ARC: Argument Representation and Coverage Analysis for Zero-Shot Long Document Summarization with Instruction Following LLMs

Anonymous EMNLP submission

Abstract

Integrating structured information has long improved the quality of abstractive summarization, particularly in retaining salient content. In this work, we focus on a specific form of structure: **argument roles**, which are crucial for summarizing documents in high-stakes domains such as law. We investigate whether instruction-tuned large language models (LLMs) adequately preserve this information. To this end, we introduce **A**rgument **R**epresentation **C**overage (ARC), a framework for measuring how well LLM-generated summaries capture salient arguments. Using ARC, we analyze summaries produced by three open-weight LLMs in two domains where argument roles are central: *long legal opinions* and *scientific articles*. Our results show that while LLMs cover salient argument roles to some extent, critical information is often omitted in generated summaries, particularly when arguments are sparsely distributed throughout the input. Further, we use ARC to uncover behavioral patterns—specifically, how the positional bias of LLM context windows and role-specific preferences impact the coverage of key arguments in generated summaries, emphasizing the need for more argument-aware summarization strategies.

1 Introduction

LLMs have made remarkable progress in text summarization, often generating summaries preferred by human evaluators in domains like news (Zhang et al., 2024; Liu et al., 2024b). However, summarization in structured, highly informative domains presents unique challenges that remain underexplored. One prominent challenge is preserving salient argument roles in the generated summaries, which can be challenging as arguments can be sparsely distributed across the input (Elaraby and Litman, 2022). The ability of LLMs to selectively retain argument roles is therefore a crucial test of their utility in generating reliable summaries in high-stakes domains.

Prior work in legal summarization has shown that explicitly modeling argument roles—either during finetuning (Fabbri et al., 2021a; Elaraby and Litman, 2022) or through post hoc re-ranking (Elaraby et al., 2023)—improves the coverage of critical argumentative content. These findings suggest that pretrained language models (PLMs) may struggle to capture structured discourse elements such as arguments without targeted supervision. However, it remains an open question whether instruction-tuned LLMs, trained with broad and often general-purpose supervision, can inherently identify and preserve salient argumentative information without explicit signals about saliency or additional tuning. In this work, we go beyond conventional summary evaluation metrics such as fluency, factuality, and coherence to address a core question: *Do LLMs effectively prioritize and preserve the most salient argumentative content in their summaries?*

To address this question, we introduce **A**rgument **R**epresentation **C**overage (ARC), a framework for evaluating how well LLM-generated summaries capture salient arguments. ARC measures coverage at three levels of granularity: (1) **Argument Set Coverage (ARC$_{\text{fullset}}$)**, assessing collective coverage of full salient argument set; (2) **Independent Role Coverage (ARC$_{\text{role}}$)**, evaluating each argument role separately as atomic units; and (3) **Subatomic Coverage (ARC$_{\text{atomic}}$)**, examining fine-grained factual units within roles. Figure 1 presents an example of how ARC operates across multiple levels of abstraction. While ARC$_{\text{fullset}}$ provides a single holistic score capturing overall argument coverage, it fails to offer fine-grained insight into which specific arguments are preserved or omitted—information critical for downstream analyses such as bias detection. ARC$_{\text{role}}$ addresses this by assessing the preservation of each argument role

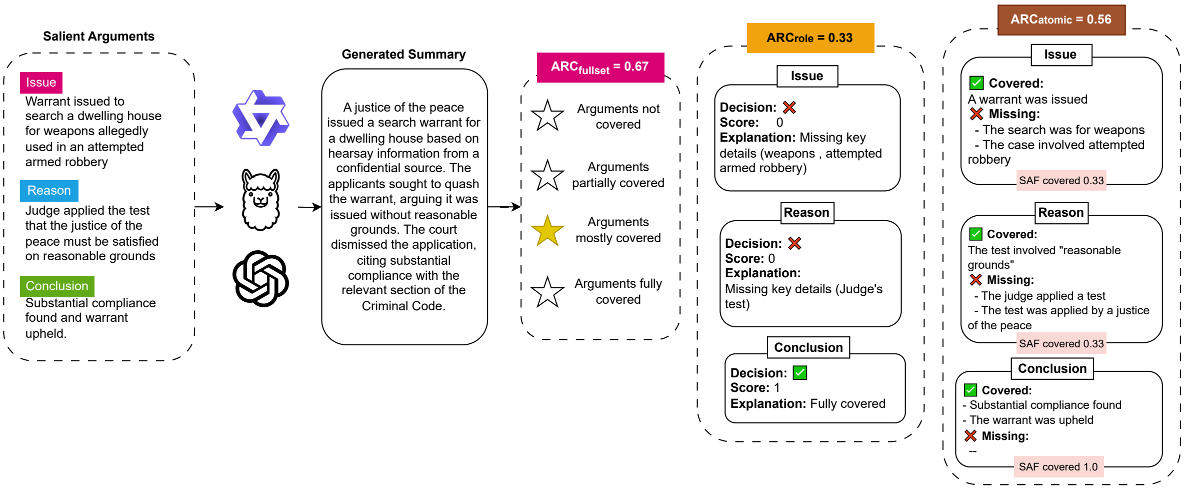
Salient Arguments
Issue
Warrant issued to search a dwelling house for weapons allegedly used in an attempted armed robbery
Reason
Judge applied the test that the justice of the peace must be satisfied on reasonable grounds
Conclusion
Substantial compliance found and warrant upheld.
Generated Summary
A justice of the peace issued a search warrant for a dwelling house based on hearsay information from a confidential source. The applicants sought to quash the warrant, arguing it was issued without reasonable grounds. The court dismissed the application, citing substantial compliance with the relevant section of the Criminal Code.
ARCfullset = 0.67
Arguments not covered
Arguments partially covered
Arguments mostly covered
Arguments fully covered
ARCrole = 0.33
Issue
Decision: ✗
Score: 0
Explanation: Missing key details (weapons , attempted armed robbery)
Reason
Decision: ✗
Score: 0
Explanation:
Missing key details (Judge's test)
Conclusion
Decision: ✓
Score: 1
Explanation: Fully covered
ARCatomic = 0.56
Issue
✓ Covered:
A warrant was issued
✗ Missing:
- The search was for weapons
- The case involved attempted robbery
SAF covered 0.33
Reason
✓ Covered:
The test involved "reasonable grounds"
✗ Missing:
- The judge applied a test
- The test was applied by a justice of the peace
SAF covered 0.33
Conclusion
✓ Covered:
- Substantial compliance found
- The warrant was upheld
✗ Missing:
--
SAF covered 1.0

Figure 1: Examples of ARC scores at multiple granularities—fullset, role, and atomic (where SAF denotes subatomic facts)—for a summary generated by LlaMA3.18B on a case from the long-legal opinions dataset.

independently, although it considers an argument entirely unsupported if even one element is missing. To mitigate this limitation, $\texttt{ARC}_{\text{atomic}}$ decomposes arguments into subatomic factual units, enabling a more continuous and informative scoring mechanism. This multi-level approach is designed to overcome the limitations of coarse Likert-scale ratings, which have been shown to suffer from low inter-annotator agreement in prior work (Elaraby et al., 2023; Krishna et al., 2023a).

Prior work highlights that LLMs often exhibit positional biases—favoring content from the beginning or end of long documents (Liu et al., 2024a; Ravaut et al., 2024a; Wan et al., 2024). Furthermore, it remains unclear whether LLMs also favor certain types of salient information that share the same structure (e.g., argument roles) during summarization. Using ARC, we conduct two key analyses: (1) *How does the position of arguments in the source document affect their inclusion in summaries?* and (2) *Are certain argument roles disproportionately favored over others?* The latter analysis is crucial, as uneven role coverage may lead to biased or incomplete summaries that misrepresent the original discourse. We conduct our experiments across two domains where argument structure plays a central role in understanding the essence of the document: *Long Legal Opinions* and *Scientific Articles*.

Our contributions are in two folds: (1) we propose ARC, a multi-granularity evaluation framework for evaluating argument coverage; and (2) we present a systematic analysis of how positional and role-specific biases affect LLMs' ability to preserve salient argumentative content in summaries.

2 Related Work

Information Saliency in LLMs. Content selection remains a core challenge in summarization. Trienes et al. (2025) found weak alignment between LLMs' saliency preferences and human judgments. While LLMs can produce summaries preferred over human references in domains like news (Zhang et al., 2024; Liu et al., 2024b), they still benefit from content planning. For example, Adams et al. (2023) showed that planning entity mentions improves the information density in GPT-4 generated summaries at the same summary length when compared to summaries generated without entitiy planning. *We extend this line of work by treating argument roles as a structured form of saliency and analyzing their preservation in LLM-generated summaries.*

Limitations of LLMs in Long-Document Summarization. LLMs face persistent issues when summarizing long texts, notably the U-shaped positional bias—favoring content at the beginning and end while neglecting the middle (Ravaut et al., 2024b). This leads to degraded faithfulness in long-form outputs (Wan et al., 2024). *We expand this analysis by quantifying how positional bias affects the coverage of salient argumentative content.*

Argument Mining and Abstractive Summarization. Incorporating argument structures into summarization has shown promise across domains,

Dataset	# Docs	Input Length	Summary Length	% Roles in Input	% Roles in Summary
CANLII	1049	122/4382/62786	17/273/2072	7.66%	66.51%
DRI	40	3460/6505/11679	67/221/298	74.14%	-

Table 1: Statistics of the datasets, including the number of documents, input document length, reference summary length (min/mean/max words), and the percentage of argument roles in the input and summary. - indicates that value can't be directly computed from the corpus.

including dialogues (Fabbri et al., 2021a), legal texts (Xu et al., 2020, 2021; Elaraby and Litman, 2022; Elaraby et al., 2023) and scientific documents (Fisas Elizalde et al., 2016). *We build on this by assessing whether instruction-tuned LLMs can cover salient arguments without the external argument role information, particularly in summarizing legal and scientific texts.*

Evaluation Metrics for Long-Form Summarization. Standard metrics often fall short in reflecting human preferences, especially for long documents (Fabbri et al., 2021b; Krishna et al., 2023b). To improve reliability, recent work has introduced unit-based metrics—such as Atomic Content Units (ACUs) (Krishna et al., 2023a) and structured factuality scores (Min et al., 2023; Yang et al., 2024)—to reduce subjectivity. *Extending this idea, we propose* ARC*, which uses argumentative structures as evaluation units and introduces subatomic granularity to assess fine-grained argument coverage.*

3 Datasets[1]

We employ two datasets that include both argument role annotations and reference summaries: CANLII (Xu et al., 2021), representing the legal domain, and DR. INVENTOR (DRI) (Fisas Elizalde et al., 2016), representing the scientific domain. An overview of dataset statistics is presented in Table 1. Both datasets consist of long-form documents paired with long-form reference summaries that average > 150 words (Krishna et al., 2023b).

3.1 Legal Opinions: CANLII

The CANLII dataset consists of 1049 legal cases annotated at the sentence level for argument roles. Notably, only 7.66% of the input text is labeled with argument roles, yet these argumentative sentences account for 66.51% of the reference summaries (Table 1). This substantial mismatch highlights a haystack-like challenge: models must accurately identify and prioritize the sparse yet highly salient argumentative content when generating summaries.

Dataset	Argument Role	Example
CANLII	**Issue**	Damage to both vehicles exceeded the insurance deductibles and both parties claim damages against each other.
	Conclusion	Fault for this accident was attributed 10% to the defendant and 90% to the plaintiff.
	Reason	Jurisdictional error is not to be equated with error of law.
DRI	**Own Claim**	Semi-Lagrangian contouring offers an elegant and effective means for surface tracking with advantages over competing methods.
	Background Claim	Accurate modeling of human motion remains a challenging task.
	Data	Animation is constrained due to hardware constraints.

Table 2: Examples of argument roles from CANLII (legal domain) and DRI (scientific domain). Colors distinguish different argument roles.

Argument roles in CANLII are annotated using the **IRC** scheme (Xu et al., 2021), which categorizes roles into three types: **Issues:** *Legal questions raised in the case.* **Reasons:** *Justifications provided for judicial decisions.* **Conclusions:** *Final rulings addressing the identified issues.* These annotations enable a fine-grained evaluation of argument coverage in generated summaries. Examples of IRCs are shown in Table 2.

3.2 Scientific Articles: DRI

The DRI dataset consists of 40 computer graphics articles, each annotated at the sentence level for 5 rhetorical roles and paired with three human-written summaries. Notably, these rhetorical roles are not necessarily argumentative, making it challenging to assess argument coverage. To address this limitation, the extended version of the dataset, SCI-ARG (Lauscher et al., 2018), enriches the DRI annotations by incorporating argument role annotations and their relations. These annotations follow a modification of the 6-argument roles described in *Toulmin model* (Toulmin, 2003), by reducing them

[1] More analysis and examples in Appendix A.

into three types: **Own Claim:** *Sentences that directly support the author's central argument.* **Background Claim:** *Sentences that reference prior research or established domain knowledge.* **Data:** *Empirical evidence that supports or refutes claims, such as experimental results or literature citations.* An example of each role is shown in Table 2.

Since the argument annotations are span-based, we map them back to complete sentences using lexical matching assigning the sentence with argument role spans if $> 50\%$ of its words falls within the sentence boundaries. A motivating feature behind selecting this corpus for our analysis is the sentence-level annotation for relevance scores on a Likert scale $(1 - 5)$, which indicates the degree of relevance to the summary. A relevance score of 4 signifies that the sentence is "relevant to the summary," while a 5 indicates it is "very relevant to the summary." In our evaluation, we focus on argument role coverage for sentences with argument roles and a Likert score of 5 (indicating high relevant argument roles to the summary). Table 1 shows that unlike legal opinions, where argument roles are sparsely distributed, scientific articles contain argumentative content throughout the document, posing a challenge of selectivity rather than retrieval. In `DRI`, sentences that contain at least 1 argument role account for 74.14% of the input text (shown in Table 1). Although the dataset does not provide gold-standard summaries annotated for argument roles, we analyze the sentences with a Likert score of 5 based on their argument role annotation. Among these sentences, 91.74% contain at least one argument role, reinforcing the strong connection between argument roles and summarization relevance in this domain.

4 The ARC framework

4.1 Overview and Notations

Given a generated summary S and a set of salient arguments $\mathbb{A} = \{a_1, a_2, \ldots, a_n\}$, we define a coverage function Φ, where: $0 \leq \Phi(v, S) \leq 1$. Here, v represents the evaluation unit against S, where: $v \in \{\mathbb{A}, a_i, m_i\}$. $\mathbb{A}$ represents the full set of salient arguments in the document. a_i is an individual argument $a_i \in \mathbb{A}$. m_i is a subatomic factual unit, where $m_i \in M_i$ and M_i is the set of atomic facts derived from argument a_i.

Full Argument Set Coverage Following Elaraby et al. (2024), we define $\Phi(\mathbb{A}, S)$ based on a Likert-scale annotation (1-4), which is then normalized to a $[0, 1]$ range: $\Phi(\mathbb{A}, S) = \frac{\ell(\mathbb{A},S)-1}{3}$

where $\ell(\mathbb{A}, S)$ is the Likert score assigned to the argument set coverage.

Independent Argument Role Coverage Each argument a_i is evaluated independently as:

$$\Phi(a_i, S) = \begin{cases} 1, & \text{if } a_i \text{ is fully preserved in } S \\ 0, & \text{if } a_i \text{ is partially/fully omitted or distorted in } S \end{cases}$$

Subatomic Argument Coverage An argument $a_i \in \mathbb{A}$ is further decomposed into subatomic fact units M_i, where for $m_j \in M_i$:

$$\Phi(m_i, S) = \begin{cases} 1, & \text{if } m_i \text{ is fully supported in } S \\ 0, & \text{if } m_i \text{ is missing or unfaithfully represented in } S \end{cases}$$

The scores obtained by Φ at different granular levels provide a deeper assessment of model behavior on the argument level. Table 3 shows `ARC` scores computation for an n number of salient arguments.

Granularity	Computation
Full Set	$\text{ARC}_{\text{fullset}}(S) = \Phi(\mathbb{A}, S)$
Individual Roles	$\text{ARC}_{\text{role}}(S) = \frac{1}{n}\sum_{i=1}^{n} \Phi(a_i, S)$
Subatomic Units	$\text{ARC}_{\text{atomic}}(S) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{M_i} \sum_{m_j \in M_i} \Phi(m_j, S)$

Table 3: `ARC` scores at different granularity levels.

4.2 Choice of Φ

Given the limited availability of large, diverse argument coverage datasets, we use the annotated dataset from Elaraby et al. (2024), comprising 90 legal opinions from the `CANLII` dataset with corresponding summaries. Each summary is annotated for argument coverage using a 4-point Likert scale at the full argument set level.[2]

Computing Φ using `LLM-judge` We use `GPT-4o` with 0 temperature sampling as an automated evaluator, following prior work that demonstrated its strong correlation with human judgments in summarization tasks (Liu et al., 2023b).

ARC$_{\text{fullset}}$: We prompt the model to assign a Likert score (1-4) based on coverage guidelines from Elaraby et al. (2024). Scores are then normalized to $[0, 1]$ for comparability with other metrics.

ARC$_{\text{role}}$: The model assigns binary scores (1 = fully supported, 0 = missing or inconsistent) for each argument $a_i \in \mathbb{A}$.

[2] See Appendix B for scale definitions.

Metric	Metric Type	Expert 1		Expert 2		Average	
		τ	ρ	τ	ρ	τ	ρ
ROUGE-1	Lexical	0.3455	0.4809	0.3382	0.4495	0.3757	0.5318
ROUGE-2		0.3233	0.4150	0.2860	0.4132	0.3291	0.4734
ROUGE-L		0.2824	0.3874	0.3463	0.4455	0.3418	0.4764
BERTScore	Semantic	0.3187	0.4092	0.2921	0.3977	0.3344	0.4756
SummaC$_{\text{ZS}}$ (sent)	Entailment	0.3676	0.5157	0.3567	0.4368	0.3204	0.4654
SummaC$_{\text{ZS}}$ (doc)		0.4512	0.4894	0.2617	0.3438	0.3747	0.4758
SummaC$_{\text{conv}}$ (sent)		0.3204	0.4654	0.3567	0.4368	0.3676	0.5157
SummaC$_{conv}$ (doc)		0.4512	0.4894	0.2617	0.3438	0.3747	0.4758
ARC$_{\text{fullset}}$($\Phi_{\texttt{GPT4-o}}$)	LLM-Judge	**0.6713**	**0.7288**	0.4072	0.4793	0.5867	0.6898
ARC$_{\text{role}}$($\Phi_{\texttt{GPT4-o}}$)		0.4884	0.6034	0.3453	0.4527	0.4474	0.5971
ARC$_{\text{atomic}}$($\Phi_{\texttt{GPT4-o}}$)		0.5806	0.7023	**0.5135**	**0.6353**	**0.6025**	**0.7560**
ARC$_{\text{role}}$($\Phi_{\texttt{DeBERTa}}$)		0.5025	0.6026	0.4593	0.5727	0.5142	0.6642
ARC$_{\text{atomic}}$($\Phi_{\texttt{DeBERTa}}$)		0.5213	0.6507	0.4347	0.5304	0.5202	0.6959

Table 4: Correlations between automatic metrics and expert judgments (τ: Kendall's tau, ρ: Pearson's r; all values statistically significant at $p < 0.01$, normalized to $[0, 1]$). **Bold** indicates highest correlation in each column.

ARC$_{\text{atomic}}$: Building on recent factuality frameworks that decompose summaries into atomic fact units for grounded evaluation (Min et al., 2023; Lee et al., 2024; Tang et al., 2024; Yang et al., 2024), ARC$_{\text{atomic}}$ instead decomposes key arguments need to be retained. This shift focuses evaluation on the *completeness* of argument roles rather than factuality evaluation. Unlike prior completeness work, which treats all information uniformly, argument roles impose a structured prioritization, which is sparse in legal cases and dense in scientific articles, enabling more targeted analysis to indicate if it's still worthy including structured information such as argument roles for a more complete summary.

Following the decomposition method of Yang et al. (2024), each argument a_i is decomposed into a set of factual units $M_i = \{m_1, ..., m_j\}$ using GPT-4o. Units are filtered by a fine-tuned entailment model (DeBERTa; (He et al.))[3] to keep only entailed facts[4]. Each unit m_i is then verified against the summary S, yielding (d, e), where $d \in \{0, 1\}$ indicates support and e specifies the error type *(missing or non-factual)*. This enables fine-grained argument coverage evaluation while distinguishing hallucination from information loss[5].

Reducing LLM-Judge Cost Evaluating argument coverage at scale using GPT-4o is costly, particularly for the legal dataset (CANLII), where each document contains numerous argument units and their subatomic facts. To mitigate this cost, we train classifiers to approximate GPT-4o judgments, thereby reducing reliance on LLM-based evaluation.

We sample 100 cases from CANLII and use GPT-4o to compute ARC$_{role}$ and ARC$_{atomic}$ for all models. This data serves as the training set for two classifiers. The first classifier, C_{role}, is a binary classifier that predicts argument-level support, trained to approximate the judgments of GPT-4o such that $\Phi_{C_{\text{role}}(a_i,S)} \approx \Phi_{\texttt{GPT-4o}(a_i,S)}$. The second classifier, C_{atomic}, is a three-way classifier that predicts whether a subatomic fact is *supported*, *missing*, or *non-factual*, ensuring that $\Phi_{C_{\text{atomic}}(m_j,S)} \approx \Phi_{\texttt{GPT-4o}(m_j,S)}$.

We explored several models, including DeBERTa (base and large), LegalBERT (Chalkidis et al., 2020), and BigBirdRoBERTa (Zaheer et al., 2021). Among all models, the DeBERTa-large demonstrated the highest performance, achieving an F1-macro score of 0.7168 for subatomic facts evaluation and 0.7938 for role evaluation against summary predictions, using 5-fold cross-validation [6].

ARC Scores Against Human Evaluation We compare ARC scores with lexical and semantic metrics from Elaraby et al. (2024), as well as entailment-based metrics, particularly SummaC (Laban et al., 2022) (both zero-shot and convolution-based), which measures alignment between source

[3] https://huggingface.co/cross-encoder/nli-deberta-v3-base

[4] Decomposition details in Appendix C.

[5] Evaluation prompts in Appendix D.

[6] Training setup and evaluation against GPT-4o are provided in Appendix E.

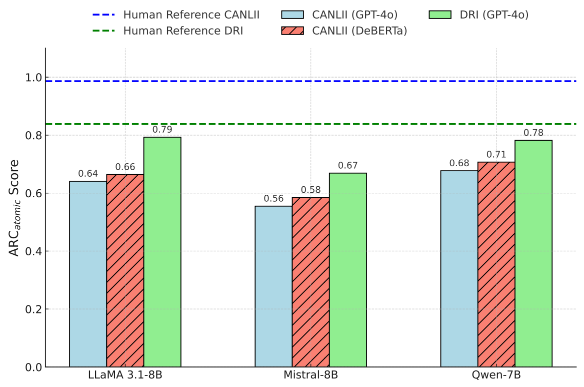

Figure 2: Average $\mathsf{ARC}_{\text{atomic}}$ across `CANLII` and `DRI` datasets.

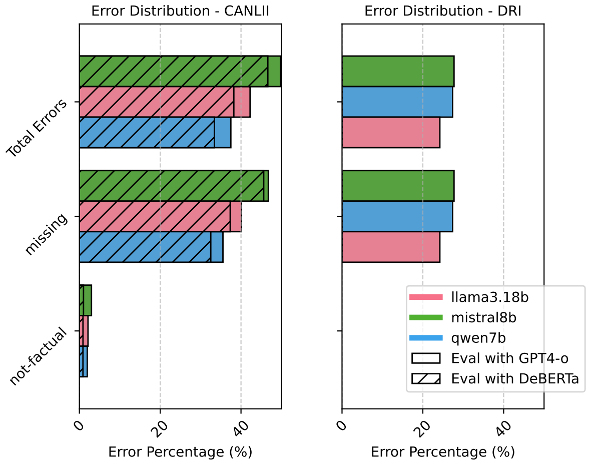

Figure 3: Error types distribution of $\mathsf{ARC}_{\text{atomic}}$.

articles and summaries. For entailment evaluation, argument roles are treated as the *hypothesis* and generated summaries as the *premise*, with SummaC computed at both document- and sentence-level granularities. Table 4 shows that ARC consistently outperforms lexical, semantic, and entailment metrics in correlating with expert annotations, with one exception: τ for $\mathsf{ARC}_{\text{role}}$ against Expert 2. The lower performance of $\mathsf{ARC}_{\text{role}}$ suggests that treating a full argument as an atomic unit introduces excessive penalization for minor omissions and inconsistencies, leading to misalignment with holistic human judgments. In contrast, $\mathsf{ARC}_{\text{atomic}}$ achieves the highest correlations with Expert 2 and the overall expert average, enabling getting a nuanced and interpretable score without losing the holistic overall coverage.

When using trained classifiers, we observe a drop in $\mathsf{ARC}_{\text{atomic}}$ performance but an improvement in $\mathsf{ARC}_{\text{role}}$, which is likely due to distributional shifts between training and human-evaluated summaries. Nonetheless, the stricter nature of $\mathsf{ARC}_{\text{role}}$ appears to amplify small deviations, reducing alignment with expert assessments. Given its strong expert correlation and interpretable error predictions, we adopt $\mathsf{ARC}_{\text{atomic}}$ as the primary metric for coverage evaluation, providing a reliable and fine-grained lens on model performance.

4.3 Obtaining generated summaries

To handle the long-form nature of argumentative texts, we employ open-weight LLMs with extended context windows: `LlaMA-3.1-8B-instruct` (`LlaMA-3.18-B`) (Grattafiori et al., 2024), `Mistral-8B-instruct(Mistral-8b)` (Jiang et al., 2024), and `Qwen2-7B` (qwe, 2024). For `Qwen2-7B`, to support context lengths exceeding 32k tokens, we integrate YARN embeddings (Peng et al.) into the model configuration prior to deployment. Inference is conducted using VLLM (Kwon et al., 2023) for scalability. Summaries are generated using 0 temperature sampling, capped at 2048 tokens to ensure fair long-form generation. Each document $d \in D$ is prompted with: `"Read the following text and summarize it: {input document}. Summarize in {reference summary word length} words. Summary:"`. The target length is dynamically set to match reference summaries for comparability. For `DRI`, the length is fixed to the longest reference summary to encourage maximal argument retention.

5 Results and Analysis

5.1 Do LLMs cover salient arguments effectively?

We compute $\mathsf{ARC}_{\text{atomic}}$[7] scores for human reference summaries on both the `CANLII` and `DRI` benchmarks using `GPT-4o`. For `CANLII`, this evaluation serves to assess the robustness of the metric by verifying whether it assigns near-perfect scores to ideal human-written summaries. For `DRI`, the goal is to measure how much human references cover salient argument roles compared to LLM-generated summaries. As shown in Figure 2, argument coverage remains imperfect across all models, with `Mistral-8B` lagging behind both `LlaMA-3.18B` and `Qwen2-7B`. Coverage for `DRI` is consistently higher and close to the reference summary compared to `CANLII` across all evaluated models. This pattern highlights the greater challenge of preserving salient argumentative information in legal texts,

[7] Full $\mathsf{ARC}_{\text{full}}$ and $\mathsf{ARC}_{\text{role}}$ results are in Appendix F.

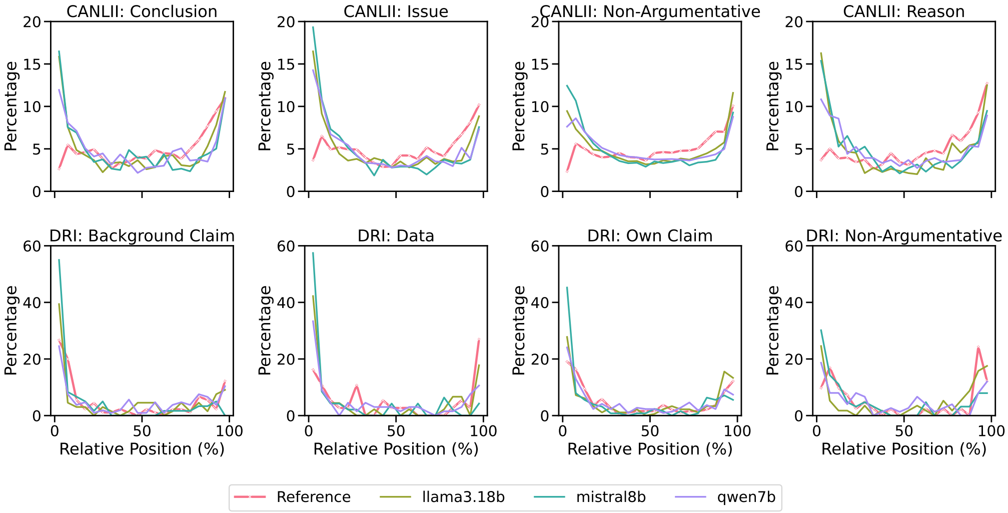

Figure 4: Source sentences relative position in the LLM context window across all models and various argument roles for both CANLII and DRI corpora.

where arguments are often sparsely distributed across lengthy and complex contexts.

Error types We analyze the types of error e extracted from $\Phi(m_j, S)$ of the $\text{ARC}_{\text{atomic}}$ evaluation. Figure 3 shows that the most frequent error is *missing*, indicating that facts in arguments are often omitted rather than misrepresented. While factual inconsistencies exist, they are less prevalent than missing information. These findings emphasize that beyond hallucination, ensuring comprehensive salient information coverage remains a critical challenge in summarization.

5.2 Do argument positions in the source affect their coverage in summaries?

Positions of arguments in the source from the perspective of LLMs Following Ravaut et al. (2024b), we start by analyzing the positions where included LLMs look at in its context window. We leverage the lexical greedy approach for source sentences identification (Ravaut et al., 2024a; Adams et al., 2023) by iteratively adding sentences in the source that maximizes ROUGE-1 score until there is no further improvement. We analyze the source sentence indices by their argument role annotations.

Figure 4 reveals a distinct U-shaped context window across all models in both datasets, with the effect being particularly pronounced in CANLII. Analyzing source sentences based on their argument role annotations suggests that argument positions are strongly influenced by this U-shaped pattern. This is especially concerning for CANLII, where reference summaries indicate that arguments do not adhere to a fixed positional pattern. In contrast, reference summaries in DRI more closely align with the LLM context window distribution.

Effect of context window positional bias on coverage We analyze the positions of salient arguments using $\text{ARC}_{\text{atomic}}$. For CANLII, we apply greedy sentence selection while restricting the selection to the annotated arguments in both the reference summary and the input document, thereby reducing computational cost and ensuring that only arguments in the reference summary are mapped to arguments in the input. For DRI, we directly select the positions of arguments in the input document with a relevance score $= 5$. We follow Ravaut et al. (2024b) by computing the mean relevant position of salient arguments and measuring the Pearson correlation ρ between this mean and $\text{ARC}_{\text{atomic}}$.

Model	CANLII	DRI
LlaMA-3.1-8B	-0.230	0.129
Mistral-8B	-0.369	-0.055
Qwen2-7B	-0.301	-0.223

Table 5: Pearson correlation (ρ) between mean relative position of salient arguments and $\text{ARC}_{\text{atomic}}$ (computed with GPT-4o) across models for CANLII and DRI. Correlations with p-value > 0.05 are shown in gray .

Table 5 [8] shows a significant negative correlation in CANLII ($p < 0.05$), indicating that LLM context windows impact argument coverage. In

[8] Correlation with full ARC scores included in Appendix G.

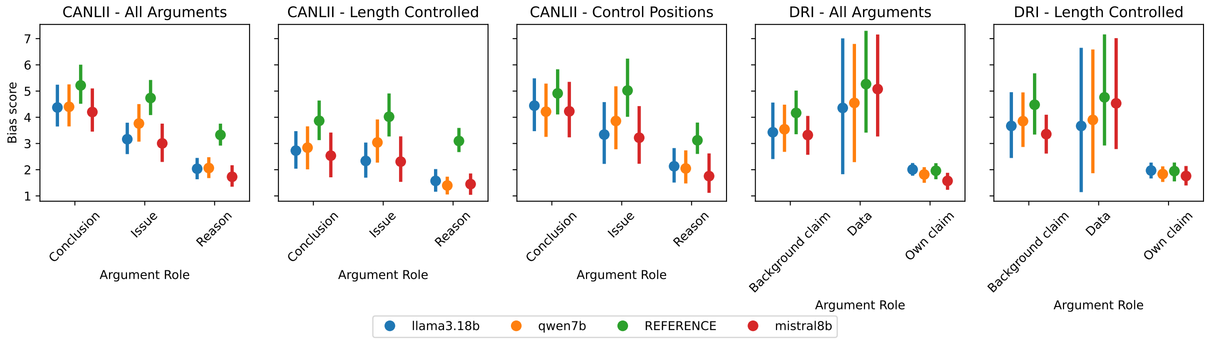

Figure 5: Bias score β across multiple argument roles (controlled length and non-controlled length) for both CANLII and DRI corpora.

DRI, correlations are weaker and not significant, with LlaMA-3.1-8B even showing a slight positive trend. These patterns align with Figure 4: reference summaries in CANLII differ in distribution from generated ones, while in DRI, argument roles better reflect the U-shaped context window bias. This suggests that context-window positional bias can hinder argument coverage, especially when arguments are sparsely distributed.

5.3 Do LLMs cover argument roles disproportionally?

We propose a bias score β. This score is grounded in the $\text{ARC}_{\text{atomic}}$ metric. We compute β for each argument role by first calculating its $\text{ARC}_{\text{atomic}}$ score and normalizing it by the role's prior frequency in the source document. This adjustment corrects for overrepresented roles, ensuring the bias score reflects true disparities in coverage. The final bias score for an argument role a is defined as:

$$\beta_a = \text{ARC}_{\text{atomic}_a} \times \frac{1}{\log\left(1 + \frac{|a|_D}{|\text{args}|_D}\right)}$$

Here, $|a|_D$ denotes the frequency of argument role a in source document D, and $|\text{args}|_D$ is the total number of arguments in D. To mitigate bias from argument length and position—especially in CANLII, where longer roles and mid-position arguments negatively affected coverage—we compute β within groups allowing up to 20% word-length variation. We also control for position by selecting CANLII articles in which at least 80% of arguments appear within the first or last 20% of the case.

Figure 5 [9] shows that salient arguments are not covered equitably across roles. In CANLII, *conclusions* are consistently better covered than *issues* and *reasons* across all groups. While length control reduces some role-related bias, *conclusion* remain the most covered, even under both length- and position-controlled settings. In contrast, for DRI, *own claims* are less covered relative to other roles, but when compared to reference summaries, their coverage is comparable. This suggests the lower bias score arises from normalization, reflecting the overrepresentation of *own claims* in the source corpus rather than a true coverage gap. Finally, while there is a bias to human reference in case of CANLII , the gap in argument coverage is notably larger for LLM outputs compared to DRI.

6 Conclusion and Future Work

We introduced ARC, a novel evaluation framework for evaluating how well LLM-generated summaries preserve salient arguments. Our multi-level formulation—spanning full-set, role-level, and subatomic coverage—correlates more strongly with human judgments compared to standard lexical, semantic, and entailment-based metrics, with subatomic evaluation showing highest overall correlation with human judgments. Using our proposed metric, we identified two key limitations in summarizing long legal opinions. First is positional bias, where LLMs tend to overrepresent content at the beginning or end of the input context, affecting the coverage of sparsely distributed arguments. Second is role bias, where models disproportionately favor covering *conclusions* over other roles such as *issues* and *reasons*. Future work can build on this work by exploring incorporating explicit argument representations during training or during prompting LLMs for a more informative summaries in high stakes domains.

[9] Appendix H includes bias analysis without prior frequency normalization to avoid denominator inflation.

Limitations

While the ARC framework enables a comprehensive multi-level evaluation of argument coverage, several limitations remain. First, our evaluation is constrained by the lack of dedicated benchmarks explicitly designed for argument coverage. Existing datasets offer limited annotation granularity, particularly at the subatomic level, which restricts the reliability of fine-grained assessments. As a result, our decomposition into atomic argument units depends on LLM-based prompting and entailment filtering, both of which may introduce inaccuracies. Second, the relatively small size of the DRI dataset (40 documents) limits the generalizability of our findings in the scientific domain. While we were constrained by the available datasets, developing larger, rigorously annotated datasets is an important direction for future work, though it is beyond the scope of this study. Finally, due to computational constraints, our evaluation only includes models that fit within available memory resources. We encountered out-of-memory issues with several larger models. Future work should consider incorporating larger-scale models to determine whether argument coverage improves with increased model capacity or if the observed limitations persist.

Ethics Statement

Our study complies with the ACL Ethics Policy. We primarily evaluate academically available datasets designed explicitly for research purposes,which we obtained through license agreement with the authors of both datasets, thus minimizing privacy risks. Additionally, our work acknowledges potential biases and inaccuracies inherent to LLM-generated outputs, including misrepresentation or omission of critical information from summaries, which could have significant implications in high-stakes domains such as law and science. Researchers and practitioners utilizing our framework should exercise caution and validate results carefully before applying these models in sensitive or consequential decision-making contexts.

References

2024. Qwen2 technical report.

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mohamed Elaraby, Huihui Xu, Morgan Gray, Kevin Ashley, and Diane Litman. 2024. Adding argumentation into human evaluation of long document abstractive summarization: A case study on legal opinions. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 28–35, Torino, Italia. ELRA and ICCL.

Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612, Toronto, Canada. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021a. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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 6866–6880, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Beatriz Fisas Elizalde, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, editors. LREC 2016. Tenth International Conference on Language Resources and Evaluation; 2016 May 23-28; Portorož, Slovenia.[Paris]: ELRA; 2016. p. 3081-8.* ELRA (European Language Resources Association).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023a. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023b. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Mining Argumentation*, page 40–46, Brussels, Belgium. Association for Computational Linguistics.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024b. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8647–8664, Mexico City, Mexico. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

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

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

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Jan Trienes, Jörg Schlötterer, Junyi Jessy Li, and Christin Seifert. 2025. Behavioral analysis of information salience in large language models. *arXiv preprint arXiv:2502.14613*.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2024. On positional bias of faithfulness for long-form summarization. *arXiv preprint arXiv:2410.23609*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Huihui Xu, Jaromír Šavelka, and Kevin D Ashley. 2020. Using argument mining for legal text summarization. In *Legal Knowledge and Information Systems*, pages 184–193. IOS Press.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 250–254.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.

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

A Extended analysis on included datasets

A.1 Examples from included datasets

Table 6 presents an excerpt from a legal opinion in the CANLII dataset, with arguments highlighted in both the input and the reference summary. Table 7 provides an excerpt from a scientific article in the DRI dataset, with highlighted arguments in the input. Although the documents are truncated for space, the examples clearly illustrate a key distinction: in CANLII, arguments constitute a smaller fraction of the input, whereas in DRI, the input is densely populated with argumentative content.

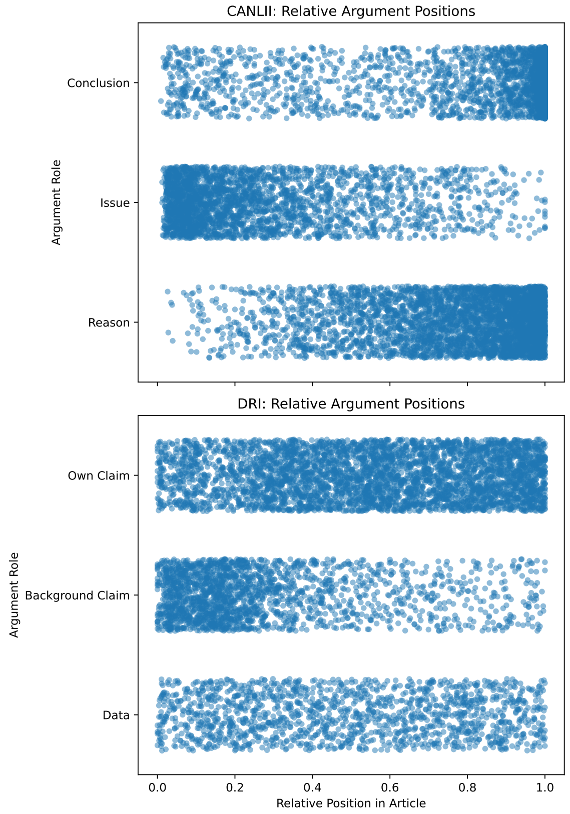

Figure 6: Distribution of argument roles in the input in both CANLII and DRI.

A.2 Distribution of arguments across the input

Figure 6 illustrates the distribution of argument roles across the source documents. In CANLII, *Conclusion* statements predominantly appear toward the end of the document, while *Issue* statements are concentrated near the beginning. In DRI, *Background claims* are more frequent at the start of the document, which aligns with the conventional structure of scientific writing where literature reviews—typically containing claims from prior work—are introduced early on.

A.3 Distribution of salient arguments

Figure 7 presents the distribution of argument roles in CANLII reference summaries and in DRI sentences annotated with a relevance Likert score of 5 (indicating very high likelihood of inclusion in a summary). In CANLII, the distribution of argument roles is relatively balanced across categories, whereas in DRI, *own claims*—statements made directly by the authors—dominate the content.

Input Article (Truncated)
Q.B. A.D. 1987 No. CS 1159 J.C.R., Regina. Applicants seek to quash a search warrant issued by a justice of the peace. ***The respondent, a justice of the peace, issued a search warrant to search a dwelling house for weapons allegedly used in an attempted armed robbery.*** *The applicants claim the warrant was unlawfully issued without proper grounds. Specifically, the sworn information relied solely on hearsay from an unidentified informant, lacking corroborating details.* ***The applicants argue that no reasonable or probable grounds were disclosed to believe the weapons would be found at the searched location.*** *They highlight that the informant's reliability was not established, nor was there an oath affirming the informant's credibility. The search warrant was issued under Section 443(1)(b) of the Criminal Code, which allows a justice to issue a warrant if reasonable grounds exist to believe evidence of an offence will be found.* ***The court explains that on applications to quash a warrant, the reviewing judge cannot substitute their opinion for that of the justice of the peace.*** *Instead, the judge must simply determine whether any evidence existed upon which the justice could be satisfied on reasonable grounds. Reliance on confidential informants is permitted, even if detailed particulars are absent, provided sufficient basis exists for reliability. Past cases (e.g., Re Lubell, Re Dodge) have accepted similar levels of disclosure to protect informant anonymity.* ***The court notes that substantial compliance with Section 443 is sufficient; perfection in drafting is not required.*** *Given practical constraints faced by peace officers preparing information, reasonable latitude must be given in interpreting the sworn information.* ***The judge concludes that although more information could have been provided, there was sufficient evidence upon which the justice could reasonably issue the warrant. Accordingly, the respondent acted within her jurisdiction, and the application to quash the warrant is dismissed.***
Reference Summary
Warrant issued to search a dwelling house for weapons allegedly used in an attempted armed robbery. *The affidavit in support referred to an unknown informant.* ***Judge applied the test that the justice of the peace 'must be satisfied on reasonable grounds.'*** ***Substantial compliance found and warrant upheld.***

Table 6: Example of an input legal document (non argumentative text are shortened for space) and its reference summary from CANLII. Highlighted sentences correspond to argumentative roles: Issue, Reason, and Conclusion.

A.4 Rhetorical roles per relevance to summary likert score

To better understand the relationship between rhetorical structure and relevance to the summary, we compute the percentage of each rhetorical role across Likert-rated sentences in the DRI corpus. As shown in Figure 8, non-argumentative content dominates among sentences rated as totally irrelevant to the summary (Likert score 1). However, as the perceived relevance increases, argumentative content becomes more prominent, with Own Claim consistently emerging as the most frequent rhetorical role across all higher-quality categories. This trend highlights a clear shift toward structured argumentative writing in more relevant reflections.

B Likert scores based on human evaluation

Table 8 shows the Likert scale from 1 to 4 definitions.

C Fact decomposition algorithm

Algorithm 1 outlines the decomposition process for an arbitrary argument $a_i \in \mathbb{A}$, performed via prompting GPT-4o.

Table 9 presents the prompt used to extract atomic facts. Table 10 provides an example decomposition from an *issue* argument role. The second fact is not supported by the original argument and is thus excluded from the final ARC score computation.

Algorithm 1: Argument Decomposition and Entailment Filtering

```
Input: Argument $a_i$, Entailment Model $\mathcal{M}$,
       Entailment Threshold $\tau$
Output: Filtered Atomic Facts $\mathcal{F}(a_i)$
Initialization:
$\mathcal{F}(a_i) \leftarrow \emptyset$     (Set of filtered atomic facts)
Decomposition:
Decompose $a_i$ into atomic facts:
  $\{m_1, m_2, \ldots, m_n\}$
foreach atomic fact $m_j$ do
    Compute entailment using
      $\mathcal{M}(m_j, a_i) \rightarrow (e, c, n)$
    if $e$ (entailment) is predicted then
        Add $m_j$ to $\mathcal{F}(a_i)$
Return: Filtered atomic facts $\mathcal{F}(a_i)$
```

D Evaluation Prompts

Evaluation prompts for $\text{ARC}_{\text{fullset}}$, ARC_{role}, and $\text{ARC}_{\text{atomic}}$ are described in Table 11, 12, 13 respectively. In each prompt, we ask the LLM to first generate a rationale before assigning a scoring decision following standard evaluation with LLMs (Liu et al., 2023a).

E Training C_{atomic} and C_{role}

We fine-tune DeBERTa, BigBirdRoBERTa, and LegalBERT using checkpoints obtained from the *HuggingFace* library (Wolf et al., 2019). All models are trained using 5-fold cross-validation based on 100 case-summary pairs for each model forming a total of 300 case-summary pairs, resulting in

Input Article (Truncated)

Our method maintains an explicit polygonal mesh that defines the surface, and an octree data structure that provides both a spatial index for the mesh and a means for efficiently approximating the signed distance to the surface. At each timestep, a new surface is constructed by extracting the zero set of an advected signed-distance function. Semi-Lagrangian backward path tracing is used to advect the signed-distance function. One of the primary advantages of this formulation is that it enables tracking of surface characteristics, such as color or texture coordinates, at negligible additional cost. We include several examples demonstrating that the method can be effectively used as part of a fluid simulation to animate complex and interesting fluid behaviors. The fundamental problem of tracking a surface as it is advected by some velocity field arises frequently in applications such as surface reconstruction, image segmentation, and fluid simulation. Unfortunately, the naive approach of simply advecting the vertices of a polygonal mesh quickly encounters problems such as tangling and self-intersection. Instead, a family of methods, known as level-set methods, has been developed for surface tracking. These methods represent the surface implicitly as the zero set of a scalar field defined over the domain. Level-set methods avoid dealing with topological changes but require high-order conservation law solvers. In contrast, our method constructs a surface directly using semi-Lagrangian contouring without solving PDEs, preserving surface detail efficiently. Using adaptive octree data structures, we can efficiently and reliably construct the new surface and corresponding signed-distance function. This allows tracking surface properties such as color or texture coordinates directly on the polygonal mesh during advection, enabling realistic animation of complex fluids. Prior methods often suffered from volume loss and smoothing artifacts, particularly in underresolved, high-curvature regions. By using an explicit surface representation, we compute exact distances near the mesh and avoid substantial interpolation errors. ... Finally, the method produces detailed, flicker-free animations of fluid behavior, demonstrating significant advantages over traditional level-set and particle-based approaches.

Reference Summary

This article presents a semi-Lagrangian surface tracking method that explicitly represents the surface as a set of polygons. The new surface and corresponding signed-distance function can be efficiently and reliably constructed using adaptive octree data structures. One of the primary advantages of this method is that it enables tracking surface characteristics, such as color or texture coordinates, or even simulation variables, accurately at negligible additional cost. These properties can be easily stored directly on the polygonal mesh and efficiently mapped onto the new surface during semi-Lagrangian advection. At each timestep, a new surface is constructed by extracting the zero set of an advected signed-distance function. The explicit representation provides advantages on computing exact signed-distance values near the mesh and storing properties on mesh vertices. It also facilitates other common operations developed for manipulating and rendering triangle meshes. To avoid the topological difficulties of directly updating an explicit surface representation, the surface is updated in time through an implicit representation. The implicit representation is then used to construct a new mesh and extracted using a contouring algorithm. For its simplicity, robustness, and speed, marching-cubes method is used for contouring. After the triangle mesh has been extracted, true distance values are assigned to the vertices of octree. This process is known as redistancing, which comprises three steps: coarsen the octree; compute exact distances at vertices; run a fast marching method over the remaining vertices. Finally, this method is able to produce detailed, flicker-free animations of complex fluid motions.

Table 7: Example from DRI showing an input scientific article and its corresponding reference summary. Sentences in the input article are highlighted according to their argument role: Own Claim, Background Claim, Data. The reference summary is unannotated.

Rating scale of the Generated Summary

1. **No arguments covered:** The generated summary did not cover the highlighted arguments in the reference summary or covered them only inadequately.

2. **Few arguments covered:** The generated summary adequately covered only a limited number of the highlighted arguments in the reference summary.

3. **Most arguments covered:** The generated summary adequately covered most of the arguments highlighted in the reference summary.

4. **All arguments covered:** The generated summary adequately covered all the highlighted arguments in the reference summary.

Table 8: Likert scale exact meaning for each score based on definitions obtained from Elaraby et al. (2024)

2220 argument-summary pairs for training C_{role} and 4914 atomic fact-summary pairs for training C_{atomic}. To handle class imbalance, we leverage weighted cross-entropy loss, where class weights are proportional to the label distribution across the dataset. This is particularly important for C_{atomic}, where the $(0, \text{non-factual})$ label is significantly underrepresented compared to others. All models are trained for 25 epochs using the Adam optimizer with an initial learning rate of 1×10^{-5}. The maximum sequence length is set to 512 for most models due to encoder constraints. However, for BigBirdRoBERTa, we expand the maximum length to 1024 to accommodate longer summaries and maintain coverage.

Table 14 shows that across all models, DeBERTa-Large obtained the best F1 macro average scores. Including encoders that can handle longer lengths didn't improve the performance of prediction.

F ARC$_{\text{full}}$ and ARC$_{\text{role}}$ results

Figure 9 presents the results for both ARC$_{\text{fullset}}$ and ARC$_{\text{role}}$ across the CANLII and DRI datasets.

The ARC$_{\text{fullset}}$ results indicate that both LlaMA-3.18B and Qwen2-7B achieve comparable levels of argument coverage from a holistic perspective. In contrast, and consistent with the ARC$_{\text{atomic}}$ results presented in Section 5,

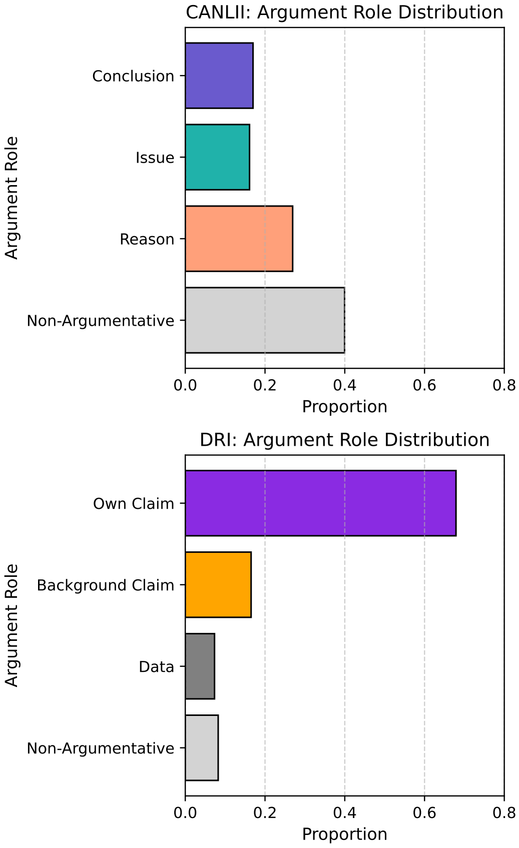

Figure 7: Argument role distributions in summaries for CANLII and DRI (for sentences with relevance score is 5). In CANLII, arguments are less densely represented compared to DRI, where own claims dominate.

Mistral-8B demonstrates slightly lower performance than both models. However, a key limitation of $\text{ARC}_{\text{fullset}}$ is its insensitivity to fine-grained differences, as it primarily relies on coarse Likert-style scores that obscure nuanced variation across model outputs.

By contrast, ARC_{role} reveals that LlaMA-3.18B exhibits a higher number of perfectly covered roles compared to both Mistral-8B and Qwen2-7B. Nonetheless, due to its stricter definition of completeness at the role level, ARC_{role} underrepresents the strength of Qwen2-7B in capturing a greater number of atomic facts than Mistral-8B, as evidenced by both $\text{ARC}_{\text{fullset}}$ and $\text{ARC}_{\text{atomic}}$.

G Correlation with source argument positions across full ARC scores

Table 15 confirms that the LLM context window significantly impacts coverage across all scores (including $\text{ARC}_{\text{fullset}}$ and ARC_{role}) in CANLII ($p < 0.05$), with a predominantly negative correlation.

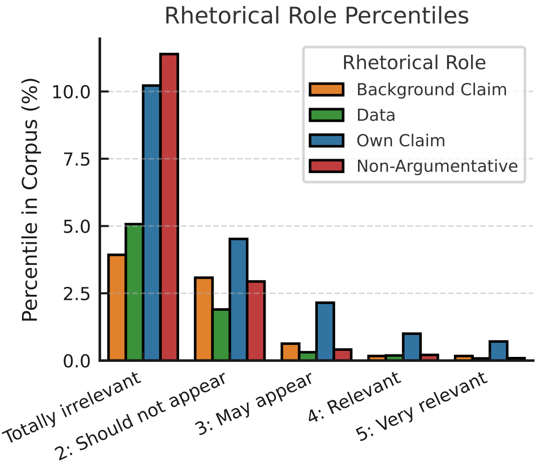

Figure 8: Rhetorical roles per each relevance score to summary from 1 to 5

Prompt Given to GPT4-o for Argument Decomposition

Task:

Extract a set of **atomic facts**—statements that can be **directly inferred** from the argument without interpretation, assumptions, or redundancy.

Guidelines:

- Extract only **explicitly stated** atomic facts.
- **Do not repeat** facts or infer from external knowledge.
- Maintain **granularity**: each fact should be **minimal yet complete**.
- Output a **valid Dictionary object** where each key is "fact1", "fact2", etc., and the values are the corresponding atomic facts.
- **No additional text or formatting**; dictionary object only.
- Each argument must yield at least **one atomic fact**.

Example Output Format:

```
{
    "fact1": "First atomic fact",
    "fact2": "Second atomic fact",
    "fact3": "Third atomic fact"
}
```

Input:

{argument}

Output:

(Dictionary object only)

Table 9: Prompt provided to GPT4-o for extracting atomic facts from arguments.

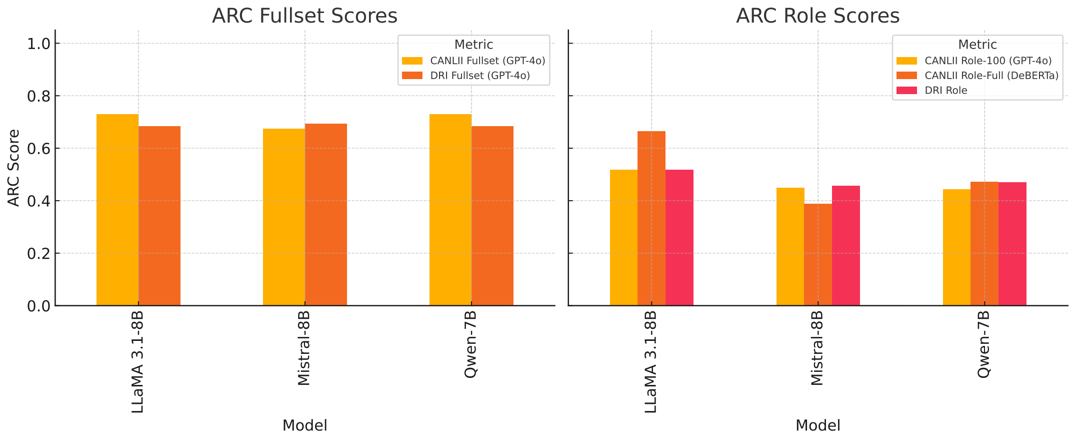

Figure 9: ARC$_{\text{full}}$ and ARC$_{\text{role}}$ results across models. Higher values indicate better argument coverage.

Argument (Issue):
FIAT. The father applied to have the mother cited for contempt for denial of access.

✓ **Fact 1:** The father applied to have the mother cited for contempt. (Entailed)

✗ **Fact 2:** The father applied for denial of access. (Not-entailed)

Table 10: Example of argument decomposition from CANLII, showing atomic facts with entailment status.

In contrast, its effect in DRI is not statistically significant, and in some cases (e.g., LLaMA-3.1-8B), it is slightly positive.

H Bias analysis for argument coverage without argument role normalization

While our normalization in computing β corrects for frequency skew, it may understate coverage for dominant argument roles in the source with inherently high raw ARC$_{\text{atomic}}$ scores. Therefore, we also examine role-specific reporting bias by directly computing $\beta_a =$ ARC$_{\text{atomic}a}$ to offer a complete picture of the role bias analysis. As shown in Figure 10, results on CANLII confirm prior findings: LLMs tend to prioritize covering *conclusion* arguments over *issue* and *reason* roles, both in the controlled and non-controlled length settings. For DRI, removing frequency normalization reveals that ARC$_{\text{atomic}_a}$ scores for roles such as background claim, own claim, and data are comparable to each other. This suggests that without normalization, the higher coverage of own claim may stem from its over-representation in the source documents, rather than a model-level preference.

Prompt Given to GPT4-o for Fullset Evaluation

Task:
Evaluate how well a given summary covers a provided set of arguments. Assign a score from 1 to 4 based on the extent of coverage, provide a clear explanation for your rating, and output the result in a specified JSON format.

Instructions:

- **Read** the provided arguments and summary carefully.
- **Rate** the extent to which the arguments are covered by the summary using the scale described in Table 8.
- **Format** your evaluation as a JSON object with:
  - "explanation": A concise explanation of your rating.
  - "rating": The assigned score (1 to 4).

Example Output Format:

```
{
  "explanation": "Place your explanation here",
   "rating": "Place your rating here"
}
```

Input:

- **Arguments:** {reference_arguments}
- **Summary:** {generated_summary}

Output:
Provide your evaluation in the specified JSON format.

Table 11: Prompt provided to GPT4-o for fullset-level argument coverage evaluation.

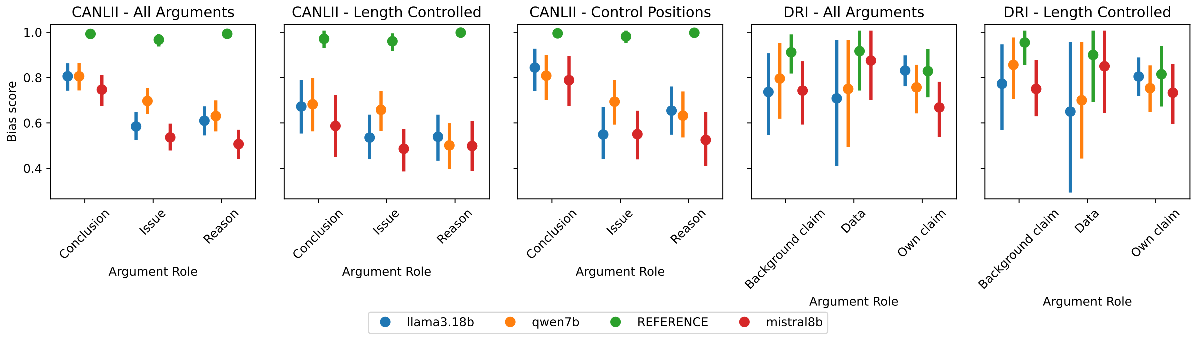

Figure 10: Bias score without any frequency normalization across multiple argument roles (controlled length and non-controlled length) for both `CANLII` and `DRI` corpus.

Prompt Given to `GPT4-o` for Argument Role Evaluation

Task:

Determine whether a summary fully supports a given argument or omits/contradicts key information.

Instructions:

- Output **1** if the summary **fully supports** the argument without omissions or contradictions.
- Output **0** if the summary **fails to support** the argument or contains **contradictory or incorrect** details (e.g., logical errors, entity mismatches).
- Respond in a **JSON object** with:
  - `"decision"`: Either `1` or `0`.
  - `"explanation"`: A brief justification, noting missing or conflicting content.

Input:

```
Argument: {argument}
Summary: {summary}
```

Output Format:

Respond **only** with a JSON object structured as:

```
{
    "explanation": "<Brief reasoning for your decision>",
   "decision": <0 or 1>
}
```

Note: Think critically before deciding. **Do not include any extra text beyond the JSON output.**

Table 12: Prompt provided to `GPT4-o` for role-level evaluation of argument support in summaries.

Prompt Given to `GPT4-o` for Atomic-Level Evaluation

Task Description:

Given an **argument** and a **summary**, evaluate whether the argument is supported by the summary and return a valid tuple in the specified format.

Explanation:

Provide a brief justification for your decision, identifying any missing, contradictory, or factually incorrect details.

Return Guidelines:

- **(1, "supported")**: The argument is **fully supported** by the summary.
- **(0, "missing")**: The argument **cannot be inferred** from the summary.
- **(0, "not-factual")**: The summary **contradicts** or misrepresents the argument.

Output Format:

Respond **only** with a JSON object, structured as:

```
{
   "explanation": "<explanation placeholder>",
      "decision": (1, "supported") or (0,
"missing") or (0, "not-factual")
}
```

Input:

```
Argument: {argument}
Summary: {summary}
```

Note: Think critically before deciding. **Do not generate any extra text beyond the JSON output.**

Table 13: Prompt provided to `GPT4-o` for atomic-level argument entailment evaluation.

Model	**F1 Macro Score** (C_{role})	**F1 Macro Score** (C_{atomic})
DeBERTa-base	0.7138	**0.7936**
DeBERTa-large	**0.7202**	**0.7936**
LegalBERT	0.6850	0.7812
BigBirdRoBERTa	0.6629	0.7891

Table 14: F1 Macro Scores for C_{role} and C_{atomic} across different models.

Metric	**Model**	CANLII	DRI
ARC$_{\text{fullset}}$	LlaMA-3.1-8B	-0.137	-0.152
	Mistral-8B	-0.042	0.250
	Qwen2-7B	-0.187	-0.174
ARC$_{\text{role}}$	LlaMA-3.1-8B	-0.200	0.202
	Mistral-8B	-0.244	-0.065
	Qwen2-7B	-0.216	-0.124
ARC$_{\text{atomic}}$	LlaMA-3.1-8B	-0.230	0.323
	Mistral-8B	-0.369	-0.055
	Qwen2-7B	-0.301	0.041

Table 15: Pearson correlation (ρ) between mean relative position of salient arguments and ARC metrics across models for CANLII and DRI. Correlations with p-value > 0.05 are shown in gray.