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

Elahe Rahimi[†‡] **Hassan Sajjad**[†] **Domenic Rosati**[†‡]
Abeer Badawi[§‡] **Elham Dolatabadi**[§‡] **Frank Rudzicz**[†‡]

[†]Dalhousie University [‡]Vector Institute [§]York University

{erahimi, HSajjad, Domenic.Rosati, frank}@dal.ca
{abeerbadawi, edolatab}@yorku.ca

Abstract

Position bias, the tendency of Large Language Models (LLMs) to select content based on its structural position in a document rather than its semantic relevance, has been viewed as a key 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 apply 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 significantly different position bias patterns than those reported by traditional metrics. Our findings suggest that these patterns 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 use 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., 2022; 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 questions about whether these models truly understand content importance or simply exploit positional shortcuts.

Initially documented as "lead bias" in news summarization, this phenomenon was considered appropriate because of the standard "inverted pyramid" structure of news articles, which emphasizes early content (Norambuena et al., 2020). However, similar positional preferences have since been reported across various neural architectures (Nallapati et al., 2017; Zhong et al., 2019) and domains (Jung et al., 2019a), suggesting broader implications beyond journalism. More recently, studies have identified a U-shaped attention pattern in which models disproportionately neglect middle sections of documents (Ravaut et al., 2024; Liu et al., 2023a), raising concerns about their ability to faithfully process and summarize long-form content.

However, characterizing these patterns as biases depends critically on accurately identifying how source content contributes to 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 approach is fundamentally inadequate for abstractive summaries, which involve extensive rephrasing. In fact, over 80% of bigrams in XSum and over 50% in CNN/DailyMail summaries are novel (Suhara and Alikaniotis, 2024), demonstrating the limitations of lexical overlap methods. Consequently, current evaluations may significantly underestimate how much source content models actually utilize.

Furthermore, treating position patterns as biases assumes they reflect model limitations rather than rational responses to document structure. Many documents naturally emphasize important information in specific locations, meaning apparent positional preferences may instead reflect effective content selection. This raises the critical need for more reliable methods to distinguish between true model limitations and appropriate structural adaptations.

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 that rely on surface-level word

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15625–15641
November 4-9, 2025

overlap, cross-encoders directly capture meaning relationships, enabling more accurate source attribution even when content is substantially rephrased. 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 documents.
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 with five state-of-the-art LLMs across six datasets, we demonstrate substantial deviations from previously reported position patterns. Our findings suggest that observed positional preferences typically reflect rational alignment with document structures and content importance rather than inherent model limitations.

Contributions We make four main contributions: **(1) Methodological:** We adapt 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 findings. **(3) Theoretical:** We demonstrate, through controlled experiments, that observed positional 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

2.1 Position Bias Characterization and Measurement

Early investigations in summarization revealed that models disproportionately select content from the beginning of news articles, a phenomenon referred to as lead bias. This behaviour was documented and countered in neural systems (Grenander et al., 2019; Xing et al., 2021). Analyses across architectures and domains confirmed that reliance on positional cues is not limited to specific models or datasets (Kedzie et al., 2018; Jung et al., 2019b; Zhong et al., 2019). More recently, position effects have been studied in large language models. Liu et al. (2024) report the "lost-in-the-middle" phenomenon, where models underutilize mid-context information. Ravaut et al. (2024) observe U-shaped utilization patterns across multiple datasets and models, showing consistent emphasis on early and late segments. Complementing these behavioural findings, Chhabra et al. (2024) propose a distributional formulation of position bias and introduce Wasserstein distance as a metric to compare system and reference positional distributions.

2.2 Attribution Challenges in Abstractive Summarization

A central limitation in studying position effects is the difficulty of attributing abstractive summaries to their sources. Heavy paraphrasing and compression obscure lexical overlap, complicating reliable content mapping (Zhang et al., 2020; See et al., 2017; Chhabra et al., 2024). Suhara and Alikaniotis (2024) address this by defining source sentences and benchmarking multiple attribution methods, finding that perplexity-based approaches perform better in highly abstractive cases while similarity-based methods are more effective in extractive ones, though both degrade under paraphrase. Xu and Durrett (2021) extend attribution analysis through decoder-level interpretation, but their reliance on attention and gradient-based techniques lacks validation against human judgments (Jain and Wallace, 2019). This illustrates why n-gram overlap metrics such as ROUGE (Lin, 2004) can underestimate content reuse (Goyal et al., 2022). Alternatives include embedding-based similarity measures such as BERTScore (Zhang et al., 2019) and content-unit frameworks (Liu et al., 2023b; Zhong et al., 2020), though their application to position bias analysis remains limited. As a result, most existing studies rely on surface-level heuristics that risk conflating genuine positional preference with rational exploitation of document structure. Our work addresses this gap by introducing cross-encoder attribution, enabling robust semantic alignment under paraphrasing and a more faithful evaluation of whether observed patterns represent true bias or justified content selection.

3 Methodology

Accurately identifying which source sentences contribute to summary content is crucial for evaluating

abstractive models, as traditional n-gram matching fails (Lin, 2004) with paraphrased content while embedding-based methods like BERTScore often misalign topically similar but factually distinct sentences. We use a cross-encoder model (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.

Our method extends the standard cross-encoder application through two components tailored for source attribution in abstractive summarization:

Pairwise Scoring. We systematically pair each summary sentence with all document sentences, scoring semantic alignment via the pre-trained `cross-encoder/stsb-roberta-base` model. This produces an $n \times m$ similarity matrix where n is the number of summary sentences and m is the number of document sentences.

Dynamic Selection Strategy. From the resulting similarity matrix, we select contributing sources for each summary sentence through adaptive thresholding that accounts for 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 μ and σ 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, depending on that instance's score distribution. If no sentences meet this criterion, we select the top-scoring sentence as a fallback to ensure attribution coverage.

Our elbow point method achieves superior performance compared to alternative thresholding approaches across multiple evaluation metrics (see Appendix A). We use the pre-trained `cross-encoder/stsb-roberta-base` model without task-specific fine-tuning to demonstrate generalizability across domains. Illustrative examples are provided in Appendix B, showing how this approach correctly identifies semantic alignments that other methods miss.

3.1 Empirical Validation

We validate our cross-encoder approach using expert annotations from Suhara and Alikaniotis (2024), 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 (ranking quality), and EMD (distributional similarity).

Our cross-encoder substantially outperforms existing attribution methods across multiple metrics (Table 1), achieving 78% precision versus 50% for bigram matching on XSum–a 56% relative improvement despite 83.82% novel bigrams in XSum summaries. Figure 1 illustrates why this accuracy matters for position bias analysis: bigram matching severely underestimates contributions from document beginnings while overestimating from endings, creating artificial position patterns. Our cross-encoder produces distributions closely aligned with human annotations, revealing that previously reported biases may partially reflect measurement artifacts rather than genuine model behaviour.

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 ($p < 0.001$).

To address potential concerns that positional patterns might reflect poor summary quality rather than meaningful content selection, we conducted comprehensive evaluations using ROUGE, BERTScore, and G-EVAL across all experimental conditions. Results consistently show high-quality

summaries across models and datasets, confirming that observed positional patterns reflect systematic content selection behaviour rather than quality artifacts (detailed results in Appendix C).

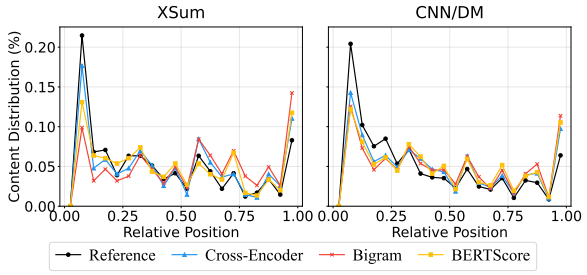

Figure 1: Position distributions by attribution method. Cross-encoder closely matches human annotations while bigram matching shows systematic distortions.

3.2 Experimental Design

Using our cross-encoder, we investigate position bias through three complementary experiments: 1) **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 random examples per dataset (CNN/DailyMail, XSum, and SAMSum), measuring how position influences selection. 3) **Long Documents:** We extend analysis to ArXiv, Multi-News, 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 for robust comparison (statistical methods detailed in Appendix F). Appendix D and Appendix E provide concrete examples of the dataset characteristics and model configurations used in our experiments. Example prompts and generation parameters can be found in Appendix H.

4 Results

4.1 Position Bias in Standard Documents

Accurate attribution reveals rightward shifts, not U-shaped bias. 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** (+0.078)	**0.37** (+0.061)	**0.43** (+0.030)
Llama-3	**0.38** (+0.056)	0.35 (+0.043)	**0.43** (+0.030)
Mistral	**0.42** (+0.098)	**0.39** (+0.083)	**0.44** (+0.033)
Phi-3	**0.40** (+0.085)	**0.40** (+0.095)	**0.45** (+0.048)
Qwen	**0.36** (+0.041)	**0.37** (+0.059)	**0.44** (+0.035)

Table 2: Mean position values across models and datasets (positions normalized to [0,1]). Bold indicates statistically significant rightward shifts compared to references ($p < 0.05$). Values in parentheses show the magnitude of shift from reference.

Table 2 quantifies these rightward shifts, with models achieving mean positions 0.041–0.098 (on [0,1]) above references in CNN/DM, similar magnitudes in XSum (+0.043 to +0.095), and smaller but consistent shifts in SAMSum (+0.030 to +0.048). This consistent pattern across all models and datasets indicating a more even distribution across positions than references, challenging assumptions of systematic positional bias.

To understand the mechanism behind these rightward shifts, we examine content selection by document sections. Figure 3 reveals that models typically extract 7-12% less content from beginning sections while incorporating 5-9% more from middle and later sections compared to human references. This redistributive pattern appears across structurally diverse content—from news articles to dialogue—confirming that models achieve more balanced document utilization rather than exhibiting positional limitations.

4.2 Context-Dependent Position Patterns

While the rightward shift appears universally, its expression varies across contexts. This variation fol-

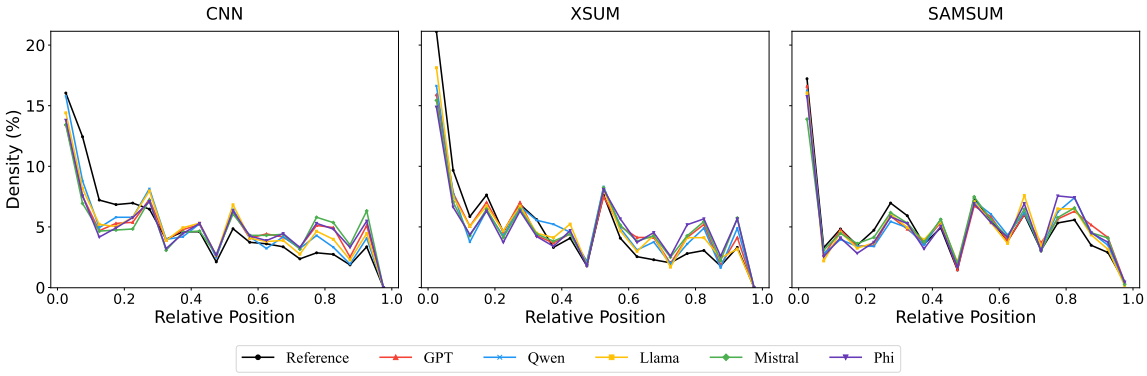

Figure 2: Position distributions comparing model-generated summaries (solid lines) with human references (dashed lines) across CNN/DailyMail, XSum, and SAMSum. Models consistently exhibit rightward shifts, selecting content from later document positions compared to human summarizers.

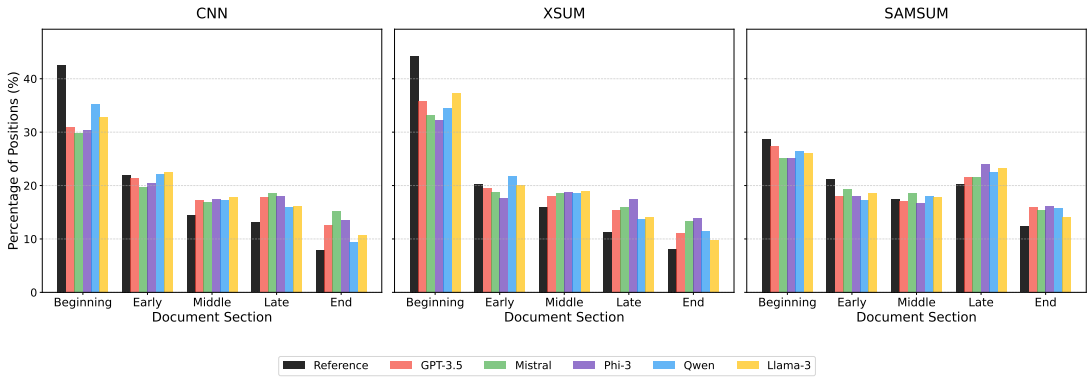

Figure 3: Content extraction by document sections. Models consistently reduce reliance on beginning sections while increasing utilization of middle and later sections compared to human references.

lows 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.

These patterns suggest models make content-based decisions that vary by context and model type, rather than showing systematic positional bias. To further validate this interpretation beyond observational evidence, our document order manipulation experiments (Section 4.3) test whether models maintain consistent selection patterns when identical content appears in different positions (Doc1+Doc2 vs. Doc2+Doc1).

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 questions: (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?

Table 3 shows that document position significantly affects selection volume. The table tracks sentences selected from Document 1 when it appears first (D1+D2) versus second (D2+D1). Most models demonstrate statistically significant position effects: positive differences indicate recency

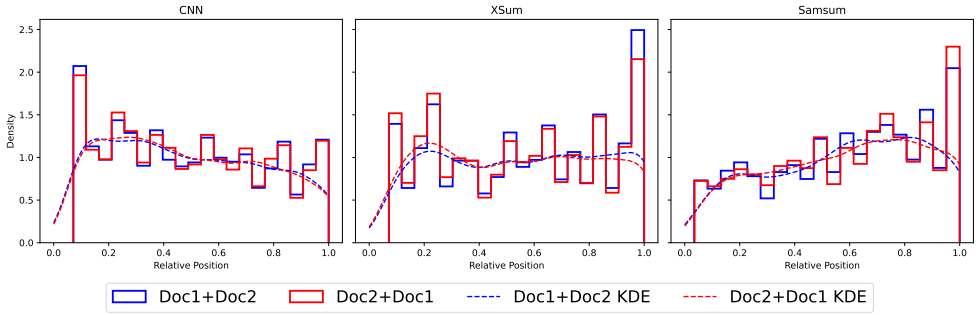

Figure 4: Llama3.2 position distributions across document configurations. The overlapping histograms demonstrate that models maintain consistent selection patterns within documents regardless of global order. Even when sentence counts differ statistically, the positions of selected content remain stable.

effects (more selection when Document 1 appears second), while negative differences indicate primacy effects. However, these effects are modest in magnitude–Mistral shows the largest effect in CNN/DM (+0.84 sentences) while GPT-3.5 shows minimal effects in SAMSum (-0.01).

Data	Model	D1+D2	D2+D1	Diff	p-value
CNN/DM	GPT-3.5	4.03	4.77	+0.74	<0.001**
	Llama-3.2	5.99	5.64	-0.35	0.002*
	Mistral	4.58	5.42	+0.84	<0.001**
	Phi-3	4.13	4.60	+0.47	<0.001**
	Qwen	4.71	5.09	+0.38	<0.001**
XSum	GPT-3.5	3.33	3.92	+0.60	<0.001**
	Llama-3.2	4.26	4.46	+0.20	0.047*
	Mistral	4.01	4.77	+0.76	<0.001**
	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	<0.001**
	Mistral	2.43	3.02	+0.59	<0.001**
	Phi-3	2.65	3.04	+0.39	<0.001**
	Qwen	2.60	2.93	+0.33	0.002**

*p < 0.05, **p < 0.001

Table 3: Sentence selection differences by document order. Each row compares how many sentences models select from documents when they appear first (D1+D2) versus second (D2+D1). Positive values indicate recency effects (preference for second document), negative values indicate primacy effects (preference for first document).

Despite significant sentence count differences, Table 4 reveals remarkable positional stabilitymodels consistently select from the same relative positions within each document regardless of global order. Of the 30 total comparisons (5 models $\times$ 3 datasets $\times$ 2 documents), 27 show no significant differences in selection patterns within documents ($p > 0.05$), representing 90% positional stability. This suggests models maintain consistent evaluation of content importance regardless of global

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 consistency within documents across configurations. P-values test whether models select sentences from the same relative positions within each document regardless of global document order. Non-significant results (p > 0.05) indicate consistent positional selection patterns.

document ordering.

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.

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. Our extended context analysis examines whether this position-independent evaluation extends to substantially longer documents and multi-document scenarios. To test whether

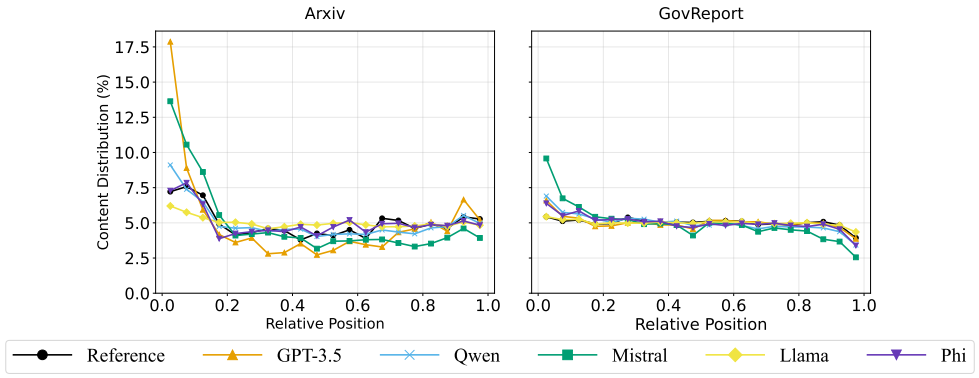

Figure 5: Document-type-dependent extraction patterns. Scientific papers show pronounced boundary bias with models over-selecting from document beginnings and ends, while government documents exhibit more uniform extraction similar to human patterns.

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 metrics across long document types. Lower values indicate closer alignment with human references. Results reveal document-structure dependency over model-size effects, with the same models showing vastly different behaviours across domains.

this position-independent evaluation extends to extended documents and multi-document scenarios, we analyze three challenging datasets with significantly longer contexts (Section 4.4).

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 content and 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 shows pronounced beginning bias, selecting 38% of summary content from the first document section versus only 27% for human references.

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 ap-

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 ± 0.187
Phi-3	0.452 (0.455)	0.460 (0.466)	3.80 (3.83)	0.028 (0.055)	0.817 ± 0.222
Llama-3	0.473 (0.471)	0.459 (0.465)	3.71 (3.83)	0.027 (0.001)	0.904 ± 0.146
Qwen	0.447 (0.436)	0.456 (0.464)	3.77 (3.83)	0.028 (0.007)	0.915 ± 0.141
Mistral	0.440 (0.381)	0.509 (0.537)	3.27 (3.66)	0.216 (0.006)	0.768 ± 0.234

Table 6: Evidence against "lost-in-the-middle" effects in multi-document summarization. Models achieve balanced global position distributions (medians 0.5), distributed source attention (high entropy), and maintain high summary quality, contradicting middle-position processing limitations.

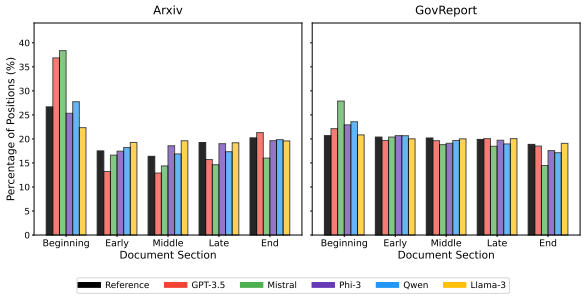

Figure 6: Content extraction by document sections. Scientific papers show boundary bias (high beginning/end extraction), than government documents.

pears 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, suggesting 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" claimsmodels 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

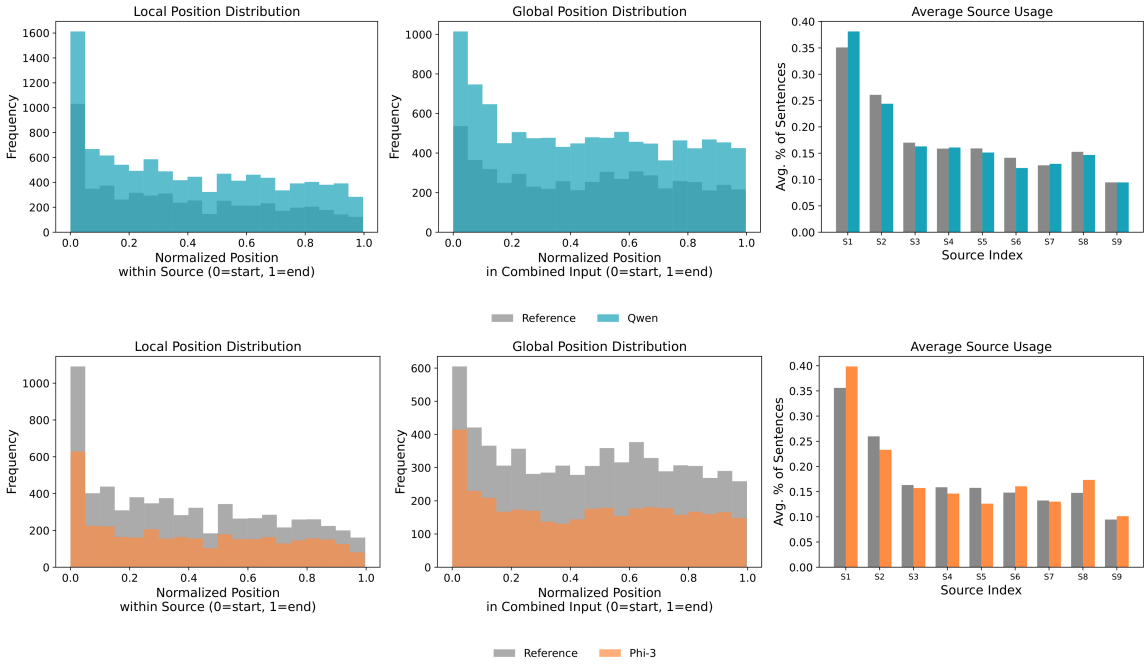

Figure 7: Multi-document position analysis for Qwen and Phi models, which successfully extract content from global middle positions, with balanced local and global position distributions.

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.

Acknowledgments

Rudzicz is supported by a Canada CIFAR Chair in AI, and a Killam Memorial Chair.

References

Anshuman Chhabra, Yang Liu, Xiang Liu, Ani Nenkova, Alan Ritter, and Zhou Yu. 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 on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. 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. 2022. News summarization and evaluation in the era of GPT-3. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3680–3695, Abu Dhabi, United Arab Emirates. 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.

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.

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

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*.

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.

Yoshi Suhara and Dimitris Alikaniotis. 2024. 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. Julians, 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.

Jiacheng Xu and Greg Durrett. 2021. Dissecting generation modes for abstractive summarization models via ablation and attribution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3717–3732. Association for Computational Linguistics.

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 Threshold Selection Validation

To validate our adaptive elbow point thresholding strategy, we conducted comparative evaluations against alternative threshold selection methods. Our approach addresses a fundamental challenge in cross-encoder attribution: determining which source sentences should be considered contributing to summary content based on their semantic similarity scores.

A.1 The Selection Challenge

Given a summary sentence and a set of document sentences, the cross-encoder produces similarity scores for each summary-source pair. The critical question becomes: which scores are high enough to indicate genuine semantic contribution versus mere topical overlap? This threshold selection directly determines attribution accuracy, as overly permissive thresholds include irrelevant sentences while overly restrictive thresholds miss valid contributions.

Attribution scores vary dramatically across summarization instanceshighly abstractive summaries exhibit uniformly low cross-encoder scores due to extensive paraphrasing, while extractive summaries show clear high-low separation patterns. Fixed thresholds fail to accommodate this variation.

A.2 Threshold Selection Methods

We evaluated four strategies for determining which cross-encoder scores indicate contributing sentences:

- **Elbow point method (ours):** For each summary sentence, we rank all document sentences by their cross-encoder scores, identify where score differences drop most sharply (the "elbow"), then select sentences that both (1) score above this elbow point and (2) exceed $\mu + 0.5\sigma$ of all scores. This adapts to instance-specific score distributions.

- **75th percentile:** Selects document sentences scoring above the 75th percentile of all cross-encoder scores for that summary sentence.

- **80th percentile:** Selects document sentences scoring above the 80th percentile of all cross-encoder scores for that summary sentence.

- **Otsu's method:** Applies Otsu's automatic threshold selection, treating attribution as a binary classification problem between contributing and non-contributing sentences.

A.3 Comparative Results

Table 7 presents comparative results against human-annotated ground truth from Suhara and Alikaniotis (2024). Our elbow point method achieves superior performance across all metrics: highest precision (0.776) and NDCG (0.857), and lowest Earth Mover's Distance (0.054). The results demonstrate that adaptive instance-specific thresholding better identifies genuinely contributing sentences compared to fixed statistical approaches.

Method	Precision	NDCG	EMD
Elbow point (ours)	**0.776**	**0.857**	**0.054**
75th percentile	0.663	0.809	0.074
80th percentile	0.707	0.819	0.060
Otsu's method	0.653	0.849	0.086

Table 7: Threshold selection method comparison on human-annotated attribution data. Bold values indicate best performance. Results show that adaptive thresholding outperforms fixed approaches for identifying semantically contributing sentences.

The superior performance of our elbow point method validates our hypothesis that effective source attribution requires adaptive thresholding that responds to the varying semantic similarity distributions across different summarization styles and content types.

B Cross-Encoder Implementation Details

B.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.

B.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 μ and σ 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.

B.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.

C Summary Quality Evaluation

To address potential concerns that positional patterns might reflect poor summary quality rather than meaningful content selection, we conducted comprehensive quality evaluations using ROUGE, BERTScore, and G-EVAL across all experimental conditions. This validation ensures that our position bias findings reflect genuine content selection behaviour rather than artifacts of low-quality summaries.

C.1 Evaluation Metrics

We employed three complementary evaluation approaches:

- **ROUGE:** Measures lexical overlap with human references through n-gram matching (ROUGE-1, ROUGE-2, ROUGE-L)
- **BERTScore:** Evaluates semantic similarity using contextual embeddings, better capturing paraphrases than ROUGE
- **G-EVAL:** GPT-4-based evaluator using Chain-of-Thought prompting to assess coherence, consistency, fluency, relevance, and completeness on a 5-point scale

C.2 Results

Table 8 presents ROUGE and BERTScore evaluations across all datasets and models. While ROUGE-2 scores are predictably low on abstractive datasets like XSum due to extensive paraphrasing, BERTScore remains consistently high ($> 87\%$

Table 9 presents G-EVAL assessments, which provide more nuanced quality evaluation through multiple dimensions. Most models achieve overall scores ≥ 4.0 across datasets, indicating high-quality summaries suitable for attribution analysis. The detailed dimensional scores reveal that models excel particularly in fluency and relevance (often > 4.5), while completeness scores are more modest (3.0–4.0), reflecting the inherent compression in summarization tasks.

C.3 Quality Validation

These comprehensive evaluations confirm that observed positional patterns reflect systematic content selection behaviour rather than quality artifacts. The consistently high BERTScore values demonstrate semantic coherence, while G-EVAL's dimensional analysis reveals that models maintain strong performance across key quality indicators. Notably, even when ROUGE scores vary (reflecting different degrees of abstractiveness), semantic quality metrics remain robust, validating our attribution-based position analysis.

The quality results support our core finding that apparent position "biases" represent sophisticated content-driven selection strategies rather than limitations in summary generation capability.

D 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

Dataset	Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
XSum	GPT-3.5	25.94	6.50	17.52	87.80
	Llama-3	27.44	6.78	19.27	87.83
	Mistral	27.94	7.43	19.24	87.93
	Qwen	27.75	6.86	18.97	88.13
	Phi-3	25.43	5.52	17.18	87.57
CNN/DM	GPT-3.5	40.84	16.79	36.56	88.18
	Llama-3	38.90	15.91	35.08	87.84
	Mistral	39.17	15.49	35.21	87.92
	Qwen	40.67	15.55	35.78	88.16
	Phi-3	39.66	14.77	34.81	88.06
SAMSum	GPT-3.5	38.62	14.34	33.16	89.76
	Llama-3	36.38	13.11	31.78	89.47
	Mistral	34.62	12.28	30.62	89.11
	Qwen	38.89	14.27	33.20	89.68
	Phi-3	35.50	11.86	31.31	89.31
ArXiv	GPT-3.5	41.35	13.26	37.21	83.98
	Llama-3	27.95	9.38	25.56	81.49
	Mistral	38.16	14.86	34.92	83.43
	Qwen	43.34	16.41	39.33	84.75
	Phi-3	43.34	16.41	39.33	84.75
GovReport	GPT-3.5	31.06	12.33	29.34	85.76
	Llama-3	50.36	17.96	48.22	84.44
	Mistral	50.29	19.87	47.46	85.89
	Qwen	49.62	17.77	47.44	84.29
	Phi-3	41.97	16.69	39.82	85.87
Multi-News	GPT-3.5	36.88	9.72	33.65	85.13
	Llama-3	38.59	10.31	35.47	83.82
	Mistral	41.09	13.12	38.07	85.25
	Qwen	40.29	10.77	36.94	84.28
	Phi-3	34.18	9.06	30.80	84.94

Table 8: ROUGE and BERTScore evaluations across all experimental conditions. BERTScore remains consistently high across datasets, indicating strong semantic quality despite varying lexical overlap patterns.

Dataset	Model	Overall Mean (SD)	Coherence Mean (SD)	Consistency Mean (SD)	Fluency Mean (SD)	Relevance Mean (SD)	Completeness Mean (SD)
XSum	GPT-3.5	4.19 (0.36)	4.03 (0.28)	4.72 (0.47)	4.11 (0.34)	4.23 (1.48)	3.88 (0.44)
	Llama-3	3.61 (0.70)	3.58 (0.66)	3.82 (1.06)	3.81 (0.51)	3.81 (1.30)	3.04 (0.63)
	Mistral	4.18 (0.34)	4.01 (0.25)	4.58 (0.56)	4.12 (0.32)	4.46 (1.14)	3.75 (0.45)
	Phi-3	4.02 (0.43)	3.81 (0.44)	4.35 (0.75)	3.96 (0.37)	4.51 (1.00)	3.44 (0.59)
	Qwen	4.18 (0.42)	3.98 (0.30)	4.59 (0.63)	4.13 (0.39)	4.35 (1.36)	3.84 (0.43)
CNN/DM	GPT-3.5	4.39 (0.30)	4.18 (0.38)	4.45 (0.50)	4.92 (0.28)	5.00 (0.00)	3.70 (0.47)
	Llama-3	4.59 (0.33)	4.64 (0.48)	4.71 (0.49)	4.98 (0.14)	4.98 (0.12)	3.80 (0.40)
	Mistral	4.36 (0.31)	4.17 (0.37)	4.41 (0.51)	4.85 (0.36)	4.98 (0.12)	3.68 (0.47)
	Phi-3	4.22 (0.32)	4.10 (0.32)	4.24 (0.49)	4.88 (0.32)	4.95 (0.22)	3.41 (0.50)
	Qwen	4.21 (0.29)	4.08 (0.27)	4.21 (0.44)	4.94 (0.23)	4.98 (0.12)	3.40 (0.49)
SAMSum	GPT-3.5	4.49 (0.35)	4.44 (0.50)	4.54 (0.51)	4.97 (0.18)	4.98 (0.13)	3.76 (0.51)
	Llama-3	4.48 (0.42)	4.40 (0.56)	4.52 (0.58)	4.97 (0.24)	4.96 (0.33)	3.78 (0.49)
	Mistral	4.57 (0.32)	4.53 (0.50)	4.64 (0.49)	4.88 (0.32)	4.99 (0.09)	3.91 (0.46)
	Phi-3	4.48 (0.35)	4.43 (0.50)	4.50 (0.55)	4.95 (0.23)	4.97 (0.17)	3.78 (0.49)
	Qwen	4.48 (0.33)	4.41 (0.49)	4.53 (0.53)	4.93 (0.26)	4.97 (0.16)	3.78 (0.45)

Table 9: G-EVAL quality assessments across standard-length datasets. Values show mean scores with standard deviations in parentheses. High overall scores (4.0) and strong performance in fluency and relevance dimensions confirm summary quality sufficient for reliable attribution analysis.

standard-length documents (142-656 tokens) and extended contexts (2,103-8,912 tokens) to test the scalability of position patterns. **(2) Domain Coverage:** Our datasets span news (CNN/DM, XSum), dialogue (SAMSum), scientific writing (ArXiv), government documents (GovReport), and multi-

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 10: Key dataset characteristics for position bias analysis

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

Table 11: Large Language Model specifications and configurations

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 10.

E 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 11.

F Statistical Testing Procedures

This appendix provides detailed explanations of the statistical methods used throughout our analysis to ensure reproducibility and methodological transparency.

F.1 Statistical Test Procedure

In our document order manipulation experiments (Doc1+Doc2 vs. Doc2+Doc1), we test whether the positions from which models select content are affected by document ordering:

1. For each **document pair** in the, we record the **sentence count** and calculate the **average position** of sentences selected from Doc1 in both input orderings (first vs second position)

2. We apply a **paired t-test** (`scipy.stats.ttest_rel`) to compare these metrics across the two orderings for Doc1 only

3. This yields **two p-values**: one for sentence count differences and one for position differences when Doc1 appears first vs second

4. Non-significant results ($p > 0.05$) indicate positional stability for Doc1 selection patterns

F.2 Position Distribution Comparisons

Throughout our analysis, we employ multiple statistical tests for robust comparison of position distributions:

- **Kolmogorov-Smirnov test:** Tests whether two distributions differ significantly
- **Two-sample t-test:** Parametric test for differences in distribution means
- **Jensen-Shannon divergence:** Symmetric measure of distributional similarity
- **Wasserstein distance:** Earth Mover's Distance measuring distributional dissimilarity
- **Kullback-Leibler divergence:** Asymmetric measure of distributional difference

Significance levels are set at $\alpha = 0.05$

F.3 Interpreting Relative Divergence Metrics

We acknowledge that divergence metrics (KL, KS, JS, WD) provide relative rather than absolute measures of alignment. These metrics lack universal thresholds defining "acceptable" or "problematic" performance. However, they become interpretable through systematic comparison against human reference distributions.

Framework for Interpretation Our interpretation framework relies on three complementary approaches:

Comparative Analysis: Lower divergence values indicate closer alignment with human position selection patterns, while higher values suggest greater deviation. For instance, on ArXiv, Phi-3 achieves KL = 0.004 and KS = 0.019, demonstrating strong alignment, whereas GPT-3.5 (KL = 0.078, KS = 0.123) shows substantially higher divergence.

Statistical Significance: We use KS test p-values (< 0.05) to identify statistically meaningful deviations from human patterns, distinguishing systematic bias from random variation.

Cross-Dataset Validation: Models showing high divergence on one dataset may perform differently on others. GPT-3.5 and Mistral exhibit significant divergence on ArXiv but demonstrate more balanced patterns on Multi-News and GovReport, indicating dataset-specific rather than model-inherent limitations.

Contextual Attribution Analysis To avoid over-interpretation of divergence metrics alone, we complement statistical measures with detailed section-level attribution analysis (§4.4.1, Figure 6). This reveals that higher ArXiv divergence stems from systematic under-selection of middle sectionsa specific, interpretable pattern rather than general summarization failure.

Our approach thus provides **nuanced interpretation**: divergence metrics identify where and when models deviate from human patterns, while attribution analysis explains why these deviations occur and whether they represent meaningful limitations.

F.4 Multi-Document Source Selection Analysis

F.4.1 Jaccard Similarity Computation

For Multi-News experiments involving multiple source documents, we compute Jaccard similarity between the set of source indices contributing to generated summaries and those in reference summaries:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where A is the set of source document indices used by the model and B is the set used in the human reference.

Example: If a generated summary uses sources $\{0, 1, 3\}$ and the reference uses $\{1, 2, 3, 4\}$, the Jaccard similarity is $\frac{|\{1,3\}|}{|\{0,1,2,3,4\}|} = \frac{2}{5} = 0.4$.

This metric quantifies overlap in source selection behaviour between models and human annotators, providing insight into content aggregation strategies in multi-document summarization.

This comprehensive statistical framework ensures that our findings reflect both statistical reliability and practical significance in understanding position bias patterns.

G 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.

H Experimental Configuration

H.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}`
- **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.

H.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

H.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.