# Not Lost After All: How Cross-Encoder Attribution Challenges Position Bias Assumptions in LLM Summarization

Anonymous ACL submission

## Abstract

Position bias—where Large Language Models (LLMs) overrepresent content from the beginnings and endings of documents while neglecting middle sections—has been considered a core limitation in automatic summarization. To measure position bias, prior studies rely heavily on n-gram matching techniques, which fail to capture semantic relationships in abstractive summaries where content is extensively rephrased. To address this limitation, we introduce a cross-encoder-based alignment method that jointly processes summary–source sentence pairs, enabling more accurate identification of semantic correspondences—even when summaries substantially rewrite the source. Experiments with five LLMs across six summarization datasets reveal markedly different position bias patterns than those reported by traditional metrics. Our findings suggest that these biases primarily reflect rational adaptations to document structure and content rather than true model limitations. Through controlled experiments and analyses across varying document lengths and multi-document settings, we show that LLMs utilize content from all positions more effectively than previously assumed, challenging common claims about "lost-in-the-middle" behaviour.

## 1 Introduction

Large language models (LLMs) have significantly advanced summarization, often producing summaries that approach human-level quality (Goyal et al., 2022a; Zhang et al., 2023). Despite this performance, *position bias*, where models preferentially select summary content from certain document locations, typically the beginning and end raises concerns about their effectiveness.

Initially documented as "lead bias" in news summarization, this phenomenon once seemed appropriate given the standard "inverted pyramid" structure of news articles, which emphasizes early content (Grenander et al., 2019; Norambuena et al., 2020). However, similar biases have since been reported across various neural models (Nallapati et al., 2017; Zhong et al., 2019) and in other domains (Jung et al., 2019a), suggesting broader implications. Recent studies identified a "U-shaped" attention pattern, where models disproportionately neglect middle sections of documents (Ravaut et al., 2024; Liu et al., 2023a), potentially highlighting limitations in summarizing long documents.

The characterization of patterns as biases depends heavily on accurately identifying each source sentence's contribution to the generated summaries. Most existing evaluations rely on n-gram matching, which counts shared word sequences between summaries and sources (Zhong et al., 2019; Ravaut et al., 2024). This method is insufficient for abstractive summaries, which often involve extensive rephrasing; notably, over 80% of bigrams in XSum and over 50% in CNN/DailyMail summaries are novel (Suhara and Alikaniotis, 2024a). Consequently, current evaluations may significantly underestimate how much source content models actually use.

Moreover, labeling position patterns as biases presupposes these patterns indicate model flaws rather than appropriate responses to content distribution. Many documents naturally emphasize important information in specific positions, meaning models' apparent biases may reflect rational content selection rather than weaknesses.

To address these concerns, we introduce a *cross-encoder approach*—a transformer-based model that jointly processes summary-source sentence pairs to explicitly measure semantic alignment. Unlike n-gram methods, cross-encoders directly capture meaning, enabling more accurate source attribution. Specifically, we investigate:

1. How improved semantic alignment alters interpretations of position bias.
2. Which position patterns emerge under precise semantic alignment in standard-length docu-

ments.

3. How these patterns shift in controlled multi-document scenarios with manipulated positions.
4. Whether biases persist in summarizing longer documents with extended context.

Through experiments involving five SotA LLMs and six different datasets, we show substantial deviations from previously reported position patterns. Our results suggest that observed biases typically reflect rational alignment with document structures and important content rather than inherent model limitations.

**Contributions** We make four main contributions: **(1) Methodological:** We introduce and validate a cross-encoder approach for source attribution in abstractive summarization that achieves substantially higher precision than traditional n-gram matching methods. **(2) Empirical:** We provide the first comprehensive analysis of position patterns using semantically-aware attribution, revealing significant deviations from previously reported bias patterns. **(3) Theoretical:** We demonstrate, through controlled experiments, that observed position preferences largely reflect underlying content importance distributions rather than systematic model limitations. **(4) Practical:** We show that models can effectively utilize content from any document position when information value justifies it, including middle sections in long documents previously thought to be "lost."

## 2 Related Work

Position bias in summarization describes model tendency to favour content from specific document locations, particularly document beginnings. This "lead bias" was first documented in news summarization, where models strongly prefer early sentences (Nallapati et al., 2016; Grenander et al., 2019; Xing et al., 2021). While initially considered appropriate for news articles that front-load key information (Norambuena et al., 2020), position bias has since been observed across different neural architectures (Nallapati et al., 2017; Zhong et al., 2019; See et al., 2017) and domains (Jung et al., 2019b; Kedzie et al., 2018). Recent research extended these findings to LLMs, documenting the "lost-in-the-middle" phenomenon where performance degrades for information in context middle positions (Liu et al., 2024; Koren and Goldberg, 2024). Studies have reported U-shaped patterns where models favour document beginnings and ends while neglecting middle sections (Ravaut et al., 2024; Chhabra et al., 2024), casting doubt on transformer architectures to process information distributed throughout long documents.

The fundamental challenge in studying position bias lies in accurately mapping summary content to source locations—a non-trivial task in abstractive summarization where content undergoes substantial semantic transformation (Zhang et al., 2020; Goyal et al., 2022b). Traditional approaches rely on lexical overlap techniques, measuring n-gram matches or word-level similarity between summaries and source segments (Lin, 2004a; Zhong et al., 2020). However, these methods struggle with paraphrasing and abstraction, potentially mischaracterizing how models utilize source content (Suhara and Alikaniotis, 2024a). Alternative approaches include embedding-based similarity measures (Zhang et al., 2019), content unit extraction methods (Liu et al., 2023b), and cross-encoder architectures that jointly process text pairs (Reimers and Gurevych, 2019), though their systematic application to position bias analysis remains limited.

Current evaluation methodologies for position bias range from simple lead overlap counts (Grusky et al., 2018) to sophisticated distribution mapping approaches that compare statistical divergence between model and reference source utilization patterns (Chhabra et al., 2024; Jung et al., 2019b). Input perturbation methods test position effects by manipulating document order (Kedzie et al., 2018; Grenander et al., 2019), though these approaches risk destroying document coherence (Chen and Bansal, 2018). Attention analysis provides another perspective by examining model internals (Jain and Wallace, 2019), yet no studies validate their attribution methods against human-annotated ground truth. We address these methodological limitations by introducing and validating cross-encoder attribution techniques that enable more precise analysis of position patterns in abstractive summarization.

## 3 Methodology

### 3.1 The Attribution Challenge

Accurately identifying which source sentences contribute to summary content is crucial for evaluating abstractive models. Traditional n-gram matching (Lin, 2004b) fails with paraphrased content, while embedding-based methods like BERTScore (Zhang et al., 2019) often misalign topically similar but factually distinct sentences, limiting attribution pre-

cision.

### 3.2 Cross-Encoder Approach.

We propose using cross-encoder models (Reimers and Gurevych, 2019) to capture semantic relationships between summary and source sentences. Unlike bi-encoders that separately encode sentences before comparing embeddings, cross-encoders jointly process concatenated summary-source pairs $[s; d_i]$ through transformer layers. This architecture enables attention mechanisms to model fine-grained semantic connections across the entire input, providing more accurate attribution for paraphrased content than separate encoding approaches.

**Dynamic Selection Strategy.** For each summary sentence $s$ and document sentences $D = \{d_1, d_2, ..., d_n\}$, we select contributing sources in two stages adapting to varying score distributions. Attribution scores vary greatly across instances: highly abstractive summaries may have uniformly low scores, while extractive summaries show clear high-low separation. Fixed thresholds fail to account for this variation, leading to over-selection in some cases and under-selection in others.

Our method first identifies where relevant content transitions to noise by finding the "elbow point"—the position in ranked attribution scores where the score difference is maximized. This boundary detection captures where marginal information gain drops most sharply (Thorndike, 1953). Among sentences scoring above this elbow point, we select those exceeding an adaptive threshold $\mu + 0.5\sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation of all scores. This statistical threshold normalizes for instance-specific score characteristics: the same raw score might indicate high relevance in one case but mediocrity in another. If no sentences meet this criterion, we select the top-scoring sentence as a fallback to ensure attribution coverage.

We use the pre-trained ‘cross-encoder/stsb-roberta-base‘ model without task-specific fine-tuning to demonstrate generalizability across domains. Appendix A provides illustrative examples showing how this approach correctly identifies semantic alignments that Bigram or BERTScore methods miss.

### 3.3 Empirical Validation of Attribution

We validate our cross-encoder approach using expert annotations from Suhara and Alikaniotis (2024b), who hired professional annotators to identify contributing source sentences across 2000 document-summary pairs from XSum and CNN/DailyMail (Krippendorff’s $\alpha = 0.8$). We evaluate using Precision, NDCG@k (ranking quality), and EMD (distributional similarity).

Table 1 demonstrates substantial improvements over existing methods. Most notably, our cross-encoder achieves 78% precision versus 50% for bigram matching on XSum—a 56% relative improvement despite 83.82% of summary bigrams being novel combinations. This highlights traditional methods’ inadequacy for abstractive content.

| Dataset | Method | Precision | NDCG | EMD↓ |
|---|---|---|---|---|
| XSum | Bigram | 0.50 | 0.67 | 0.14 |
| | BERTScore | 0.69 | 0.77 | 0.06 |
| | Cross-Encoder | **0.78** | **0.86** | **0.05** |
| CNN/DM | Bigram | 0.59 | 0.85 | 0.10 |
| | BERTScore | 0.72 | 0.85 | 0.09 |
| | Cross-Encoder | **0.78** | **0.91** | **0.07** |

Table 1: Source attribution performance. All improvements statistically significant ($p < 0.001$).

Crucially, Figure 1 reveals that bigram matching systematically distorts position patterns—underestimating contributions from document beginnings while overestimating from endings. Our cross-encoder produces distributions closely aligned with human annotations, suggesting previously reported U-shaped biases may partially reflect measurement artifacts rather than genuine model behaviours.

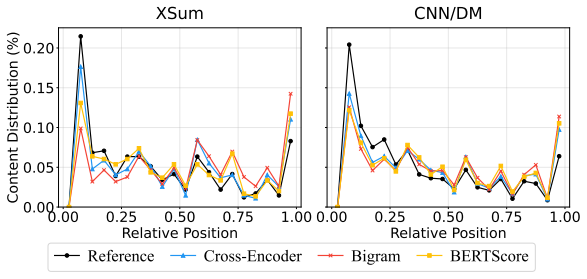


Figure 1: Position distributions by attribution method. Cross-encoder closely matches human annotations while bigram matching shows systematic distortions.

### 3.4 Experimental Design

Using our cross-encoder, we investigate position bias through three complementary experiments: 1)

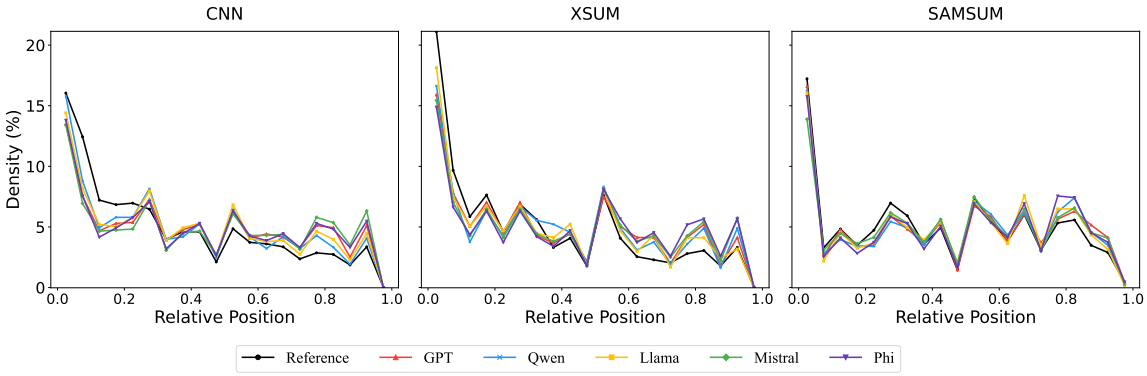


Figure 2: Position distributions comparing model-generated summaries (solid lines) with human references (dashed lines) across CNN/DailyMail, XSum, and SAMSum. Models consistently exhibit rightward shifts, selecting content from later document positions compared to human summarizers.

**Standard Documents:** We analyze position distributions across CNN/DailyMail, XSum, and SAMSum, comparing human references with outputs from five LLMs (Phi-3, GPT-3.5-Turbo, Llama-3.2-1B, Mistral-7B, Qwen-2.5-7B) to distinguish domain-specific patterns from general patterns. 2) **Controlled Order Manipulation:** To isolate position effects from content importance, we create document pairs in alternate orders (Doc1+Doc2 vs. Doc2+Doc1) using 500 examples per dataset (CNN/DailyMail, XSum, and SAMSum), measuring how position influences selection. 3) **Long Documents:** We extend analysis to ArXiv, MultiNews, and GovReport to determine whether position patterns scale with length or represent architectural limitations. In all experiments, we normalize positions to [0,1], analyze both continuous distributions and sectional breakdowns, and apply multiple statistical tests (KS (Massey Jr, 1951), Mann-Whitney U (Mann and Whitney, 1947), $t$-test (Student, 1908)).Appendix B and Appendix C provide concrete examples of the dataset characteristics and model configurations used in our experiments. Example prompts and generation parameters can be found in Appendix E.

## 4 Results

### 4.1 Position Bias in Standard Documents

**Accurate attribution reveals rightward shifts, not U-shaped bias.** To our second research question, we analyze position patterns using cross-encoder attribution across three standard-length datasets. Our findings fundamentally challenge previous characterizations of position bias in LLM summarization. Figure 2 reveals that, while all summaries appropriately select more content from document beginnings (where important information typically concentrates), models systematically select content from later document positions than human references across all datasets. This reflects rational information seeking rather than bias, with models demonstrating more balanced content use than human summarizers. These findings directly contradict the widely-reported U-shaped attention hypothesis, where models allegedly favour beginnings and ends while neglecting middle sections.

| Model | CNN/DM | XSum | SAMSum |
|---|---|---|---|
| Reference | 0.32 | 0.31 | 0.40 |
| GPT-3.5 | **0.40** | **0.37** | **0.43** |
| Llama-3 | **0.38** | 0.35 | **0.43** |
| Mistral | **0.42** | **0.39** | **0.44** |
| Phi-3 | **0.40** | **0.40** | **0.45** |
| Qwen | **0.36** | **0.37** | **0.44** |

Table 2: Mean position values across models and datasets. Bold indicates statistically significant rightward shifts compared to references ($p < 0.05$).

Table 2 quantifies these rightward shifts, with models achieving mean positions 0.041-0.098 points higher than references in CNN/DM. This systematic pattern holds across all five LLMs and three diverse datasets, indicating more balanced content selection than human summarizers rather than systematic bias toward document boundaries.

The sectional analysis in Figure 3 reveals the mechanism underlying these shifts: models extract 7-12% less content from beginning sections while incorporating 5-9% more from middle and later sections. This redistributive pattern appears across structurally diverse content—from news articles to dialogue—suggesting that accurate semantic attribution reveals sophisticated content selection strate-

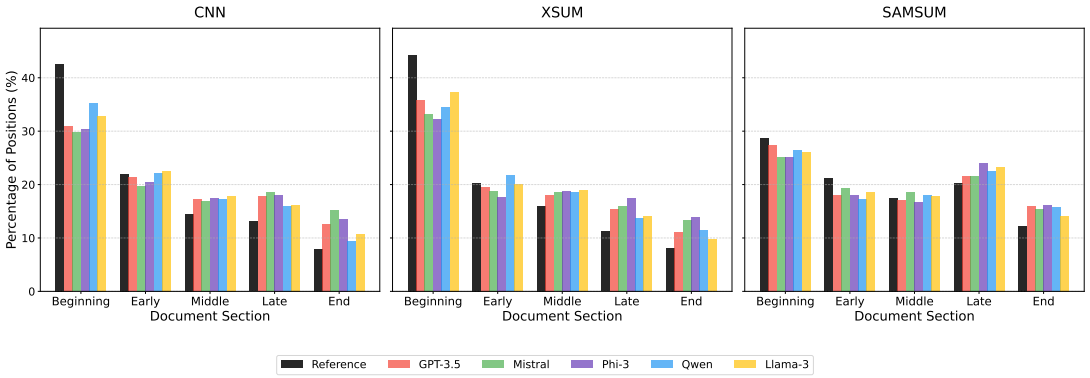


Figure 3: Content extraction by document sections. Models consistently reduce reliance on beginning sections while increasing utilization of middle and later sections compared to human references.

gies rather than positional limitations.

### 4.2 Interaction Effects: Pattern Variations

While the rightward shift appears universally, its expression varies across contexts. This variation follows a three-factor interaction pattern that explains the diversity in reported position bias findings: 1) **Universal tendency toward balanced selection.** All models show rightward shifts compared to humans, suggesting neural architectures naturally distribute attention more evenly across documents. 2) **Content-dependent modulation.** This tendency manifests differently across domains: strongly in news (CNN/DM: +0.041 to +0.098), variably in abstractive tasks (XSum: +0.043 to +0.095), and consistently in dialogue (SAMSum: +0.030 to +0.048). 3) **Architecture-specific differences.** Model variations become pronounced in highly abstractive contexts, where Phi-3 shows the strongest rebalancing (+0.095) while Llama-3's shift is insignificant.

Rather than viewing position patterns as fixed biases, these findings suggest they emerge from rational content assessment that vary based on document structure, task demands, and architecture.

While these correlational findings reveal consistent position patterns, they leave a key question unanswered: do models select content based on position or because important information happens to appear in certain locations? In Phase 2, we address this confound through controlled experiments where we rearrange identical content into different positions. This design tests whether identical information receives different treatment based solely on its position.

### 4.3 Document Order Manipulation

Previous studies test position bias by shuffling sentences (Kedzie et al., 2018), which destroys document structure. Instead, we concatenate two documents in different orders: Doc1+Doc2 versus Doc2+Doc1. This preserves coherence while testing whether models treat identical content differently based on its sequential position.

We examine two critical aspects: (1) Does document position affect how many sentences models select from each document? (2) Do models select sentences from the same positions within documents regardless of global order? Our findings reveal a nuanced pattern where sentence counts show statistical significance but position distributions demonstrate remarkable stability.

| Data | Model | D1+D2 | D2+D1 | Diff | p-value |
|---|---|---|---|---|---|
| CNN/DM | GPT-3.5 | 4.03 | 4.77 | +0.74 | 2.7e-09** |
| | Llama-3.2 | 5.99 | 5.64 | -0.35 | 0.002* |
| | Mistral | 4.58 | 5.42 | +0.84 | 1.7e-12** |
| | Phi-3 | 4.13 | 4.60 | +0.47 | 9.7e-05** |
| | Qwen | 4.71 | 5.09 | +0.38 | 0.0006** |
| XSum | GPT-3.5 | 3.33 | 3.92 | +0.60 | 1.5e-08** |
| | Llama-3.2 | 4.26 | 4.46 | +0.20 | 0.047* |
| | Mistral | 4.01 | 4.77 | +0.76 | 1.3e-13** |
| | Phi-3 | 4.03 | 4.15 | +0.12 | 0.257 |
| | Qwen | 4.54 | 4.77 | +0.23 | 0.028* |
| SAMSum | GPT-3.5 | 2.01 | 2.00 | -0.01 | 0.895 |
| | Llama-3.2 | 3.41 | 2.58 | -0.84 | 6.7e-10** |
| | Mistral | 2.43 | 3.02 | +0.59 | 8.4e-09** |
| | Phi-3 | 2.65 | 3.04 | +0.39 | 0.0002** |
| | Qwen | 2.60 | 2.93 | +0.33 | 0.0015** |

*p < 0.05, **p < 0.001

Table 3: Sentence count differences across configs.

Our findings demonstrate a key insight: while document order can produce statistically detectable effects on sentence counts, models maintain remarkable consistency in the positions from which

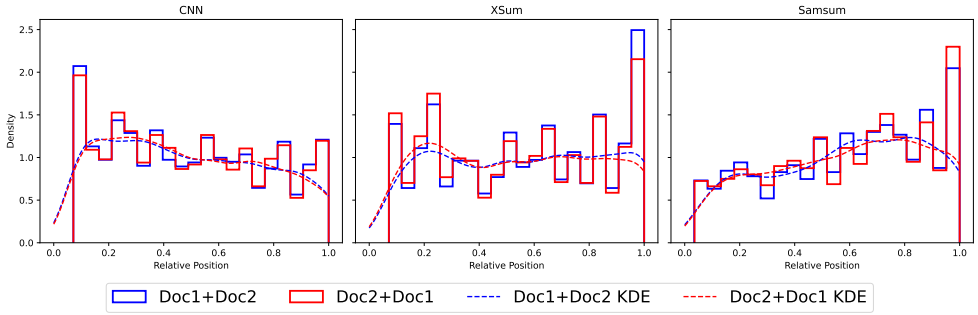


Figure 4: Llama3.2 position distributions across document configurations. The overlapping histograms demonstrate that models maintain consistent selection patterns within documents regardless of global order. Even when sentence counts differ statistically, the positions of selected content remain stable.

| Data | Model | Doc1 $p$-val | Sig? | Doc2 $p$-val | Sig? |
|---|---|---|---|---|---|
| CNN/DM | GPT-3.5 | 0.038 | Yes* | 0.405 | No |
| | Llama-3.2 | 0.511 | No | 1.000 | No |
| | Mistral | 0.079 | No | 0.918 | No |
| | Phi-3 | 0.180 | No | 0.230 | No |
| | Qwen | 0.739 | No | 0.988 | No |
| XSum | GPT-3.5 | 0.480 | No | 0.018 | Yes* |
| | Llama-3.2 | 0.017 | Yes* | 0.327 | No |
| | Mistral | 0.581 | No | 0.411 | No |
| | Phi-3 | 0.366 | No | 0.802 | No |
| | Qwen | 0.949 | No | 0.803 | No |
| SAMSum | GPT-3.5 | 0.233 | No | 0.960 | No |
| | Llama-3.2 | 0.454 | No | 0.990 | No |
| | Mistral | 0.908 | No | 0.513 | No |
| | Phi-3 | 0.406 | No | 0.699 | No |
| | Qwen | 0.126 | No | 0.244 | No |

*$p < 0.05$

Table 4: Position distribution stability in documents.

they extract content. The 90% consistency rate in position distributions suggests that models effectively identify and extract informative content regardless of global document ordering.

These results establish that, in controlled two-document settings, position effects are modest and do not fundamentally alter content assessment. However, this raises important questions about longer contexts where "lost-in-the-middle" effects are widely reported. Phase 3 examines whether this position-independent evaluation extends to substantially longer documents and multi-document scenarios. Figure 4 visualizes this stability. The overlapping distributions confirm that models evaluate content based on intrinsic information rather than global position, even when they adjust selection volume in response to document ordering.

### 4.4 Position Bias in Extended Contexts

To investigate whether position patterns scale to longer inputs, we analyze three challenging datasets: ArXiv (scientific papers), GovReport (government documents), and Multi-News (multi-document collections). This addresses our fourth research question: do position patterns persist in extended contexts where "lost-in-the-middle" effects are commonly reported?

#### 4.4.1 Context-Dependent Position Effects

Figure 5 reveals a striking pattern: position bias varies dramatically by document type, not just length. Scientific papers show substantial model-reference divergence, while government documents exhibit remarkable alignment for some models.

Table 5 quantifies these differences, revealing three key insights: 1) **Document structure matters more than length.** The same model shows vastly different behaviours across document types. GPT-3.5 exhibits high divergence in scientific papers (KS = 0.123, $p < 0.001$) but near-perfect alignment in government documents (KS = 0.017, $p = 0.230$). 2) **Size doesn't predict performance.** Smaller models often outperform larger ones. Phi-3 (3B parameters) shows the best ArXiv alignment (KS = 0.019, $p = 0.794$), while GPT-3.5 shows the worst, challenging assumptions about scale and bias. 3) **Models adapt to document conventions.** Rather than exhibiting fixed biases, models demonstrate sophisticated adaptation to different information structures, suggesting content-driven rather than position-driven selection.

Figure 6 provides section-level analysis. In scientific papers, models over-extract from document boundaries—Mistral selects 38% from the first 20% versus 27% for humans. Government documents show more uniform extraction, with Llama-3 nearly

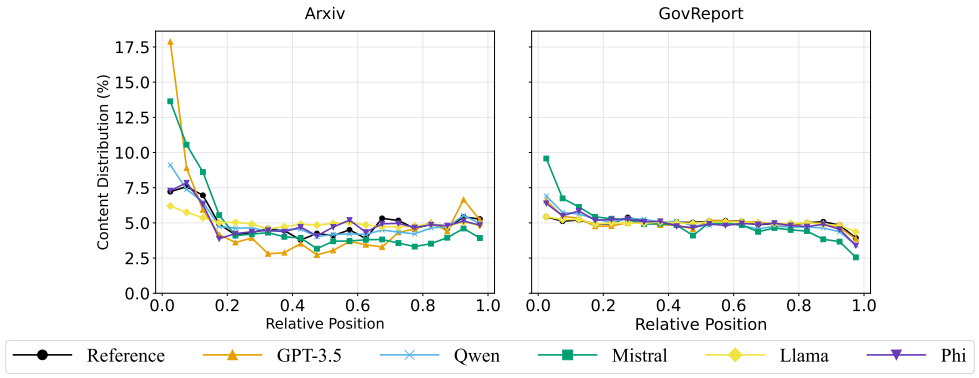


Figure 5: Position distributions in long documents. Scientific papers (ArXiv) show more pronounced model-reference differences than government documents (GovReport), indicating that position patterns depend on document structure rather than length alone.

| Model | ArXiv (Scientific Papers) | | | | GovReport (Government Docs) | | | |
|---|---|---|---|---|---|---|---|---|
| | KL | JS | WD | KS (p-val) | KL | JS | WD | KS (p-val) |
| GPT-3.5 | 0.078 | 0.018 | 0.050 | 0.123 (<0.001) | 0.002 | <0.001 | 0.006 | 0.017 (0.230) |
| Llama-3 | 0.012 | 0.003 | 0.016 | 0.046 (0.002) | <0.001 | <0.001 | 0.003 | 0.005 (0.986) |
| Mistral | 0.045 | 0.011 | 0.078 | 0.119 (<0.001) | 0.024 | 0.006 | 0.053 | 0.077 (<0.001) |
| Phi-3 | 0.004 | 0.001 | 0.007 | 0.019 (0.794) | 0.002 | <0.001 | 0.016 | 0.027 (0.001) |
| Qwen | 0.006 | 0.001 | 0.014 | 0.026 (0.289) | 0.004 | <0.001 | 0.022 | 0.033 (0.001) |

KL = Kullback-Leibler; JS = Jensen-Shannon; WD = Wasserstein; KS = Kolmogorov-Smirnov

Table 5: Position distribution divergence in long documents. Lower values indicate closer alignment with humans.

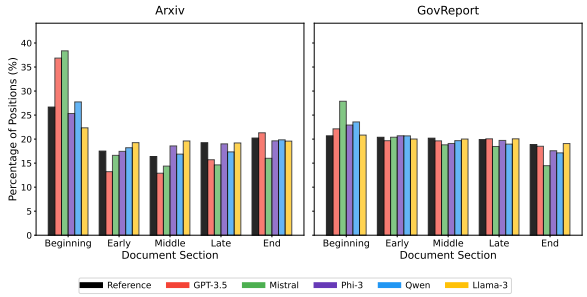


Figure 6: Content extraction by document sections. Scientific papers show boundary bias (high beginning/end extraction), than government documents.

matching human patterns across all sections.

#### 4.4.2 Refuting "Lost-in-the-Middle"

Multi-News provides a naturalistic test of "lost-in-the-middle" claims. Unlike artificial manipulations, this dataset requires models to integrate across multiple sources where important content naturally appears throughout the sequence.

Table 6 shows successful middle position use. Key findings: 1) **Middle position extraction:** All models show median global positions near 0.5, indicating substantial middle content use. GPT-3.5 (median = 0.453) and Phi-3 (median = 0.455) center precisely on middle positions. 2) **Distributed source attention:** High entropy values (3.27-3.85) show models attend broadly across sources rather than focusing on a few. Most models match human entropy patterns (3.83-3.84). 3) **Quality maintained:** Despite distributed attention, models achieve high content overlap with references. Qwen (0.915 Jaccard) and Llama-3 (0.904) show that middle focus doesn't compromise quality.

Figure 7 visualizes this success. Both Qwen and Phi show balanced local and global position distributions, contradicting claims that models cannot effectively process middle content in long sequences.

#### 4.4.3 Implications: Rethinking Position Bias

Our extended context analysis reveals that position bias is neither universal nor primarily length-dependent. Instead, it reflects: 1) **Document-specific adaptation:** Models adjust to different information structures (scientific vs. government writing), showing sophisticated content assessment rather than rigid positional preferences. 2) **Quality over position:** In multi-document settings where middle positions contain crucial information, models successfully extract and utilize this content while maintaining high summary quality. 3) **Architecture-content interactions:** Different models excel with different document types, sug-

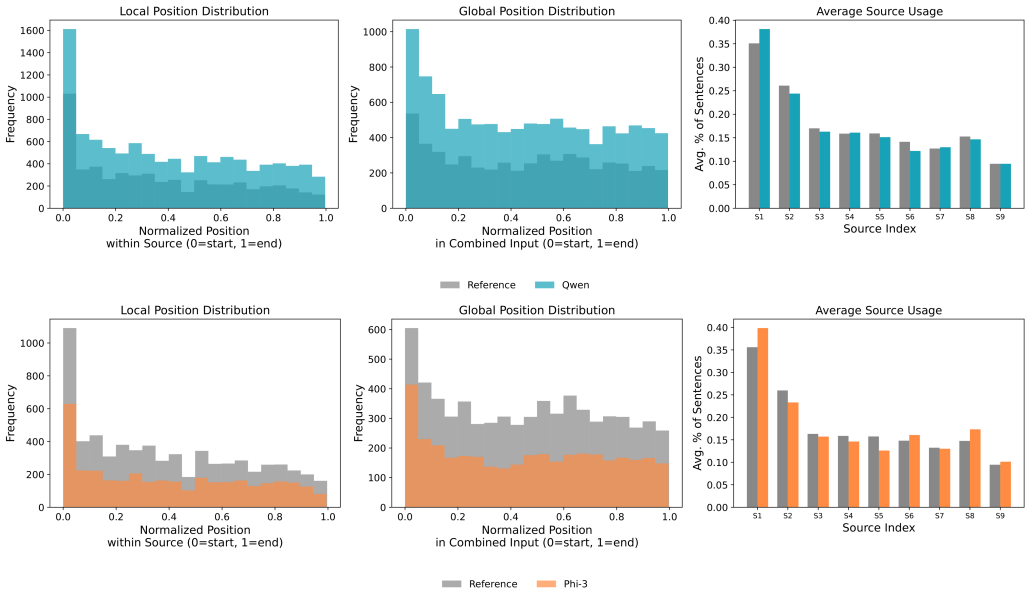


Figure 7: Multi-document position analysis for Qwen and Phi models, which successfully extract content from global middle positions, with balanced local and global position distributions.

| Model | Global Position Mean (Median) | Global Reference Mean (Median) | Source Entropy (Reference) | KS Statistic (p-value) | Jaccard Similarity |
|---|---|---|---|---|---|
| GPT-3.5 | 0.458 (0.453) | 0.458 (0.459) | 3.85 (3.84) | 0.022 (0.147) | $0.871 \pm 0.187$ |
| Phi-3 | 0.452 (0.455) | 0.460 (0.466) | 3.80 (3.83) | 0.028 (0.055) | $0.817 \pm 0.222$ |
| Llama-3 | 0.473 (0.471) | 0.459 (0.465) | 3.71 (3.83) | 0.027 (0.001) | $0.904 \pm 0.146$ |
| Qwen | 0.447 (0.436) | 0.456 (0.464) | 3.77 (3.83) | 0.028 (0.007) | $0.915 \pm 0.141$ |
| Mistral | 0.440 (0.381) | 0.509 (0.537) | 3.27 (3.66) | 0.216 (0.006) | $0.768 \pm 0.234$ |

Table 6: Multi-document position statistics. All models successfully utilize middle positions.

gesting that "bias" patterns reflect architectural strengths rather than fundamental limitations.

These findings challenge the characterization of position bias as a universal model limitation. Instead, they suggest that LLMs implement adaptive summarization strategies that prioritize content over position, even in extended contexts where such limitations might be expected.

## 5 Conclusion

This paper fundamentally reframes position bias in LLM summarization through improved semantic attribution. Using cross-encoder methods, we demonstrate that reported position biases largely reflect rational content assessment rather than architectural limitations. We challenge these core assumptions across five models and multiple datasets. First, the widely-cited U-shaped attention pattern does not hold—models show rightward shifts toward more balanced content use compared to humans. Second, controlled position manipulation reveals minimal systematic effects: 90% of comparisons show no significant differences in where models select content, even when sentence counts vary. Third, extended context analysis refutes “lost-in-the-middle” claims—models successfully extract from global middle positions (median $\sim$ 0.5) in multi-document settings while maintaining quality. Most importantly, position patterns prove context-dependent rather than universal. Models that struggle with scientific papers excel with government documents, demonstrating adaptive strategies that prioritize content structure over positional heuristics. This suggests that "bias" reflects sophisticated document-type recognition rather than processing limitations. These results shift the research focus from bias mitigation to content assessment enhancement. Future work should develop semantic evaluation frameworks that reveal model capabilities obscured by traditional metrics. Our cross-encoder approach provides such a foundation, showing that concerns about positional limitations may be overstated when models possess robust content evaluation mechanisms.

## 6 Limitations

While our work offers important insights into position bias through improved semantic attribution, several limitations present opportunities for future research in this area.

First, though our cross-encoder approach demonstrates substantial improvement over traditional methods (achieving 78% precision compared to 50% for bigram matching on XSum), attribution remains challenging for highly abstractive summaries. The complexity of mapping semantic relationships in extensively rewritten content means that even our enhanced methodology cannot perfectly capture all summary-source connections, particularly in cases of extreme abstraction or implicit inferencing.

Second, our findings establish strong correlational patterns between content selection and document position, though fully isolating causal mechanisms presents inherent challenges. Though our document-order manipulation experiments demonstrate consistent position preferences despite reordering, establishing definitive causal relationships between position and content selection remains difficult within the constraints of natural language, where content importance and position are often intrinsically linked in well-formed documents.

Third, our study examines five diverse models and six datasets spanning multiple domains, providing a robust foundation for our conclusions. Nevertheless, the LLM landscape continues to evolve rapidly, and extending this analysis to additional architectural families and specialized domains would further validate the generalizability of our findings. The significant variation we observed across document types—particularly between scientific papers and government documents—suggests rich territory for exploring how position patterns interact with different document structures and conventions.

## References


Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. 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.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev Wang. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022a. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

Tanya Goyal, Junyi Jessy Zhang, and Greg Durrett. 2022b. News summarization and evaluation in the era of gpt-3. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3680–3695. Association for Computational Linguistics.

Max Grenander, Yue Dong, David M. Blei, and Kathleen McKeown. 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.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Luyang Huang, Shuyang Cao, Barun Paranjape, Saurabh Chopra, and Wen-tau Yih Hassan Suleman.

2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019a. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. 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 3324–3335, Hong Kong, China. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lianhui Yang, Hyungsuk Jung, Sunghun Choi, Hwanhee Lee, Seung-won Hwang, and Eunjeong L. Park. 2019b. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6035–6045.

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.

Avi Koren and Yoav Goldberg. 2024. Long context compression with activation beacon. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

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

Chin-Yew Lin. 2004b. 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. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Nelson F. Liu, Matei Zaharia, and Christopher Ré. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:574–590.

Xiangyang Liu, Yansong Chen, Jiacheng Yao, Fangzhou Fan, and Dongyan Zhou. 2023b. Towards improving faithfulness in abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9295–9310, Toronto, Canada. Association for Computational Linguistics.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Bryan Norambuena, Enrique Horta, Axel Soto, and Daniel Cabrero. 2020. Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. In *Proceedings of the 2020 Computation + Journalism Symposium*.

Mathieu Ravaut, Siqi Jaunet, Yi Tay, Dara Bahri, Donald Dugan, Mitchell Weiss, and Rahma Chaabouni Saurous. 2024. On context utilization in summarization with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

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.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Yoshi Suhara and Dimitris Alikaniotis. 2024a. Source identification in abstractive summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 212–224, St. Julian's, Malta. Association for Computational Linguistics.

Yoshi Suhara and Dimitris Alikaniotis. 2024b. Source identification in abstractive summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 212–224, St. Julian's, Malta. Association for Computational Linguistics.

Robert L Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18(4):267–276.

Bo Xing, Zeyao Wang, and Zhichun Yin. 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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. A closer look at data bias in neural extractive summarization models. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208. Association for Computational Linguistics.

# A Cross-Encoder Implementation Details

## A.1 Model Architecture and Processing

Our cross-encoder approach utilizes the pre-trained `cross-encoder/stsb-roberta-base` model for several methodological reasons. This model processes concatenated summary-source sentence pairs $[s; d_i]$ through shared transformer layers, enabling joint attention across both texts. We selected this specific architecture based on three considerations: (1) its training on semantic textual similarity tasks aligns with our attribution objectives, (2) the RoBERTa-base size provides computational tractability for large-scale experiments while maintaining representational capacity, and (3) using a general-purpose model without domain-specific fine-tuning demonstrates the robustness of our approach across diverse datasets. Unlike bi-encoders that separately encode sentences before similarity computation, this joint processing architecture enables attention mechanisms to model semantic relationships across the entire input sequence.

## A.2 Dynamic Selection Strategy

For each summary sentence $s$ and document sentences $D = \{d_1, d_2, ..., d_n\}$, our attribution method operates in two stages:

1. **Elbow Point Detection:** We identify the position in ranked attribution scores where the score difference is maximized, capturing where marginal information gain drops most sharply.

2. **Adaptive Thresholding:** Among sentences scoring above the elbow point, we select those exceeding $\mu + 0.5\sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation of all scores.

If no sentences meet this criterion, we select the top-scoring sentence as a fallback to ensure attribution coverage.

## A.3 Illustrative Example: Semantic Nuance Detection

To demonstrate the superior capability of our cross-encoder approach, consider this real example from XSum:

**Source Document:** "Chief Secretary to the Treasury Danny Alexander, former Lib Dem leader Charles Kennedy and John Thurso were beaten by the SNP... Mr Kennedy, who lost Ross, Skye and Lochaber to Ian Blackford, said the 2015 election's defeat of Lib Dems and Labour in Scotland would become known as the 'night of the long sgian dubhs'..."

**Generated Summary:** "High profile Liberal Democrats have lost three strongholds in the Highlands and Islands."

**Ground Truth Attribution:** Sentences 0 and 3 (human annotated)

**Method Comparison:**

- **Bigram Matching:** Selected sentence 8 ("He said the Liberal Democrats should hold their heads high...") with only 8.3% overlap. Achieved 0% precision and recall.
- **BERTScore:** Selected sentences 1, 3, 10, 11 based on embedding similarity. Achieved 25% precision due to topical similarity without semantic correspondence.
- **Cross-Encoder:** Correctly identified sentence 3 with a score of 0.999, achieving 100% precision. The model captured that "defeat of Lib Dems" semantically corresponds to "lost three strongholds," despite completely different surface forms.

This example illustrates how traditional methods fail with abstractive content: bigram matching finds no meaningful connections, while BERTScore conflates topical similarity with semantic correspondence. Our cross-encoder successfully identifies the semantic relationship between "defeat" and "lost strongholds," demonstrating its superiority for abstractive summarization evaluation.

## B Dataset Statistics

Our evaluation spans six diverse datasets with varying structural and domain characteristics. Three key aspects distinguish our experimental design: **(1) Document Length Diversity:** We analyze both standard-length documents (142-656 tokens) and extended contexts (2,103-8,912 tokens) to test scalability of position patterns. **(2) Domain Coverage:** Our datasets span news (CNN/DM, XSum), dialogue (SAMSum), scientific writing (ArXiv), government documents (GovReport), and multi-document scenarios (Multi-News) to ensure generalizability across text types. **(3) Abstractiveness Levels:** XSum represents highly abstractive summarization (21 tokens, single sentence), while CNN/DM and others allow more extractive approaches, enabling us to test how summarization style affects position bias patterns. Complete statistics are provided in Table 7.

## C Model Specifications

We evaluate five state-of-the-art language models representing different scales and architectural approaches. Our selection ensures comprehensive coverage across model sizes (1B to 175B parameters), organizations (OpenAI, Meta, Microsoft, Mistral AI, Alibaba), and context capabilities (16K to 131K tokens). All models use instruct-tuned versions to ensure optimal summarization performance. Detailed specifications are shown in Table 8.

## D Multi-News Source Distribution

Multi-News contains instances with varying numbers of source articles (1-9 news articles per instance). To ensure robust analysis across different complexities, we systematically sampled at least 20 instances for each source count when possible, resulting in balanced representation across document configurations. This distribution allows us to test position bias across varying document complexities, from single-source instances (equivalent to standard summarization) to complex multi-source scenarios where content importance is distributed throughout the sequence.

## E Experimental Configuration

### E.1 Prompting Strategies

We employ dataset-specific prompts designed to optimize summarization quality while maintaining consistency across models. All prompts position the model as a "professional summarizer" to encourage high-quality output.

**Phase 1 - Standard Documents** For CNN/DailyMail, XSum, and SAMSum:

```
You are a professional summarizer.
Summarize the following text in {n}
sentences.
```

where `{n}` represents the average summary length (CNN/DM: 3, XSum: 1, SAMSum: 1).

**Phase 2 - Document Order Manipulation** For two-document concatenation experiments:

```
You are a professional summarizer. The
following are two unrelated articles.
Summarize the key point of each article
in a coherent manner.
Article 1: {article1}
Article 2: {article2}
```

**Phase 3 - Extended Contexts**

- **ArXiv:** `You are a professional summarizer. Summarize the scientific paper. Paper: {article}`

Table 7: Key dataset characteristics for position bias analysis

| Dataset | Domain | Samples | Document Length (tokens) |
|---|---|---|---|
| CNN/DailyMail (Hermann et al., 2015) | News | 1,000 | 994.56 |
| XSum (Narayan et al., 2018) | News | 1,000 | 566.79 |
| SAMSum (Gliwa et al., 2019) | Dialogue | 819 | 175.54 |
| ArXiv (Cohan et al., 2018) | Scientific | 200 | 8,940.00 |
| GovReport (Huang et al., 2021) | Government | 200 | 11,025.02 |
| Multi-News (Fabbri et al., 2019) | Multi-Document | 157 | 2,998.52 |

Table 8: Large Language Model specifications and configurations

| Model | Parameters | Context Window | Organization | Release Date |
|---|---|---|---|---|
| GPT-3.5-turbo | 175B | 16,385 tokens | OpenAI | March 2023 |
| Llama-3.2-1B-Instruct | 1B | 131,072 tokens | Meta | September 2024 |
| Mistral-7B-Instruct-v0.2 | 7B | 32,768 tokens | Mistral AI | December 2023 |
| Phi-3-mini-128k-Instruct | 3.8B | 128,000 tokens | Microsoft | April 2024 |
| Qwen-2.5-7B-Instruct | 7B | 32,768 tokens | Alibaba | September 2024 |

- **GovReport:** `You are a professional summarizer. Summarize the government report. Report: {article}`
- **Multi-News:** `You are a professional summarizer. Summarize each article news in a coherent manner. Paper: {article}`

These prompts balance specificity with generality, providing clear task framing without biasing content selection toward particular document positions.

### E.2 Generation Parameters

Following Ravaut et al. (2024), we employ consistent generation parameters across all models:

- Temperature: 0.3
- Top-k: 50
- Max tokens: Adaptive based on dataset (50-250 tokens)
- Stop sequences: Model-specific defaults

### E.3 Computational Infrastructure

All experiments were conducted on NVIDIA A40 GPUs with 48GB memory. API-based models (GPT-3.5) utilized rate limiting of 60 requests per minute to ensure reproducibility.

### E.4 Statistical Testing

Position distribution comparisons employ multiple statistical tests for robustness:

- Kolmogorov-Smirnov test for distribution equality
- Mann-Whitney U test for median differences
- Two-sample t-test for mean differences
- Jensen-Shannon divergence for distributional similarity

Significance levels are set at $\alpha = 0.05$ with Bonferroni correction for multiple comparisons where applicable.