating open-source instruction-tuned models. Our selection includes *Qwen2.5B-instruct 1.5B, 3B, 7B, 14B*, *Google Gemma3-instruct 1B, 4B*, *MistralAI-Ministral-8B-Instruct-2410, Mistral-7B-Instruct-v0.3*, *Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct*.[2] We specifically prioritized instruction-finetuned models, as prior research has highlighted instruction tuning as a key factor enabling strong zero-shot summarization capabilities (Zhang et al., 2024).

Prompt Design To ensure consistent outputs, we designed standardized English and French prompts guiding models to generate neutral, comprehen-

[2] We also used GPT-4o; its results are provided in Appendix I.

sive conversation summaries. Detailed prompt templates are provided in Appendix D.

Quality Control We implemented a rigorous quality assurance process for the generated summaries. We used `GPT-4o` to evaluate summaries on Coherence, Factuality, and Conciseness, following recent research validating LLMs as effective automatic evaluators (Zheng et al., 2023; Liu et al., 2023). On average, the summaries received a score of 4.0 on a 0-5 rating scale. Full evaluation details are provided in Appendix C.

3.5 Baselines

To ensure our metric reliably captures positional bias, we conducted a controlled evaluation using synthetic summaries with known bias patterns. In particular, we generated nine extractive summaries with predetermined positional biases by varying the proportion and location of content extracted from the source: Leading-X% summaries use the first X% of the text, Recency-X% use the last X%, and Middle-Random-X% draw from a middle segment; the proportions used were 15%, 25%, and 35%. We give the detailed methodology and statistical validation for baselines in Appendix E.

4 PoSum-Bench Methodology

4.1 Positional Bias

Existing methods for quantifying positional bias, such as n-gram distribution, are limited in granularity and fail to capture semantic alignment between source content and summaries (Wu et al., 2023; Zhu et al., 2021a; Chhabra et al., 2024). To address this, we propose a sentence-level framework that quantifies bias by measuring semantic preservation across dialogue positions. We define two core bias types—leading bias (favoring initial turns) and recency bias (favoring final turns)—and introduce a comprehensive positional bias index, offering a more interpretable, fine-grained perspective on positional behavior in conversation summarization.

Prerequisite: Identifying Skipped Sentences. To identify sentences underrepresented or omitted in the generated summary, we use a semantic similarity-based method with sentence embeddings and dynamic thresholding.

Let c_i denote the i-th sentence in the original conversation, and s_j the j-th sentence in the summary. The semantic similarity $score_{i,j}$ between c_i and s_j is computed via their embeddings:

$$score_{i,j} = \text{sim}\left(\text{emb}(c_i), \text{emb}(s_j)\right), \tag{4}$$

where $\text{emb}(\cdot)$ is the embedding function, $\text{sim}(\cdot,\cdot)$ is a similarity metric (e.g., cosine similarity).

To enable meaningful comparison, we normalize similarity scores using softmax:

$$\hat{score}_{i,j} = \frac{e^{score_{i,j}}}{\sum_{k=1}^{m} e^{score_{i,k}}}, \tag{5}$$

where m is the number of sentences in the summary, and we apply max-pooling to capture the best alignment for each conversation sentence:

$$score_i^{\max} = \max_{j=1,\ldots,m} \hat{score}_{i,j}. \tag{6}$$

To account for varying content distributions, we apply a quantile transformation to the max-scores:

$$\tilde{score}_i = Q(score_i^{\max}), \tag{7}$$

where $Q(\cdot)$ normalizes the score distribution.

Finally, a threshold based on the mean and standard deviation of the transformed scores identifies skipped sentences:

$$\tau = \mu - \alpha\sigma, \tag{8}$$

where μ and σ are the mean and standard deviation of the transformed scores and α is the sentence exclusion threshold parameter that controls the sensitivity of skipped sentence detection.[3] Sentences with $\tilde{score}_i < \tau$ are considered *skipped*, indicating underrepresentation in the summary. This τ is dynamic and robust, as it is calculated based on the statistical properties of each conversation-summary pair. Unlike fixed thresholds, our approach automatically adjusts to different dialogue types and conversation lengths.

These skipped sentences are then used to evaluate positional bias and content coverage in summaries.

4.2 Bias Measurement

Let n represent the total number of sentences in the conversation, and let k denote the number of skipped sentences. For each skipped sentence c_i, indexed from 0, its **Softmax weight** w_i is defined as:

$$w_i = \frac{e^{l_i}}{\sum_{j=0}^{k-1} e^{l_j}}, \tag{9}$$

[3] We set α = 1.0 as the default value in our experiments, balancing between detection accuracy and false positives.

where l_i denotes the word count of the skipped sentence c_i. This weight reflects the relative significance of each skipped sentence in the context, with longer sentences being assigned higher weights.

Leading Bias

To measure **Leading Bias**, we first define the **log-normalized position** for a skipped sentence c_i at position pos_i within the whole conversation. This position reflects the relative position of the sentence in the sequence, and is given by:

$$P(c_i) = \frac{\ln(\text{pos}_i + 2)}{\ln(n + 1)}, \tag{10}$$

where pos_i is the index of the skipped sentence in the full context (with pos_i starting from 0). The log-normalization ensures that positions closer to the beginning of the sequence receive a higher weight, while positions further along the context are adjusted accordingly.

For the list of skipped sentences, we similarly define the **log-normalized position** as:

$$P_{\text{skipped}}(i) = \frac{\ln(i + 2)}{\ln(k + 1)}, \tag{11}$$

where i is the index of the skipped sentence in the skipped sentence list (also starting from 0). The use of logarithmic scaling ensures a more gradual weighting of positions in the list.

Thus, the **Leading Bias** B_l is calculated as the weighted average of the ratio between the log-normalized positions of the sentences in the full context and the skipped sentence list, weighted by the sentence's Softmax weight:

$$B_l = \frac{1}{k} \sum_{i=0}^{k-1} \left(\frac{P(c_i)}{P_{\text{skipped}}(i)} \cdot w_i \right), \tag{12}$$

This formula quantifies the extent to which the model overemphasizes sentences from the beginning of the context (i.e., the early turns of the conversation), as compared to sentences within the skipped subset. A higher B_l indicates stronger leading bias.

Recency Bias

In contrast to Leading Bias, **Recency Bias** reflects the model's tendency to overweight sentences that appear near the end of the conversation. To account for this, we reverse the positions in the context, so that sentences closer to the end are given higher weight. For a skipped sentence c_i at position i in the whole conversation, the **reversed log-normalized position** is defined as:

$$P_{\text{rev}}(c_i) = \frac{\ln(n - \text{pos}_i + 1)}{\ln(n + 1)}, \tag{13}$$

where $n - \text{pos}_i + 1$ represents the reverse position of c_i in the full context. This ensures that sentences towards the end are assigned higher values, reflecting their proximity to the end of the conversation.

For the skipped sentence list, we similarly define the **reversed log-normalized position** as:

$$P_{\text{skipped}}^{\text{rev}}(i) = \frac{\ln(k - i + 1)}{\ln(k + 1)}, \tag{14}$$

This normalization works similarly to the standard log-normalized position, but reflects the reversed order for the skipped sentences.

Finally, the **Recency Bias** B_r is computed as the weighted average of the ratio between the reversed log-normalized positions of the sentences in the full context and the reversed positions in the skipped sentence list, weighted by the Softmax weights of the sentences:

$$B_r = \frac{1}{k} \sum_{i=0}^{k-1} \left(\frac{P_{\text{rev}}(c_i)}{P_{\text{skipped}}^{\text{rev}}(i)} \cdot w_i \right). \tag{15}$$

This metric quantifies the degree to which the model overemphasizes the latter parts of the conversation, with a higher B_r indicating stronger recency bias.

4.2.1 Overall Positional Bias Index

To quantify the extent and orientation of positional bias in model-generated summaries, we introduce two complementary metrics: **Bias Magnitude** and **Bias Direction**.

Bias Magnitude This metric quantifies the absolute degree of positional bias, independent of direction:

$$B_{\text{magnitude}} = |B_l - B_r| \cdot \log(e + T), \tag{16}$$

where B_l and B_r are the leading and recency bias scores, respectively, and T is the total number of tokens in the conversation. The logarithmic scaling normalizes the metric across varying conversation lengths, accounting for the increased difficulty of maintaining positional neutrality in longer contexts.

Model/Baseline	Short Text 1,594 instances		Medium Text 390 instances		Long Text 290 instances	
	Bias Mag.	Lead/Rec/Neut %	Bias Mag.	Lead/Rec/Neut %	Bias Mag.	Lead/Rec/Neut %
Llama-3.2-1B-Instruct	1.39	43.5/22.8/33.7	0.07	52.9/47.1/0.0	0.03	55.3/44.7/0.0
Llama-3.2-3B-Instruct	1.44	38.2/27.3/34.5	0.06	55.4/44.6/0.0	0.03	53.2/46.8/0.0
Gemma-3-1b-it	1.62	42.7/27.2/30.1	0.07	49.5/50.5/0.0	0.03	47.4/52.6/0.0
Gemma-3-4b-it	1.35	39.0/28.9/32.1	0.07	44.9/55.1/0.0	0.03	54.0/46.0/0.0
Qwen2.5-1.5B-Instruct	1.49	40.5/25.8/33.6	0.07	52.4/47.6/0.0	0.03	53.8/46.2/0.0
Qwen2.5-3B-Instruct	1.44	36.8/31.7/31.5	0.07	50.1/49.9/0.0	0.03	52.1/47.9/0.0
Qwen2.5-7B-Instruct	1.34	37.6/30.2/32.2	0.07	50.4/49.6/0.0	0.03	54.1/45.9/0.0
Qwen2.5-14B-Instruct	1.57	46.7/27.2/26.1	0.07	52.4/47.6/0.0	0.03	50.3/49.7/0.0
Mistral-7B-Instruct-v0.3	1.43	39.0/30.7/30.3	0.08	50.3/49.7/0.0	0.03	48.8/51.2/0.0
Mistral-8B-Instruct-2410	1.50	43.5/28.7/27.8	0.07	52.5/47.5/0.0	0.03	53.9/46.1/0.0
Leading-15%	1.48	69.4/19.4/11.2	0.35	70.4/29.6/0.0	0.14	65.2/34.8/0.0
Leading-25%	1.45	76.7/12.6/10.7	0.33	72.8/27.2/0.0	0.14	72.1/27.9/0.0
Leading-35%	1.55	82.0/7.0/11.0	0.24	78.9/21.1/0.0	0.09	81.4/18.6/0.0
Recency-15%	1.46	28.8/61.1/10.1	0.32	44.5/55.5/0.0	0.16	35.9/64.1/0.0
Recency-25%	1.47	19.1/71.2/9.7	0.31	39.6/60.4/0.0	0.18	28.6/71.4/0.0
Recency-35%	1.51	11.4/78.0/10.5	0.23	27.8/72.2/0.0	0.11	17.2/82.8/0.0
Middle-Random-15%	1.86	50.0/42.8/7.2	0.26	54.8/45.2/0.0	0.11	46.2/53.8/0.0
Middle-Random-25%	1.90	48.0/45.9/6.1	0.26	61.2/38.8/0.0	0.11	46.2/53.8/0.0
Middle-Random-35%	1.96	48.8/45.2/6.0	0.14	59.6/40.4/0.0	0.05	46.9/53.1/0.0

Table 2: Positional Bias Analysis for English instances. Text length categories determined by K-means clustering: Short (667 tokens), Medium (10,564 tokens), Long (20,549 tokens). Bias Mag. = Average Magnitude of positional bias calculated as $|B_l - B_r| \cdot \ln(e + T)$. Lead/Rec/Neut % = Percentage of instances showing leading bias vs. recency bias vs. neutral (no bias).

Bias Direction To assess the model's tendency toward earlier or later content, we define a directional indicator based on the sign of the bias difference:

$$\text{Direction} = \begin{cases} +1, & \text{if } B_l > B_r \quad \text{(leading bias)} \\ -1, & \text{if } B_l < B_r \quad \text{(recency bias)} \end{cases} \tag{17}$$

5 Experimental Results

5.1 Conversation Length Bias Distribution

We systematically analyzed positional bias in several large language models on conversation summarization tasks, focusing on preferences for content at the beginning (leading bias) or end (recency bias) across different text lengths. Results in Tables 2 and 3 reveal: (1) **Short texts show strong positional bias**: In short conversations, most models exhibit clear bias, with bias magnitudes between 1.34 and 1.62. For example, Gemma-3-1b-it shows a strong leading bias (42.7% leading vs. 27.2% recency), indicating a preference for summarizing early conversation content. (2) **Bias weakens in longer texts**: In medium and long texts, bias magnitude drops to below 0.1, and leading/recency distributions become more balanced (e.g., Mistral-7B in long texts: 48.8% leading vs. 51.2% recency), suggesting that richer content reduces positional preference.

Notably, the observed reduction in positional bias for longer conversations should be interpreted with caution. This trend may not necessarily reflect more balanced attention across the input but could instead result from challenges the model faces in summarizing longer texts. In such cases, limited coverage of original conversation content may obscure underlying positional preferences. As our study does not directly assess summary quality, we refrain from drawing conclusions about whether the reduced bias indicates genuine improvement. Future research incorporating quality evaluations will be important to clarify this relationship.

The baselines exhibit clear and expected positional biases, especially in shorter texts, aligning closely with our main experimental findings. For instance, the Leading-X% extraction increasingly favors leading positions as extraction length (X) increases, with leading bias exceeding 80% at 35% extraction in short conversations. Conversely, the Recency-X% extraction similarly displays strong recency bias patterns, particularly at higher extraction ratios. The Middle-Random-X% baselines consistently remain balanced across both languages and lengths, confirming the neutrality of content drawn from the middle sections. Tables 10, 11 in the Appendix present two examples of extractive baselines along with their individual bias scores.

Model/Baseline	Bias Mag.	Lead/Rec/Neut %
Llama-3.2-1B-Instruct	0.79	43.3/55.1/1.6
Llama-3.2-3B-Instruct	0.85	51.7/46.1/2.2
Gemma-3-1b-it	0.80	45.3/53.1/1.6
Gemma-3-4b-it	0.83	47.5/51.1/1.4
Qwen2.5-1.5B-Instruct	0.79	46.3/51.3/2.4
Qwen2.5-3B-Instruct	0.77	47.1/50.9/2.0
Qwen2.5-7B-Instruct	0.74	43.9/54.1/2.0
Qwen2.5-14B-Instruct	0.77	44.1/53.7/2.2
Mistral-7B-Instruct-v0.3	0.77	45.4/52.0/2.6
Mistral-8B-Instruct-2410	0.79	46.7/51.9/1.4
Leading-15%	1.26	59.7/37.9/2.4
Leading-25%	1.35	70.7/28.1/1.2
Leading-35%	1.30	81.8/17.6/0.6
Recency-15%	1.30	35.1/63.1/1.8
Recency-25%	1.23	25.5/72.7/1.8
Recency-35%	1.11	25.9/73.1/1.0
Middle-Random-15%	1.51	43.9/55.3/0.8
Middle-Random-25%	1.38	45.5/52.9/1.6
Middle-Random-35%	1.18	49.9/49.3/0.8

Table 3: Positional Bias Analysis for French. French results only cover short text instances (499 in total). One medium-length instance was excluded from our study.

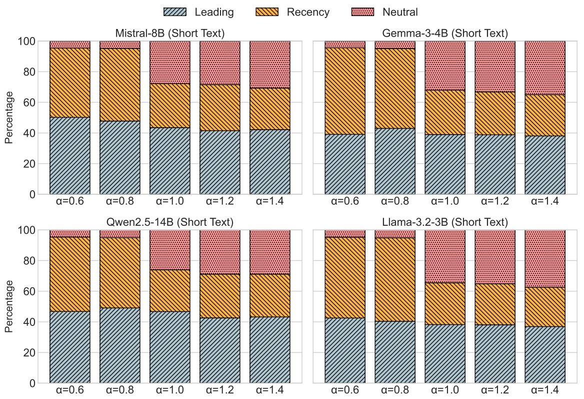

Figure 2: Distribution of positional bias types (Leading, Recency, Neutral) across varying α thresholds for short texts. Each subplot corresponds to a different language model.

5.2 Language Bias Distribution

We analyzed positional bias in conversation summarization across English and French short-text scenarios. Overall, both languages show similar trends, with models favoring content at the beginning or end of the conversation, though differences in bias strength and consistency were observed: (1) Consistent bias across languages: Most models show clear positional preferences in both languages. For example, Qwen2.5-14B shows 46.7% leading / 27.2% recency in English and 44.1% leading / 53.7% recency in French. (2) Lower bias magnitude and fewer neutral cases in French: Bias magnitude for Mistral-7B-Instruct-v0.3 drops from 1.43 (EN) to 0.77 (FR), and neutral instances decrease from 30.3% (EN) to 2.6% (FR), indicating French summaries tend to select start/end content more decisively. (3) More pronounced recency bias for French: most models show recency bias in more than 50% of instances, whereas the results for English are more evenly distributed across three bias types.

5.3 Impact of the Sentence Exclusion Threshold

This experiment examines how varying the sentence exclusion threshold, governed by α, influences bias distributions across four models: Mistral-8B, Gemma-3-4B, Qwen2.5-14B, and Llama-3.2-3B. As shown in Figure 2, each model exhibits different bias patterns even under the same α value. For example, at $\alpha = 0.6$, Mistral-8B favors both Leading and Recency, whereas Gemma-3-4B leans more heavily toward Recency. As α increases, all models undergo notable distribution shifts, particularly around $\alpha = 1.0$, where Leading and Recency proportions converge and Neutral increases significantly—indicating a move toward greater neutrality. This trend aligns with the role of α as a sentence inclusion threshold: higher values lower the exclusion threshold, resulting in more sentences being classified as Neutral.

6 Mitigation Discussion

For future work, we propose two strategies to mitigate positional biases: (1) Prompt engineering, which involves crafting the input prompt to encourage the model to consider content from all parts of the conversation more evenly. (2) Objective-oriented reinforcement learning, where the model is fine-tuned with a bias-aware reward function that explicitly penalizes excessive focus on either early or late content. Nonetheless, they represent promising avenues for future research to reduce positional bias and enhance the overall balance and fidelity of conversational summaries.

7 Conclusion

We introduced PoSum-Bench, a benchmark for evaluating positional bias in conversational summarization across English and French datasets. Our novel sentence-level semantic similarity metric quantifies the direction and magnitude of positional bias, enabling cross-lingual, reference-free analysis of summaries. Through extensive experiments, we demonstrated that PoSum-Bench effectively captures positional bias patterns, revealing variations with conversation length and context.

PoSum-Bench offers a standardized framework for assessing and mitigating bias, providing a foundation for developing more balanced and unbiased conversational summarization models.

Limitations

Reflecting on our methodology and experiments, we identify the following limitations: (1) we used basic prompts for summary generation and did not explore their impact on summary bias; (2) due to computational constraints, we excluded samples with turns exceeding 500 rounds; (3) we defined long, medium, and short conversations in a simplistic manner and could refine this categorization using frequency-based approaches.

Ethics Statement

All datasets used in this study are publicly available, and our PoSum-Bench dataset is openly accessible. Datasets are fully anonymized, with no personal information processed. For transparency, we note that Claude 3.7 was used only for text polishing in manuscript preparation. We acknowledge potential risks in our approach, including the possibility that optimization for positional bias metrics alone might compromise other important qualities of summarization systems, and that our findings may not generalize equally across all languages and cultural contexts given our dataset limitations. This benchmark aims to measure and evaluate positional bias in conversational summarization, contributing to the development of more fair and representative NLP systems by making these resources widely available to the research community.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Bèze, Renato De Mori, and Eric Arbillot. 2012. DECODA: a call-centre human-human spoken conversation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1343–1347, Istanbul, Turkey. European Language Resources Association (ELRA).

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 1–11, Mexico City, Mexico. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5453–5460. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Ruifang He, Liangliang Zhao, and Huanyu Liu. 2020. TWEETSUM: Event oriented social summarization dataset. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5731–5736, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive meeting summarization: A survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Han Wu, Mingjie Zhan, Haochen Tan, Zhaohui Hou, Ding Liang, and Linqi Song. 2023. VCSUM: A versatile Chinese meeting summarization dataset. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6065–6079, Toronto, Canada. Association for Computational Linguistics.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 948–954, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. NarraSum: A large-scale dataset for abstractive narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 182–197, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021a. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021b. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1462–1471, New York, NY, USA. Association for Computing Machinery.

A Computational Cost

Our experiments required minimal computational resources for data processing and bias calculation, which run efficiently on CPU. For summary generation across 10 LLMs, we utilized pairs of NVIDIA A100 40GB GPUs (2 GPUs per model family) running in parallel. The entire summary generation process took approximately 16 hours of wall-clock time, with the longest individual model family requiring 15.5 hours. Sentence embedding generation was performed on a single RTX 3090 GPU, completing in approximately 30 minutes. In total, our experiments required approximately 100 GPU-hours, a reasonable computational investment that makes our benchmark accessible to researchers who can leverage our pre-computed results.

B Sentence Transformers

To calculate the semantic similarity between conversations and their summaries, we employed pre-trained sentence transformer models to represent sentences in a dense vector space (Reimers and Gurevych, 2019). Specifically, we utilized `sentence-transformers/all-MiniLM-L6-v2` (Wang et al., 2021) from the HuggingFace Transformers library for its optimal balance between computational efficiency and performance quality.

C Quality Criteria

Coverage

- **Fully Satisfied:** The summary comprehensively includes all main ideas and key details from the original text without significant omissions.
- **Partially Satisfied:** The summary includes the main information but omits some important details or secondary points, leading to an incomplete representation.
- **Not Satisfied:** The summary fails to capture the main information from the original text, omitting substantial key details and resulting in severely insufficient coverage.

Factuality

- **Fully Satisfied:** All statements in the summary are consistent with the original text, providing accurate information without factual errors or distortions.
- **Partially Satisfied:** Most information in the summary is correct, but there are minor factual discrepancies or slightly imprecise descriptions affecting overall accuracy.
- **Not Satisfied:** The summary contains numerous factual errors or false information, significantly deviating from the original content and impairing the reader's understanding.

Conciseness

- **Fully Satisfied:** The summary is highly concise, retaining only necessary information, avoiding redundancy and verbosity, and presenting content clearly and succinctly.
- **Partially Satisfied:** The summary conveys the main information but includes some overly lengthy parts or unnecessary details, diminishing the effectiveness of information delivery.
- **Not Satisfied:** The summary is verbose and repetitive, containing excessive unnecessary content and failing to effectively distill and convey the key points of the original text.

GPT-4o Evaluation Prompt We used the following prompt in Table 4 to evaluate summaries with GPT-4o, instructing the model to assess each summary on a scale of 0-5 across three dimensions: coverage, factuality, and conciseness.

The evaluations were performed by prompting GPT-4o with the conversation transcript, the generated summary, and the above evaluation criteria. For each summary, GPT-4o assigned scores for coverage, factuality, and conciseness on a scale from 0 to 5, where higher scores indicate better performance. We then calculated the average scores across all evaluated instances for each model and metric; the results are shown in Table 5.

D Prompts for Generation

To ensure the uniformity and correctness, for English and French conversation data, we adopt the following prompts in Table 6 across all the models.

E Positional Bias Methodology Validation

To rigorously validate our positional bias metrics, we conducted a series of controlled experiments with artificially constructed extractive summaries exhibiting known bias patterns. This approach allowed us to verify that our metrics correctly identify and quantify different types of positional bias before applying them to LLM-generated summaries. Tables 10 and 11 show two examples from the dataset of controlled experiments.

E.1 Experimental Design

We implemented an extractive pipeline that creates three distinct types of biased extractions:

- **Leading Extractions:** Selecting the first N% of sentences from each conversation, creating summaries with expected leading bias.
- **Recency Extractions:** Selecting the last N% of sentences, creating summaries with expected recency bias.
- **Middle Random Extractions:** Randomly selecting N% of sentences from the middle portions (excluding first and last sentences), creating more balanced summaries.

For each extraction type, we tested three extraction ratios (15%, 25%, and 35%) to assess our metrics' sensitivity to bias magnitude. We randomly sampled 50% of our dataset for leading extraction and the remaining 50% for recency or middle-random extraction, ensuring statistical independence between samples while maintaining computational efficiency. Importantly, we ensured that different extraction methods were applied to non-overlapping subsets of the data, maintaining the independence assumption required for the Mann-Whitney U test used in our statistical analysis.

E.2 Results and Analysis

E.2.1 Extractive Summary Analysis

Tables 2, 3 present our positional bias analysis across different extraction strategies and text lengths. The controlled experiment results (lower sections of both English and French data) clearly demonstrate the effectiveness of our positional bias detection methodology.

For leading extractions, we observe a consistent pattern of high leading bias classifications that increases with extraction ratio. In English short texts, leading bias classifications increase from 69.4% (at 15% extraction) to 82.0% (at 35% extraction), while recency bias classifications correspondingly decrease from 19.4% to 7.0%. Similar patterns appear in medium and long texts, with leading bias classifications reaching 78.9% and 81.4% respectively at 35% extraction. The French dataset exhibits comparable trends, with leading bias classifications increasing from 59.7% (at 15% extraction) to 81.8% (at 35% extraction).

Conversely, recency extractions show the expected opposite pattern. In English short texts, recency bias classifications increase from 61.1% (at 15% extraction) to 78.0% (at 35% extraction), while leading bias classifications decrease from 28.8% to 11.4%. The medium and long text categories display similar patterns, with recency bias classifications reaching 72.2% and 82.8% respectively at 35% extraction. The French dataset follows similar trends, with recency bias classifications increasing from 63.1% (at 15% extraction) to 73.1% (at 35% extraction).

Middle-random extractions, which serve as our control group, demonstrate more balanced classifications between leading and recency bias. For English short texts, the leading-to-recency ratios remain relatively stable across extraction ratios (50.0%/42.8% at 15%, 48.8%/45.2% at 35%). This balance, though slightly favoring leading bias in some instances, particularly in medium texts, confirms that our method does not systematically favor either bias type when content is more evenly dis-

Prompt for Summary Evaluation

System Prompt: Given the following criteria, evaluate the summary and provide a score for each category on a scale of 0–5: **Categories:** Coverage, Factuality, Conciseness

Coverage

- **5: Excellent** — All main ideas and key details included.
- **4: Very Good** — Almost all main ideas and most details.
- **3: Good** — Most main ideas and several details.
- **2: Fair** — Some main ideas, many details missing.
- **1: Poor** — Few ideas and minimal detail.
- **0: Unsatisfactory** — Substantial omissions.

Factuality

- **5: Excellent** — Completely accurate.
- **4: Very Good** — Minor inconsistencies.
- **3: Good** — A few minor errors.
- **2: Fair** — Several inaccuracies.
- **1: Poor** — Many inaccuracies.
- **0: Unsatisfactory** — Numerous factual errors.

Conciseness

- **5: Excellent** — No redundancy.
- **4: Very Good** — Minimal unnecessary content.
- **3: Good** — Some unnecessary content.
- **2: Fair** — Noticeable redundancy.
- **1: Poor** — Excessively verbose.
- **0: Unsatisfactory** — Extremely verbose and repetitive.

Evaluation Format (Markdown):

- **Coverage:** [score] - [brief justification]
- **Factuality:** [score] - [brief justification]
- **Conciseness:** [score] - [brief justification]

Ensure your evaluation is objective and based solely on the provided criteria.

User Prompt: Original Conversations: {conversation}, Corresponding Summary: {summary}

Table 4: Prompt Used for Quality Evaluation

Model	Coverage	Factuality	Conciseness	Overall
Qwen2.5-1.5B-Instruct	2.95	4.29	3.62	3.62
Qwen2.5-14B-Instruct	3.99	4.89	4.44	4.44
Qwen2.5-3B-Instruct	3.70	4.77	4.15	4.21
Qwen2.5-7B-Instruct	3.91	4.86	4.31	4.36
Gemma-3-1b-it	2.12	3.78	4.01	3.30
Gemma-3-4b-it	3.87	4.93	4.64	4.48
Llama-3.2-1B-Instruct	2.95	3.76	3.24	3.32
Llama-3.2-3B-Instruct	3.57	4.65	3.94	4.12*
Mistral-8B-Instruct-2410	3.51	4.64	4.02	3.95*
Mistral-7B-Instruct-v0.3	3.46	4.87	4.24	4.16*
Average	**3.40**	**4.54**	**4.06**	**4.00**

Table 5: GPT-4o scores across 10 Models (evaluated on a sample of 100 instances). Each metric is measured on a scale from 0 to 5, where higher scores indicate better performance. * Some overall scores were approximated based on a subsample of the data due to budget restrictions.

Table 6: Prompts Used for Summary Generation in English and French.

Bilingual Prompt Format for Summary Generation

EN: {"system prompt": You are a professional summary writer, and you are asked to write a summary of the following text. Please only return the generated summary.
"User Input": Now, please summarize {*conversations*}, only return summarized answer in plain text format, starting with "SUMMARY:"}

FR: {"system prompt": Vous êtes un rédacteur professionnel de résumés, et on vous demande d'écrire un résumé du texte suivant. Veuillez ne renvoyer que le résumé généré.
"User Input": Maintenant, veuillez résumer {*conversations*}, ne renvoyer que la réponse résumée au format texte brut, en commençant par "RÉSUMÉ:"}

tributed. In the French dataset, middle-random extractions show a similar balanced pattern, with classifications hovering near 50% for both leading and recency bias at higher extraction rates.

Interestingly, the bias magnitude (Bias Mag.) generally decreases as text length increases, which is expected as longer texts provide more context and potentially dilute position-specific information. For instance, in English leading extractions at 35%, the bias magnitude decreases from 1.55 (short texts) to 0.24 (medium texts) to 0.09 (long texts).

This controlled experiment validates that our positional bias metrics successfully identify artificial biases introduced through position-specific extraction, with classification distributions clearly matching the expected patterns across different extraction strategies, ratios, and languages.

Figure 4 further illustrates these findings by visualizing the distribution of bias classifications across our different datasets, corroborating the tabular results and demonstrating the robustness of our bias detection methodology across diverse conversational contexts.

E.2.2 Statistical Significance Testing

We applied two-tailed Mann-Whitney U tests (appropriate for non-normally distributed data) to analyze differences between extraction types. Our analysis focused on comparing how different extraction methods perform at capturing the same content positions, rather than comparing each method's performance in its specialized domain. Table 7 summarizes the key findings:

- **Leading vs Recency (Leading Score):** Leading extractions significantly outperformed recency extractions at capturing leading content across all extraction ratios ($p < 10^{-22}$), with significance increasing at higher extraction ratios.
- **Leading vs Recency (Recency Score):** Sim-

ilarly, recency extractions significantly outperformed leading extractions at capturing recency content (p < 10^{-18}), confirming the effectiveness of our position-sensitive metrics.

- **Bias Magnitude Comparisons:** Middle random extractions exhibited dramatically smaller bias magnitudes compared to both leading and recency extractions (p < 10^{-149}), indicating their more balanced content representation.

- **Extraction Ratio Sensitivity:** Statistical significance strengthened with increasing extraction ratios (from 15% to 35%), supporting our method's ability to detect different degrees of positional bias.

These results validate that our positional bias metrics effectively distinguish between different bias patterns: leading extractions produce significantly higher leading bias scores (B_l), recency extractions produce higher recency bias scores (B_r), and middle random extractions show more balanced metrics with significantly smaller bias magnitude. Importantly, the extremely low p-values (often below 10^{-100}) demonstrate the robust discriminative power of our metrics across different extraction conditions.

F Bias Direction Scores Distribution Across Different Sub-Datasets

The results of the experiment are presented in Figure 3. The distribution of bias direction scores (Leading Score minus Recency Score) across datasets highlights distinct positional biases depending on conversational context. For longer, formal meetings (ICSI, MeetingBank), distributions cluster symmetrically around zero, indicating balanced coverage across conversations. Conversely, shorter informal dialogues (TweetSum, SummEdits, DialogueSUM) display marked positive skewness, reflecting a pronounced leading bias—summaries disproportionately favor early content. In contrast, customer service interactions (DECODA) exhibit a slight negative skew, suggesting mild recency bias likely due to conversation resolutions typically occurring near the end. These findings confirm that positional biases in summarization vary systematically by dialogue length, language, and conversational context.

G Bias Direction Scores Distribution in Controlled Experiments

The experimental results for the controlled setup are shown in Figure 4. Controlled experiments, in which summaries were artificially constrained to specific conversation segments, clearly demonstrate the sensitivity of our bias metric. Summaries derived exclusively from the initial 15% (Leading-15%) show strong positive skewness, while those from the final 35% (Recency-35%) are significantly negatively skewed, confirming the metric's accurate detection of imposed biases. Summaries from middle sections (Middle-25%) produce balanced distributions around zero, affirming the neutrality of the metric when no positional bias is enforced. These controlled scenarios validate the robustness and reliability of our positional bias quantification approach.

H Attention Analysis

Based on the visualizations presented in Figure 5, we selected representative examples from various models and analyzed their multi-head attention maps. Notably, the attention heatmaps associated with forward-leaning summaries exhibit clear distinctions from those corresponding to backward-leaning summaries. Although attention heatmaps alone may not provide conclusive evidence of a model's tendency toward leading or recency bias, they offer valuable auxiliary insights that can support such interpretations.

I GPT4o Summary Bias Result

GPT-4o is one of the flagship LLM models which has been instruction-fintuned. We utlized API service that OpenAI provided and the version was 2024-11-20 to provide a more comprehensive comparison between closed-source and open-sourced modes.

I.1 Positional Bias Analysis for GPT-4o

Table 8 presents the positional bias analysis for GPT-4o across different languages and text lengths. Several notable findings emerge when comparing these results with the open-source models in Table 2.

For short English conversations, GPT-4o exhibits a strong leading bias (61.2% leading vs. 29.2% recency), with an average bias magnitude of 1.49. This leading bias is notably stronger than most open-source models we evaluated, which typically

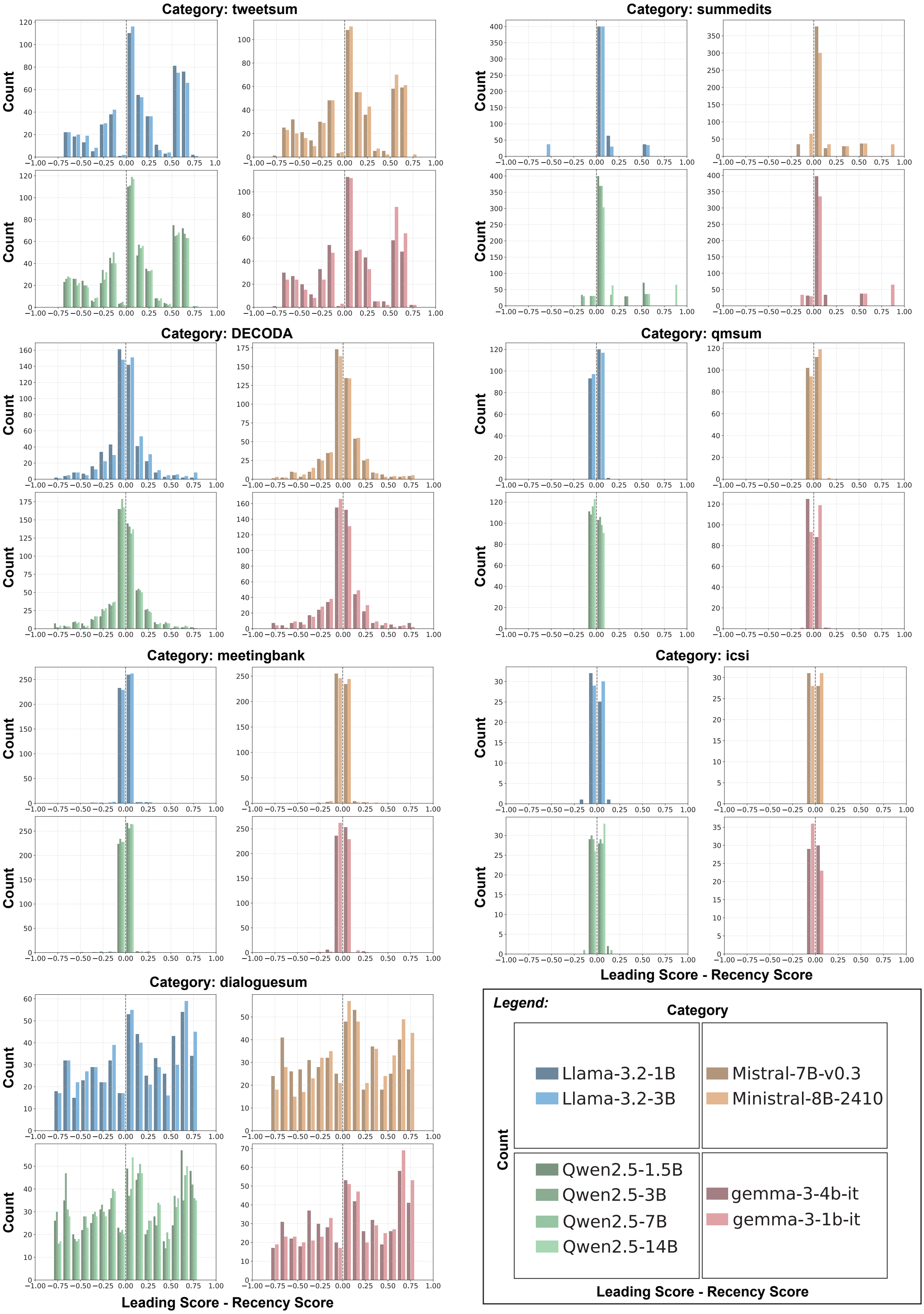

Figure 3: Distribution of bias direction scores (Leading Score - Recency Score) across various models and subdatasets.

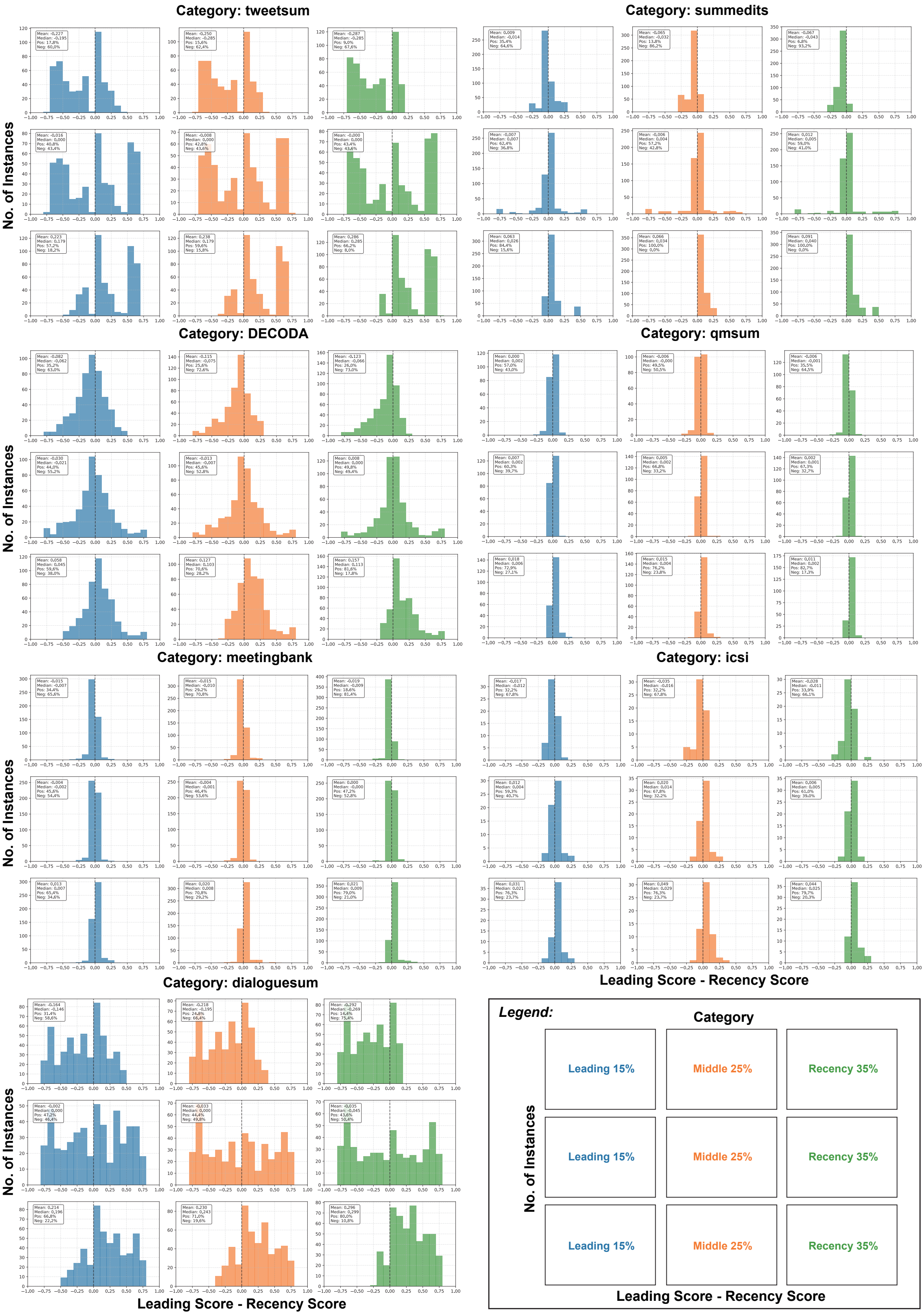

Figure 4: Distribution of bias direction scores (Leading Score - Recency Score) across different extraction methods and sub-datasets.

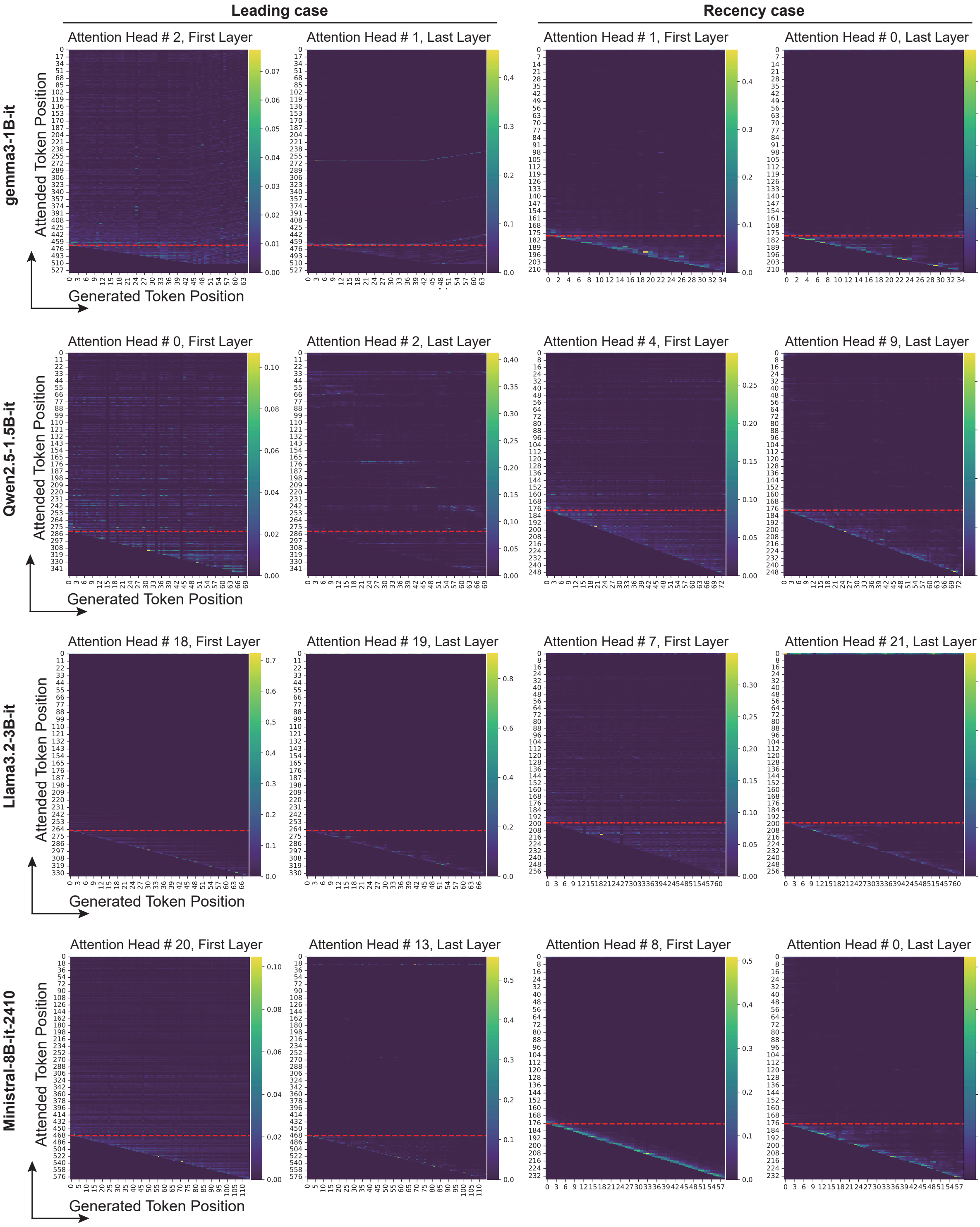

Figure 5: Attention Analysis Across Different Model Family

show more balanced distributions between leading and recency bias for short texts. For instance, Qwen2.5-7B-Instruct has a 37.6%/30.2%/32.2% lead/recency/neutral distribution, and Gemma-3-4b-it shows a 39.0%/28.9%/32.1% distribution. The bias magnitude for GPT-4o (1.49) is comparable to the average of open-source models but its leading bias percentage is substantially higher.

Interestingly, for medium and long English texts, GPT-4o demonstrates a more balanced positional bias profile. In medium texts, it shows a slight leading preference (52.4% vs. 47.6%), while in long texts, it actually exhibits a slight recency preference (46.6% vs. 53.4%). This pattern is consistent with most open-source models we tested, suggesting that longer contexts generally lead to more balanced positional information utilization across different model architectures and training paradigms.

For French texts, GPT-4o shows a relatively balanced distribution with a slight recency preference (46.7% leading vs. 50.7% recency), which differs from some open-source models like Llama-3.2-3B-Instruct that exhibit a stronger leading bias (51.7%/46.1%/2.2%). The bias magnitude for GPT-4o in French (0.70) is slightly lower than most open-source models, potentially indicating more uniform information utilization across conversation positions.

These results suggest that despite GPT-4o's advanced capabilities, it still exhibits significant positional biases, particularly for shorter conversations. The stronger leading bias in short English texts indicates that even state-of-the-art closed-source models tend to prioritize information from the beginning of conversations when generating summaries. This finding reinforces our broader observation that positional bias remains an important consideration across the entire spectrum of current LLMs, regardless of their source or sophistication level.

J Unified Json Format

Figure 7 illustrates the standardized format we adopted for storing the processed data, which not only facilitates our experiments but also serves as a reference for other researchers.

K Positional Bias Direction Across Different Models

Figure 6 presents heatmaps illustrating the positional bias direction across various language models. We applied Ward clustering based on Euclidean distance to group models (rows) exhibiting similar positional bias patterns, revealing several distinct clusters that merit discussion.

K.1 Overall Observations (Full Dataset, n = 2,773)

In Figure 6 Panel A, we observe that models are grouped into several major clusters based on their positional bias behavior, with models from various families distributed across different clusters:

- **Upper cluster:** Models consistently demonstrate positive bias values (reddish regions) across different token positions, with the bias being relatively consistent across sequence lengths. This suggests these models, regardless of family, tend to prioritize information at the beginning of sequences (leading bias) across various context lengths. While the bias remains positive, there appears to be slight variation in intensity across token positions, suggesting subtle changes in bias strength as sequence length changes.

- **Middle cluster:** These models exhibit a more nuanced pattern: neutral to slightly negative bias at higher token positions (3-5, representing > 1,000 tokens). This possibly indicates a crucial mechanism that changes which leads to shifting from leading to recency bias as sequence length increases, a phenomenon that warrants further investigation.

- **Lower cluster:** The distinctive feature of this cluster is the strong negative bias (deep blue) at lower token positions (2-3, representing 100-1,000 tokens), while maintaining relatively neutral or slightly positive bias at higher positions. This suggests these models significantly favor information toward the end of sequences when processing shorter inputs, but this recency bias diminishes with longer sequences.

The clustering reveals that positional bias behavior doesn't strictly align with model families but rather represents fundamental differences in how models process sequential information across different context lengths. The token number axis (horizontal) shows how bias patterns evolve from shorter to longer sequences (as token values are log10-transformed, with 2 representing 100 tokens

Table 7: Controlled Extraction Experiment Results - Mann-Whitney U Test Comparison

Extraction Type	**Percentage**	**Leading Score**	**Recency Score**	**Bias Magnitude**	**Statistical Significance (p-values)**	
					vs Leading	**vs Recency**
Leading	15%	0.442	0.346	0.096	—	1.35×10^{-22}***
	25%	0.432	0.293	0.139	—	6.89×10^{-37}***
	35%	0.440	0.278	0.162	—	1.39×10^{-45}***
Recency	15%	0.316	0.396	0.080	1.06×10^{-18}***	—
	25%	0.299	0.405	0.105	2.67×10^{-33}***	—
	35%	0.289	0.412	0.123	1.21×10^{-46}***	—
Middle Random	15%	0.375	0.353	0.022	6.05×10^{-160}***	3.99×10^{-149}***
	25%	0.358	0.344	0.014	1.61×10^{-172}***	2.27×10^{-166}***
	35%	0.361	0.342	0.018	2.25×10^{-183}***	4.58×10^{-181}***

$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: The table organizes data by extraction method (Leading, Recency, Middle Random).
"vs Leading" p-values compare Leading Score between the given method and Leading extraction.
"vs Recency" p-values compare Recency Score between the given method and Recency extraction.
For Middle Random vs Leading/Recency comparisons, p-values represent bias magnitude differences.
All comparisons show statistically significant differences, confirming that extraction position significantly affects content bias.

Table 8: Positional Bias Analysis for GPT-4o Across Languages and Text Lengths

Language	**Text Length Category**	**Bias Mag.**	**Lead/Rec/Neut %**	**Max Magnitude**	**Sample Count**
English	Short (∼667 tokens)	1.49	61.2/29.2/9.5	4.94	1,594
	Medium (∼10,564 tokens)	0.07	52.4/47.6/0.0	0.79	389
	Long (∼20,549 tokens)	0.02	46.6/53.4/0.0	0.26	290
French	Short (∼667 tokens)	0.70	46.7/50.7/2.6	4.37	499

Note: Bias Mag. = Average Magnitude of positional bias calculated as $|B_l - B_r| \cdot \ln(e + T)$.
Lead/Rec/Neut % = Percentage of samples showing leading bias vs. recency bias vs. neutral (no bias).
Max Magnitude represents the highest bias magnitude observed in each category.
Text length categories are consistent with those used for open-source models in our previous analysis, determined by K-means clustering of token counts.

and 5 representing 100,000 tokens), with most models showing some degree of sequence-length-dependent bias behavior.

K.2 Comparison Between English (n = 2,273) and French (n = 500) Datasets

The comparison between Panels B and C in Figure 6 reveals the influence of language on positional bias:

- **Consistency:** Certain models exhibit similar positional bias patterns across both English and French datasets, indicating that their positional bias behavior remains relatively stable across linguistic contexts.

- **Variability:** Some models display significantly different positional bias patterns between language datasets. In the French dataset (Panel C), the positional bias structures of certain models become more pronounced, and the clustering appears more distinct.

- **French-specific characteristics:** Panel C demonstrates a more evident hierarchical structure in the French dataset, with more pronounced differences between models, particularly in regions with lower Token values (shorter sequences).

K.3 Model Family Characteristics

Our analysis reveals distinct patterns across the four model families:

- **Qwen family:** Qwen models generally demonstrate consistent positional bias patterns, with their bias behavior changing systematically across different sequence lengths.

- **Llama 3.2 family:** These models exhibit diverse positional bias behaviors, distributed across different cluster groups, indicating sig-

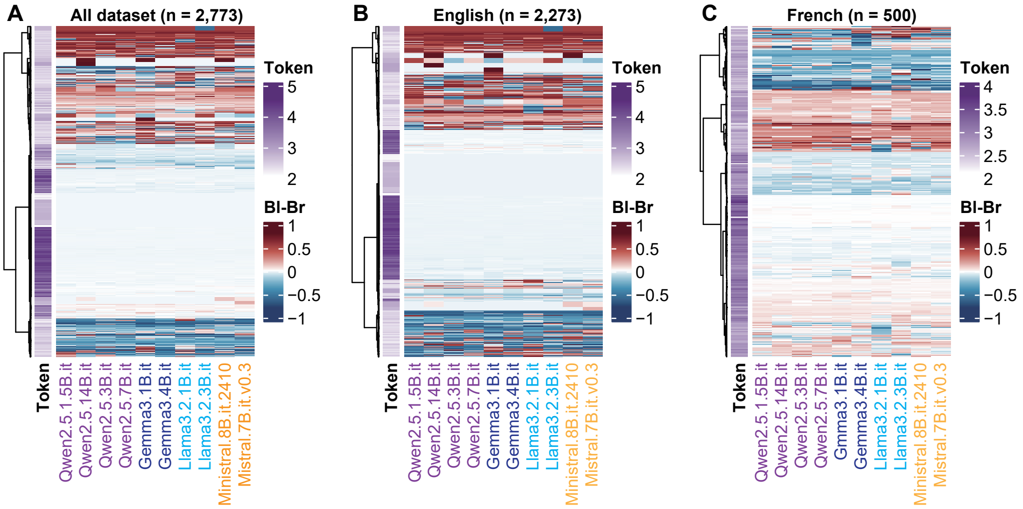

Figure 6: Heatmaps of positional bias direction across models on the PoSum dataset. Panel A: full dataset ($n = 2,773$); Panel B: English subset ($n = 2,273$); Panel C: French subset ($n = 500$). Color indicates Bl–Br scores (Leading Bias minus Recency Bias), with red showing a leading bias (focus on sequence start) and blue indicating a recency bias (focus on sequence end). Token counts are $\log_{10}$-transformed (range: 100 to 100,000 tokens). Rows are clustered using Euclidean distance with Ward's method.

nificant within-family variation in how they process positional information.

- **Mistral family:** Mistral models show considerable variability in positional bias, with different bias patterns across sequence lengths.
- **Gemma family:** Gemma models typically cluster in similar regions, suggesting that this family may have more consistent position-processing mechanisms across different sequence lengths.

These findings suggest that positional bias is a complex product of model architecture, training data, and inference processes. Different model families exhibit distinct clustering characteristics, and these bias properties show a degree of language dependency. Most importantly, the bias direction systematically varies with sequence length (token position), suggesting that models employ different strategies for information prioritization depending on the context length they process.

Table 9: Impact of Threshold Parameter (α) on Positional Bias Distribution Across Text Lengths

Model	Length	Positional Bias Distribution (Leading%/Recency%/Neutral%)										
		$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1.0$	$\alpha = 1.1$	$\alpha = 1.2$	$\alpha = 1.3$	$\alpha = 1.4$	$\alpha = 1.5$
English												
Mistral-7B	Short	44.4/51.7/3.9	43.9/52.1/4.1	43.7/52.3/4.1	41.2/54.3/4.5	41.1/54.4/4.5	39.0/30.7/30.3	38.9/29.9/31.2	39.1/29.7/31.2	37.7/28.7/33.7	37.2/29.0/33.8	31.2/24.1/44.8
	Medium	56.3/43.8/0.0	55.4/44.6/0.0	52.4/47.6/0.0	51.9/48.1/0.0	53.5/46.5/0.0	50.5/49.5/0.0	49.7/50.3/0.0	50.0/50.0/0.0	49.5/50.5/0.0	48.6/51.4/0.0	48.9/51.1/0.0
	Long	48.5/51.5/0.0	47.3/52.7/0.0	49.4/50.6/0.0	49.4/50.6/0.0	47.3/52.7/0.0	48.5/51.5/0.0	46.1/53.9/0.0	46.7/53.3/0.0	47.3/52.7/0.0	50.6/49.4/0.0	50.0/50.0/0.0
Mistral-8B	Short	50.2/45.6/4.3	50.2/45.3/4.5	50.0/45.5/4.5	47.8/47.4/4.8	45.8/49.3/4.8	43.5/28.7/27.8	41.4/30.3/28.2	41.5/30.1/28.4	42.6/26.7/30.7	42.2/27.1/30.8	35.3/23.2/41.5
	Medium	55.3/44.7/0.0	55.0/45.0/0.0	53.7/46.3/0.0	51.9/48.1/0.0	55.0/45.0/0.0	52.5/47.5/0.0	55.6/44.4/0.0	56.6/43.4/0.0	53.2/46.8/0.0	54.0/46.0/0.0	52.7/47.3/0.0
	Long	52.2/47.8/0.0	52.9/47.1/0.0	53.2/46.8/0.0	50.9/49.1/0.0	52.2/47.8/0.0	53.9/46.1/0.0	52.6/47.4/0.0	54.9/45.1/0.0	53.6/46.4/0.0	53.6/46.4/0.0	49.8/50.2/0.0
Gemma-3-1B	Short	46.2/49.3/4.5	48.0/47.4/4.6	46.0/49.4/4.5	45.9/49.2/4.8	45.2/49.9/4.8	42.7/27.2/30.1	40.6/26.4/33.0	42.5/26.5/31.0	40.8/25.8/33.4	40.7/25.8/33.6	33.5/22.1/44.4
	Medium	48.0/52.0/0.0	48.7/51.3/0.0	45.7/54.3/0.0	50.5/49.5/0.0	45.9/54.1/0.0	49.5/50.5/0.0	49.0/51.0/0.0	48.7/51.3/0.0	46.7/53.3/0.0	46.9/53.1/0.0	45.4/54.6/0.0
	Long	40.8/59.2/0.0	41.1/58.9/0.0	40.4/59.6/0.0	43.6/56.4/0.0	46.3/53.7/0.0	47.4/52.6/0.0	48.8/51.2/0.0	51.6/48.4/0.0	49.8/50.2/0.0	48.4/51.6/0.0	49.1/50.9/0.0
Gemma-3-4B	Short	39.4/56.2/4.4	39.1/56.5/4.4	41.5/54.1/4.4	43.0/52.1/4.9	43.2/51.9/4.9	39.0/28.9/32.1	38.8/28.4/32.7	38.8/28.0/33.1	38.2/26.9/34.9	38.1/26.9/35.1	31.6/22.8/45.7
	Medium	52.3/47.7/0.0	54.1/45.9/0.0	52.6/47.4/0.0	50.8/49.2/0.0	50.0/50.0/0.0	44.9/55.1/0.0	48.2/51.8/0.0	49.2/50.8/0.0	50.8/49.2/0.0	48.7/51.3/0.0	49.2/50.8/0.0
	Long	47.0/53.0/0.0	47.4/52.6/0.0	49.8/50.2/0.0	50.9/49.1/0.0	51.6/48.4/0.0	54.0/46.0/0.0	52.6/47.4/0.0	53.7/46.3/0.0	54.0/46.0/0.0	53.7/46.3/0.0	54.4/45.6/0.0
Qwen2.5-1.5B	Short	41.7/54.6/3.7	40.9/55.2/3.9	43.3/52.8/4.0	42.8/52.8/4.3	43.3/52.4/4.3	40.5/25.8/33.6	40.6/25.2/34.2	40.8/25.0/34.2	39.5/23.8/36.8	39.1/24.0/37.0	32.4/19.9/47.7
	Medium	59.6/40.4/0.0	56.3/43.7/0.0	56.0/44.0/0.0	53.5/46.5/0.0	53.2/46.8/0.0	52.4/47.6/0.0	52.4/47.6/0.0	51.4/48.6/0.0	50.1/49.9/0.0	48.3/51.7/0.0	47.0/53.0/0.0
	Long	51.7/48.3/0.0	51.7/48.3/0.0	53.4/46.6/0.0	52.8/47.2/0.0	54.8/45.2/0.0	53.8/46.2/0.0	57.9/42.1/0.0	55.9/44.1/0.0	57.9/42.1/0.0	53.1/46.9/0.0	52.1/47.9/0.0
Qwen2.5-3B	Short	42.2/54.0/3.8	41.5/54.7/3.8	41.1/55.1/3.8	39.0/56.9/4.1	38.6/57.3/4.1	36.8/31.7/31.5	39.0/29.2/31.8	38.8/29.2/31.9	35.8/29.8/34.4	35.1/30.4/34.5	28.3/26.5/45.2
	Medium	55.3/44.7/0.0	55.3/44.7/0.0	59.4/40.6/0.0	54.2/45.8/0.0	53.7/46.3/0.0	52.9/47.1/0.0	50.4/49.6/0.0	50.4/49.6/0.0	50.9/49.1/0.0	49.9/50.1/0.0	51.4/48.6/0.0
	Long	51.4/48.6/0.0	50.3/49.7/0.0	52.4/47.6/0.0	53.4/46.6/0.0	53.8/46.2/0.0	52.1/47.9/0.0	55.5/44.5/0.0	52.8/47.2/0.0	54.5/45.5/0.0	54.1/45.9/0.0	50.7/49.3/0.0
Qwen2.5-7B	Short	44.3/51.1/4.6	43.9/51.3/4.6	40.0/55.4/4.6	41.7/53.4/5.0	41.3/53.8/5.0	37.6/30.2/32.2	39.6/27.8/32.6	40.3/27.0/32.7	38.8/26.0/35.1	38.6/26.2/35.2	31.2/22.8/46.0
	Medium	54.5/45.5/0.0	52.4/47.6/0.0	51.9/48.1/0.0	49.9/50.1/0.0	50.4/49.6/0.0	50.4/49.6/0.0	51.7/48.3/0.0	52.2/47.8/0.0	53.2/46.8/0.0	50.9/49.1/0.0	51.9/48.1/0.0
	Long	56.9/43.1/0.0	56.2/43.8/0.0	55.5/44.5/0.0	55.9/44.1/0.0	54.1/45.9/0.0	54.1/45.9/0.0	57.6/42.4/0.0	56.2/43.8/0.0	55.2/44.8/0.0	54.5/45.5/0.0	50.0/50.0/0.0
Qwen2.5-14B	Short	45.2/50.3/4.6	46.8/48.5/4.7	49.2/46.1/4.7	49.1/45.9/5.1	48.6/46.4/5.1	46.7/27.2/26.1	44.7/28.8/26.5	42.6/28.5/28.9	43.5/27.7/28.7	43.2/27.9/28.9	36.0/24.5/39.5
	Medium	56.8/43.2/0.0	56.6/43.4/0.0	54.8/45.2/0.0	56.0/44.0/0.0	54.8/45.2/0.0	52.7/47.3/0.0	50.1/49.9/0.0	52.7/47.3/0.0	52.2/47.8/0.0	53.7/46.3/0.0	48.8/51.2/0.0
	Long	52.4/47.6/0.0	56.2/43.8/0.0	55.9/44.1/0.0	52.4/47.6/0.0	51.4/48.6/0.0	50.3/49.7/0.0	52.1/47.9/0.0	49.0/51.0/0.0	50.3/49.7/0.0	50.7/49.3/0.0	50.3/49.7/0.0
Llama-3.2-1B	Short	47.0/49.1/3.9	47.2/48.9/3.9	45.0/51.0/4.0	44.5/51.3/4.2	44.2/51.6/4.2	43.5/22.8/33.7	43.2/22.8/34.1	43.2/22.6/34.1	41.7/21.8/36.6	41.7/21.7/36.6	34.6/18.3/47.1
	Medium	52.7/47.3/0.0	50.6/49.4/0.0	52.4/47.6/0.0	54.9/45.1/0.0	51.6/48.4/0.0	52.9/47.1/0.0	50.1/49.9/0.0	50.1/49.9/0.0	52.4/47.6/0.0	53.4/46.6/0.0	52.4/47.6/0.0
	Long	51.1/48.9/0.0	52.8/47.2/0.0	54.2/45.8/0.0	55.6/44.4/0.0	55.6/44.4/0.0	55.3/44.7/0.0	57.0/43.0/0.0	55.6/44.4/0.0	55.6/44.4/0.0	50.7/49.3/0.0	46.5/53.5/0.0
Llama-3.2-3B	Short	43.4/52.1/4.6	42.5/52.7/4.8	40.2/55.1/4.7	40.4/54.5/5.1	40.3/54.5/5.1	38.2/27.3/34.5	38.1/26.9/34.9	38.1/26.8/35.1	37.0/25.7/37.3	36.9/25.7/37.5	29.6/22.3/48.1
	Medium	53.4/46.6/0.0	50.6/49.4/0.0	50.6/49.4/0.0	51.1/48.9/0.0	52.7/47.3/0.0	55.4/44.6/0.0	53.4/46.6/0.0	52.7/47.3/0.0	53.7/46.3/0.0	53.9/46.1/0.0	53.9/46.1/0.0
	Long	50.4/49.6/0.0	48.6/51.4/0.0	47.9/52.1/0.0	50.4/49.6/0.0	50.7/49.3/0.0	53.2/46.8/0.0	52.8/47.2/0.0	52.1/47.9/0.0	51.1/48.9/0.0	53.2/46.8/0.0	49.3/50.7/0.0
French												
Mistral-7B	Short	42.4/57.4/0.2	43.2/56.2/0.6	43.6/55.6/0.8	42.6/55.6/1.8	43.2/55.0/1.8	45.4/52.0/2.6	45.6/51.6/2.8	47.8/49.6/2.6	47.8/49.4/2.8	49.2/48.2/2.6	50.2/46.8/3.0
Mistral-8B	Short	42.7/57.1/0.2	42.9/56.1/1.0	45.1/53.7/1.2	43.7/54.7/1.6	44.5/54.1/1.4	46.7/51.9/1.4	48.9/49.3/1.8	50.7/47.3/2.0	50.5/47.3/2.2	51.5/46.5/2.0	52.1/45.5/2.4
Gemma-3-1B	Short	46.2/53.6/0.2	48.0/51.4/0.6	46.0/53.4/0.6	45.9/52.7/1.4	45.2/53.2/1.6	42.7/55.1/2.2	40.6/57.5/1.9	42.7/55.5/1.8	49.9/47.9/2.2	40.7/57.2/2.1	50.3/47.1/2.6
Gemma-3-4B	Short	39.4/60.2/0.4	39.1/60.2/0.6	41.5/57.1/1.4	43.0/56.2/0.8	43.2/51.9/4.9	39.0/54.7/6.3	38.8/56.8/4.4	39.1/58.5/2.4	38.2/59.9/1.9	38.1/59.8/2.1	48.9/49.1/2.0
Qwen2.5-1.5B	Short	42.9/56.9/0.2	41.1/58.3/0.6	43.3/55.1/1.6	45.1/53.7/1.2	43.3/55.5/1.2	40.5/57.3/2.2	40.6/56.9/2.5	40.9/56.3/2.8	49.1/48.1/2.8	39.1/58.7/2.2	48.9/47.9/3.2
Qwen2.5-3B	Short	45.1/54.5/0.4	46.5/52.3/1.2	46.1/52.7/1.2	46.1/52.1/1.8	44.5/54.1/1.4	46.7/51.9/1.4	48.5/49.3/2.2	50.9/47.1/2.0	50.9/46.7/2.4	51.5/46.3/2.2	51.5/45.3/3.2
Qwen2.5-7B	Short	44.4/55.2/0.4	46.5/52.3/1.2	48.9/49.9/1.2	42.3/55.7/2.0	42.7/55.5/1.8	44.3/53.7/2.0	48.9/49.1/2.0	48.9/48.9/2.2	48.9/48.9/2.2	50.1/47.7/2.2	48.3/49.1/2.6
Qwen2.5-14B	Short	39.7/60.1/0.2	39.8/60.0/0.2	39.8/59.8/0.4	42.7/55.5/1.8	42.7/55.5/1.8	39.0/59.4/1.6	40.6/57.3/2.1	39.1/58.5/2.4	48.1/49.3/2.6	39.2/58.5/2.3	48.5/48.7/2.8
Llama-3.2-1B	Short	40.5/59.3/0.2	41.1/58.3/0.6	44.5/54.7/0.8	44.1/54.5/1.4	42.7/55.5/1.8	43.1/55.3/1.6	43.7/54.9/1.4	44.1/54.5/1.4	45.1/53.1/1.8	46.5/51.7/1.8	45.9/51.7/2.4
Llama-3.2-3B	Short	46.9/52.9/0.2	47.7/51.3/1.0	48.9/49.9/1.2	50.3/48.1/1.6	51.1/47.5/1.4	51.7/46.1/2.2	53.1/44.7/2.2	52.9/45.3/1.8	52.5/45.5/2.0	52.7/45.3/2.0	53.1/44.3/2.6

Note: Text length categories determined by K-means clustering: Short: ∼667 tokens, Medium: ∼10,564 tokens, Long: ∼20,549 tokens.

Leading%/Recency%/Neutral% represents the proportion of samples showing leading bias, recency bias, or neutral bias.

The threshold for identifying skipped sentences is defined as $\mu - \alpha\sigma$, where μ and σ are the mean and standard deviation of the normalized similarity scores.

French dataset contains only short text samples, hence the absence of data for Medium and Long categories.

Figure 7: JSON structure of a sample PoSum benchmark instance for positional bias evaluation

```
{
  "id": "multi_domain_enfr_042",
  "conversations": [
    "personA: Hello, I’d like to schedule a meeting for next week.",
    "personB: Sure, what day works best for you?",
    "personA: How about Wednesday afternoon?",
    "personB: That works. I’ll send over a calendar invite."
  ],
  "summary": "Person A and Person B coordinate to schedule a meeting for Wednesday
      afternoon.",
  "llm_generated_summary": [
    {
      "model_name": "Qweb2.5-3B-instruct",
      "gen_summary": "The participants agree to set up a meeting next Wednesday
          afternoon.",
      "similarity_scores": [0.92, 0.87, 0.89, 0.66],
      "similarity_threshold": "mean-0.8*std",
      "missed_sentences_index": [2]
    },
    {
      "model_name": "LLaMA-3-13B",
      "gen_summary": "They discuss scheduling a meeting and settle on Wednesday.",
      "similarity_scores": [0.85, 0.83, 0.80, 0.66],
      "similarity_threshold": "mean-0.8*std",
      "missed_sentences_index": [3]
    }
  ]
}
```

Baseline-Leading Extraction

ID: tweet_restructured_2766

Conversations:

- 379392: Disgusted at high numbers of people without kids using parent &amp; child spaces at @sainsburys Crayford. It needs better oversight please.'
- sainsburys: @379392 ... on the honesty and integrity of our customers. Thanks, Karen 2/2
- 379392: @sainsburys Not quite true. The car park is meant to be patrolled with charges for those abusing the system. Never seen it patrolled, when is this done? https://t.co/Ala69HT4t8'
- sainsburys: @379392 I'm afraid the store is now closed, but I've emailed your feedback on to the Store Manager to be reviewed. Thanks, Naomi.
- 379392: @sainsburys I guess that means my feedback was filed in the bin. I'd like a written response please.
- sainsburys: @379392 I'm afraid this isn't something that we'd get a written response out for. The car park is owned by horizon and is monitored by CCTV by...1/3.

Leading 15% Extraction Summary:

- 379392: Disgusted at high numbers of people without kids using parent &amp; child spaces at @sainsburys Crayford. It needs better oversight please.

Leading 25% Extraction Summary:

- 379392: Disgusted at high numbers of people without kids using parent &amp; child spaces at @sainsburys Crayford. It needs better oversight please.
- sainsburys: @379392 ... on the honesty and integrity of our customers. Thanks, Karen 2/2

Leading 35% Extraction Summary:

- 379392: Disgusted at high numbers of people without kids using parent &amp; child spaces at @sainsburys Crayford. It needs better oversight please.
- sainsburys: @379392 ... on the honesty and integrity of our customers. Thanks, Karen 2/2

Scores and Skipped Sentences:

- **Leading 15%:** Leading Score: 0.56, Recency Score: 0.92, Ignored Indices: [1]
- **Leading 25%:** Leading Score: 1, Recency Score: 0.36, Ignored Indices: [5]
- **Leading 35%:** Leading Score: 1, Recency Score: 0.36, Ignored Indices: [5]

Table 10: Baseline Example: Leading Extraction

Baseline-Recency Extraction

ID: summedits_sales_call_structured_446

Conversations:

- Sales Agent: Good morning/afternoon, may I speak with [Client's Name], please?
- Client: Yes, speaking.
- Sales Agent: Hi, my name is [Sales Agent's Name], and I represent BestInsuranceXYZ. I'm calling you today because we have some great insurance products that could be of interest to you. Do you currently have any insurance products with us?
- Client: No, I don't have any insurance with your company. What kind of insurance do you offer?

. . .

- Client: I'm mostly interested in health insurance. What plans do you have available?

. . .

- Client: Yes, please. Can you also tell me about the pricing for each plan?
- Client: Yes, those prices are reasonable. Are there any deductibles or copayments?
- Client: Yes, let's do it.

. . .

- Sales Agent: Great! I'll send the necessary paperwork to your email, and we can confirm your enrollment once you've completed and submitted it. Thank you for choosing BestInsuranceXYZ for your health insurance needs.

Recency 15% Extraction Summary:

- Client: Yes, let's do it.
- Sales Agent: Great! I'll send the necessary paperwork to your email, and we can confirm your enrollment once you've completed and submitted it. Thank you for choosing BestInsuranceXYZ for your health insurance needs.

Recency 25% Extraction Summary:

- Sales Agent: Yes, there is a deductible for each plan, which varies based on the plan you select. There may also be copayments for certain medical expenses, but we can go over those details when we get you enrolled in a plan. Would you like to move forward with the enrollment process?
- Client: Yes, let's do it.
- Sales Agent: Great! I'll send the necessary paperwork to your email, and we can confirm your enrollment once you've completed and submitted it. Thank you for choosing BestInsuranceXYZ for your health insurance needs.

Recency 35% Extraction Summary:

- Sales Agent: Of course. Our basic plan starts at $100 a month, our standard plan starts at $150 a month, and our premium plan starts at $200 a month. The pricing may vary based on your specific needs and circumstances, but those are our starting rates. Are you comfortable with those prices?
- Client: Yes, those prices are reasonable. Are there any deductibles or copayments?

. . .

- Sales Agent: Great! I'll send the necessary paperwork to your email, and we can confirm your enrollment once you've completed and submitted it. Thank you for choosing BestInsuranceXYZ for your health insurance needs.

Scores and Skipped Sentences:

- **Recency 15%:** Leading Score: 0.87, Recency Score: 0.68, Ignored Indices: [8]
- **Recency 25%:** Leading Score: 0.61, Recency Score: 0.91, Ignored Indices: [3]
- **Recency 35%:** Leading Score: 0.68, Recency Score: 0.87, Ignored Indices: [4]

Table 11: Baseline Example: Recency Extraction