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

Mohamed Elaraby, Diane Litman
University of Pittsburgh
Pittsburgh, PA, USA
{mse30,dlitman}@pitt.edu

Abstract

We introduce **A**rgument **R**epresentation **C**overage (ARC), a bottom-up evaluation framework that assesses how well summaries preserve salient arguments, a crucial issue in summarizing high-stakes domains such as law. ARC provides an interpretable lens by distinguishing between different information types to be covered and by separating omissions from factual errors. Using ARC, we evaluate summaries from eight open-weight LLMs in two domains where argument roles are central: *long legal opinions* and *scientific articles*. Our results show that while LLMs capture some salient roles, they frequently omit critical information, particularly when arguments are sparsely distributed across the input. Moreover, ARC uncovers systematic patterns—showing how context window positional bias and role-specific preferences shape argument coverage—providing actionable guidance for developing more complete and reliable summarization strategies.

1 Introduction

LLMs have achieved remarkable progress in abstractive summarization, producing summaries that are fluent, coherent, and often preferred by human evaluators in domains such as news (Zhang et al., 2024; Liu et al., 2024b). This progress has shifted the focus of summarization research from *how to generate fluent summaries* to *how to properly evaluate them*. Evaluation is especially critical because, despite their fluency, LLMs frequently hallucinate or omit key content (Huang et al., 2025; Kim et al.), undermining their reliability in high-stakes domains.

In this work, we focus on the evaluation of long-context summaries generated by instruction-following LLMs, focusing on whether they preserve **salient argumentative content**. In crucial domains such as law, the most essential information is conveyed through *argument roles* (the components of reasoning and main claims in the text). Preserving these roles is particularly challenging in long documents, where arguments are sparsely distributed across thousands of tokens (Elaraby and Litman, 2022). The ability of LLMs to identify and retain argument roles is therefore a crucial test of their utility for reliable summarization in high-stakes settings.

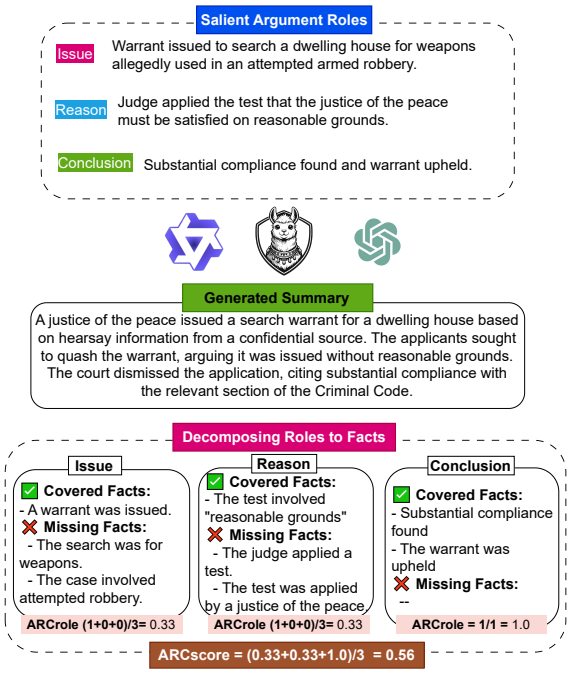

Figure 1: Overview of the ARC framework, which computes coverage hierarchically from atomic facts to roles to an overall summary $\text{ARC}_{\text{score}}$. Example shown from a LlaMA3.18B summary of a long legal opinion.

Why argument-based evaluation? Prior work has shown that explicitly modeling argument roles can improve summarization quality (Fabbri et al., 2021a; Elaraby and Litman, 2022; Elaraby et al., 2023). Building on this, argument-based evaluation allows us to address a fundamental research question: **RQ1: Can instruction-following LLMs**

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3626–3643
March 24-29, 2026

inherently identify and preserve argument roles without explicit supervision? Additonally, we see two key benefits of argument-based evaluation where existing metrics fall short: (1) *interpreting which types of information are missing or misrepresented in generated summaries*, and (2) *guiding future alignment strategies for domain-specific summarization through a reproducible methodology*. In this work, we focus on the first point by developing argument-based evaluation methods that make coverage interpretable.

We introduce **A**rgument **R**epresentation **C**overage (ARC), [1] a framework for evaluating how well LLM-generated summaries capture salient arguments. ARC leverages atomic fact decomposition of argument roles to measure coverage *hierarchically*. As illustrated in Figure 1 (decomposing roles to facts), all salient arguments are first decomposed into atomic facts to assess **fact-level coverage**. Each atomic fact is then compared against the generated summary and labeled as either *supported*, *missing*, or *contradicted*. These binary fact-level judgments are aggregated to compute **role-level coverage (ARC_{role})**, providing fine-grained insight into how well each argument role is preserved. Finally, role-level scores are combined to produce **summary-level coverage ($\text{ARC}_{\text{score}}$)**, a holistic measure of how effectively a summary preserves salient argument roles. We apply ARC to study coverage in two domains where argument structure is essential for understanding the core of the document: *long legal opinions* and *scientific articles*.

Prior work shows that LLMs often exhibit positional biases—favoring content from the beginning or end of long documents (Liu et al., 2024a; Ravaut et al., 2024; Wan et al., 2025). It is less clear, however, whether they also bias toward certain types of salient information that share the same structure (e.g., argument roles) during summarization. With ARC, we investigate additional two research questions: **RQ2: How does the position of arguments in the source document affect their inclusion in summaries?** and **RQ3: Are certain argument roles disproportionately favored over others?** To answer the latter, we leverage ARC_{role} to define a bias score that quantifies role-specific biases in saliency coverage.

Our analysis with ARC across eight open-weight LLMs reveals: (1) instruction-following models still struggle with saliency coverage, particularly when argument roles are sparsely distributed, underscoring the need for further alignment; (2) positional bias in the context window negatively impacts coverage; and (3) bias analysis shows that LLMs exhibit systematic preferences for certain argument roles, particularly in long legal opinions.

[1]The full codebase, along with instructions for obtaining the data, is publicly available at https://github.com/EngSalem/ARCScore.git.

2 Related Work

Information Saliency in LLMs' Summaries. 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 entity planning. In similar veins, Gantt et al. (2024); Walden et al. (2025) show that LLMs generate more focused and complete event-centered summaries when generation is explicitly conditioned on both the input document(s) and structured event representations. *We extend this line of work by treating argument roles as a structured form of saliency and analyzing their preservation in LLM-generated summaries.*

LLMs in Long-Document Summarization. LLMs face persistent issues when summarizing long texts, notably the U-shaped positional bias (Liu et al., 2024a)—favoring content at the beginning and end while neglecting the middle (Ravaut et al., 2024). This leads to degraded faithfulness in long-form outputs (Wan et al., 2025). *We expand this analysis by quantifying how positional bias affects the coverage of salient argumentative content.*

Argument Mining in Abstractive Summarization. Incorporating argument structures into summarization has shown promise across domains, 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 et al., 2016). *We build on this by assessing whether instruction-following LLMs can cover salient arguments without the external argument role information*

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.

CANLII (Legal opinions)
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 (Scientific documents)
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 and DRI. Colors distinguish different argument roles.

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

3 Datasets

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 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., 2023). More analysis and examples are given in Appendix A.

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.

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. 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 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 be-

hind 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

Decomposing Roles to Facts. For each argument role $r_i \in R$, where R is the set of salient argument roles (roles that are essential to be included in the summary), we follow the decomposition algorithm of Yang et al. (2024b). Specifically, we employ a strong LLM (GPT4-o) to decompose r_i into a set of atomic facts $\mathbb{F}_r = \{f_1, f_2, \ldots, f_{|\mathbb{F}_r|}\}$. These facts are further filtered using an entailment check against r_i to eliminate over-generated facts. [2]

Role-Level Coverage. Given a generated summary S, we define the indicator function:

$$\delta(f_i, S) = \begin{cases} 1 & \text{if atomic fact } f_i \text{ is correctly covered in } S, \\ 0 & \text{otherwise.} \end{cases}$$

The role-level coverage for r is then:

$$\mathsf{ARC}_{\text{role}}(r, S) = \frac{1}{|\mathbb{F}_r|} \sum_{i=1}^{|\mathbb{F}_r|} \delta(f_i, S).$$

Summary-Level Coverage. The overall coverage score, $\mathsf{ARC}_{\text{score}}$, across all roles is computed as aggregation of $\mathsf{ARC}_{\text{role}}$ across R:

$$\mathsf{ARC}_{\text{score}}(S) = \frac{1}{|R|} \sum_{r \in R} \mathsf{ARC}_{\text{role}}(r, S).$$

4.2 Evaluating Factual Units (δ)

The main goal of the ARC framework is to provide an interpretable evaluation of coverage in generated summaries. To do so, the δ function distinguishes between two error types: *missing facts* (information omitted) and *incorrectly covered facts* (information misrepresented). [3] Although both reduce completeness, separating them reveals whether errors stem from failing to prioritize salient content or from factual mistakes during generation. We implement δ using an LLM judge. [4] We employ zero-shot LLMs, which have shown strong correlation with human judgments in summary evaluation (Liu et al., 2023). Given an atomic fact and a summary, we prompt the LLM to return a pair (d, e), where $d \in \{0, 1\}$ indicates whether the fact is supported (*covered*) or not supported (*missing or incorrect*), and e provides a categorical interpretation of the error type *(no error, missing, non-factual)*.

4.3 Benchmarking $\mathsf{ARC}_{\text{score}}$

To evaluate the effectiveness of $\mathsf{ARC}_{\text{score}}$ in capturing holistic summary coverage, we benchmark it against existing automatic metrics using expert-annotated data.

Expert-Annotated Data. Human-annotated datasets for fine-grained coverage evaluation are extremely scarce. To our knowledge, the only available resource is the dataset of Elaraby et al. (2024), which includes 90 legal opinion summaries annotated for salient argument coverage on a 4-point Likert scale. [5] Two legal experts independently annotated the data, achieving a moderate agreement of $\kappa = 0.46$ (quadratic weighted). To improve the robustness of the annotations, we filtered out pairs with > 1 point of disagreement, resulting in 87 article–summary pairs with improved agreement ($\kappa = 0.605$). We use the expert average as the gold standard to compare against automatic metrics.

Automatic Metrics for Holistic Coverage. We benchmark $\mathsf{ARC}_{\text{score}}$ against a suite of strong automatic metrics. For all baselines, we treat the set of salient arguments as the *hypothesis* and the generated summary as the *premise/reference*. We first

[2] Decomposition algorithm in Appendix B.

[3] Examples of atomic facts are included in Appendix C.

[4] See Appendix D for the evaluation prompt.

[5] See Appendix E for scale definitions.

Metric	Expert Avg.		Interpretability	
	τ	ρ	Info Types	Errors
ROUGE-1	0.391	0.539	✗	✗
ROUGE-2	0.336	0.475	✗	✗
ROUGE-L	0.345	0.479	✗	✗
BERTScore	0.354	0.517	✗	✗
SummaC$_{ZS}$ (sent)	0.387	0.537	✗	✗
SummaC$_{ZS}$ (doc)	0.375	0.476	✗	✗
SummaC$_{conv}$ (sent)	0.345	0.459	✗	✗
SummaC$_{conv}$ (doc)	0.352	0.311	✗	✗
FactScore (GPT4-o)	0.405	0.549	✗	✗
ARC$_{score}$ (GPT4-o)	**0.465**	**0.593**	✓	✓

Table 3: Metric correlations (Kendall's τ, Pearson's ρ) with expert judgments using 87 articles ($\kappa = 0.605$). All rows: $p < 0.05$.

Judge Model	τ	ρ
Proprietary models		
GPT-4o	0.465	0.593
GPT-4o-mini	0.446	0.582
Open-weight Instruction-following models		
Qwen-2.5-7B-Instruct	0.329	0.480
Qwen-2.5-14B-Instruct	0.454	0.601
Llama-3.1-8B-Instruct	*0.463*	*0.610*
Mistral-8B-Instruct	0.443	0.599
Open-weight Reasoning models		
QwQ-32B-AwQ	0.437	0.596
DeepSeek-R1-Distill-Qwen-14B	**0.509**	**0.638**

Table 4: Expert-average correlations (Kendall's τ, Pearson's ρ) for judges in ARC$_{score}$. *Italicized* = best instruction-following, underlined = best proprietary, **bold** = overall best. $p < 0.05$ for all rows.

include the standard metrics reported in Elaraby et al. (2024): ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. Next, we add entailment-based baselines from Laban et al. (2022), where SummaC computes entailment between hypothesis and reference at both sentence-level and document-level granularity. We include both (zeroshot zs and convolution conv versions of SummaC). Finally, we incorporate an LLM-judge baseline: FactScore (Min et al., 2023), which—similar to ARC$_{score}$—relies on decomposition for fact-level evaluation. We adapt FactScore for coverage by treating the generated summary as the knowledge base and the salient arguments as hypotheses to be tested. For fairness, both ARC$_{score}$ and FactScore use GPT4-o as the evaluator. We report correlation with expert averages using Kendall's τ and Pearson's ρ.

Table 3 shows that among automatic metics, ARC$_{score}$ achieves the highest correlation with expert judgments,[6] outperforming standard evaluation metrics and LLM-based decomposition-based metrics such as FactScore. Crucially, ARC$_{score}$ improves the holistic view of coverage over baselines, while also providing interpretability that existing metrics lack. Unlike FactScore, which provides only fact-level coverage, ARC enables multi-level analysis: (i) role-level coverage via ARC$_{role}$, and (ii) error-type analysis distinguishing between *missing* and *incorrect* facts. This decomposition allows us to diagnose not only how much coverage is achieved, but also which roles are underrepresented and why errors occur, making ARC a more effective diagnostic tool for granular coverage evaluation.

Alternative LLM-judges. To make ARC more scalable and accessible, we explore alternatives to GPT-4o for fact-level evaluation that maintain high correlation with expert ratings while reducing inference cost and avoiding reliance on proprietary APIs that may be deprecated. We focus on two categories of open-weight models: (i) *instruction-following models.* We evaluate smaller open-weight models that fit within our compute and memory budgets, including Qwen-2.5-Instruct (7B and 14B), Llama-3.1-8B-Instruct, and Mistral-8B-Instruct. (ii) *Reasoning models.* We further examine open-weight reasoning models, specifically QwQ-32B-AwQ and DeepSeek-R1-Distill-Qwen-14B, which have demonstrated competitive performance with proprietary instruction-following and reasoning models on complex tasks, including verification. Table 4 shows that DeepSeek-R1-Distill-Qwen-14B achieves the strongest overall correlation, surpassing even proprietary models such as GPT-4o. Among open-weight instruction-following models, Llama-3.1-8B-Instruct attains the highest correlation, outperforming GPT-4o-mini. We employ both proprietary (GPT-4o) and open-weight judge (DeepSeek-R1-Distill-Qwen-14B) to compute ARC$_{score}$. To balance cost and scalability, GPT-4o is applied to a representative subset of 100 CANLII (CANLII$_{100}$) cases and the full DRI corpus, while DeepSeek-R1-Distill-Qwen-14B is used for both CANLII$_{100}$ and the full CANLII (1049 articles). Our goal is to ensure that the sensitivity to the underlying judge (i.e., disagreements at the fact level) does not alter the overall coverage trends

[6] See Appendix F for correlations with each expert.

Model	CANLII$_{100}$		CANLII	DRI	
	GPT4-o	DeepSeek	DeepSeek	GPT4-o	DeepSeek
Reference	**0.986**	**0.977**	**0.982**	**0.838**	**0.813**
Qwen-2.5-14B	*0.677*	*0.707*	0.722	0.772	0.797
Qwen-2.5-7B	0.676	0.691	*0.725*	*0.799*	*0.806*
Qwen-2.5-3B	0.648	0.702	0.710	0.730	0.759
Qwen-2.5-1.5B	0.472	0.512	0.539	0.668	0.628
Mistral-8B	0.554	0.609	0.608	0.669	0.697
LLaMA-3.1-8B	0.641	0.642	0.678	0.749	0.793
LLaMA-3.2-3B	0.573	0.603	0.639	0.705	0.740
LLaMA-3.2-1B	0.451	0.530	0.545	0.674	0.631

Table 5: Average ARC$_{\text{score}}$ across CANLII$_{100}$, full CANLII, and DRI. Models are ordered by size. **Bold** indicates the human reference (upper bound), and *italic* marks the best-performing model in each dataset.

	CANLII$_{100}$				CANLII		DRI			
Model	MF	FE	MF	FE	MF	FE	MF	FE	MF	FE
Qwen-2.5-14B	**35.7**	2.0	**31.4**	3.2	**28.7**	3.1	**36.5**	0.2	**36.7**	0.5
Qwen-2.5-7B	**35.5**	2.0	**31.2**	3.6	**27.8**	3.2	**33.5**	0.5	**34.7**	1.0
Qwen-2.5-3B	**36.2**	2.8	**30.5**	3.4	**29.5**	3.6	**36.6**	0.8	**36.6**	0.7
Qwen-2.5-1.5B	**56.3**	3.6	**49.1**	5.5	**46.9**	4.9	**49.9**	0.4	**47.8**	1.1
Mistral-8B	**46.8**	3.0	**42.1**	3.1	**40.4**	3.7	**43.3**	0.4	**44.0**	1.3
LLaMA-3.1-8B	**40.1**	2.1	**36.9**	3.6	**33.1**	3.3	**41.1**	0.3	**41.7**	0.8
LLaMA-3.2-3B	**45.4**	3.0	**40.1**	4.0	**36.5**	4.2	**43.8**	0.3	**43.3**	1.3
LLaMA-3.2-1B	**56.3**	4.6	**49.1**	5.5	**45.1**	5.7	**49.9**	0.7	**50.0**	1.3

Table 6: Error rates (%) normalized by the total number of facts per dataset. Each dataset reports **Missing Facts (MF)** and **Factual Errors (FE)**. Columns shaded in gray correspond to **DeepSeek-R1-Distill-Qwen14B** as the judge. Higher error values are **bolded**.

observed across models.

5 Analyzing Summary Coverage with ARC

In this section, we use ARC to evaluate summaries produced by long-context LLMs on the CANLII and DRI datasets.

5.1 Obtaining Generated Summaries

We evaluate eight long-context open-weight LLMs from the LLaMA (Grattafiori et al., 2024), Mistral (Jiang et al., 2024) and Qwen (Yang et al., 2024a; Team, 2024) families that meet our computational constraints. Specifically, we include LLaMA-3.1-8B-Instruct, LLaMA-3.2-1B and 3B, Qwen-2.5-Instruct-1.5B, 3B, 7B, 14B, and Mistral-8B. We vary model sizes to examine whether scaling improves coverage performance. Inference is conducted using VLLM (Kwon et al., 2023) for efficiency and scalability, with Qwen models further extended to $130k$ tokens via RoPE scaling (factor 4.0).

Summaries are generated using greedy decoding ($T = 0$), capped at 2048 tokens to ensure fair long-form generation. Each document $d \in D$ is prompted with the following instruction: "Read the following text and summarize it: {input document}. Summarize in {reference summary word length} words. Summary:" We deliberately use a generic summarization prompt to evaluate LLMs without giving them any prior indication of which information is salient.[7] This setup mirrors the human annotation process, where experts also lacked explicit guidance on saliency when producing reference summaries. The target length is dynamically matched to the reference summary for comparability. For DRI, the target length is fixed to the longest reference summary to encourage maximal argument retention.

5.2 RQ1: Can LLM summaries cover salient arguments without explicit supervision?

We compute ARC$_{\text{score}}$ for all model-generated summaries described in Section 5.1, as well as for human reference summaries on both the CANLII and DRI benchmarks. For CANLII, we report results on both the smaller 100-article subset (CANLII$_{100}$) scored with GPT-4o and the full set (CANLII) scored with DeepSeek-R1-Distill-Qwen14B, thereby testing the robustness of ARC$_{\text{score}}$ across different judges and verifying that it assigns near-perfect scores to expert-written summaries. For DRI, because role-level annotations in summaries are unavailable, the goal is to assess how well human references capture salient argument roles compared to LLM-generated outputs.

Table 5 shows that argument coverage remains incomplete across all models and domains, indicating that LLMs could still benefit from argument-aware supervision. Although the two judges yield slightly different ARC$_{\text{score}}$ values, the overall ranking of models is consistent. [8] Two clear patterns emerge from the results. (1) *Larger models generally achieve higher coverage.* Within the Qwen-2.5 family, performance improves when scaling from 1.5B to 3B and 7B across both judges. A similar trend holds for LLaMA-3.2, where the 3B model surpasses the 1B variant. [9] (2) *Coverage is more challenging in legal opinions.* On DRI, the strongest model (Qwen-2.5-7B) achieves 0.799/0.806, closely approaching the human ref-

[7] See Appendix G for experiments with an argument-aware prompt.

[8] For CANLII$_{100}$, Kendall's $\tau = 0.94$ (between the judges scores); for DRI, $\tau = 0.83$; both $p < 0.05$.

[9] This scaling effect is examined only for models with 1.5B–14B parameters.

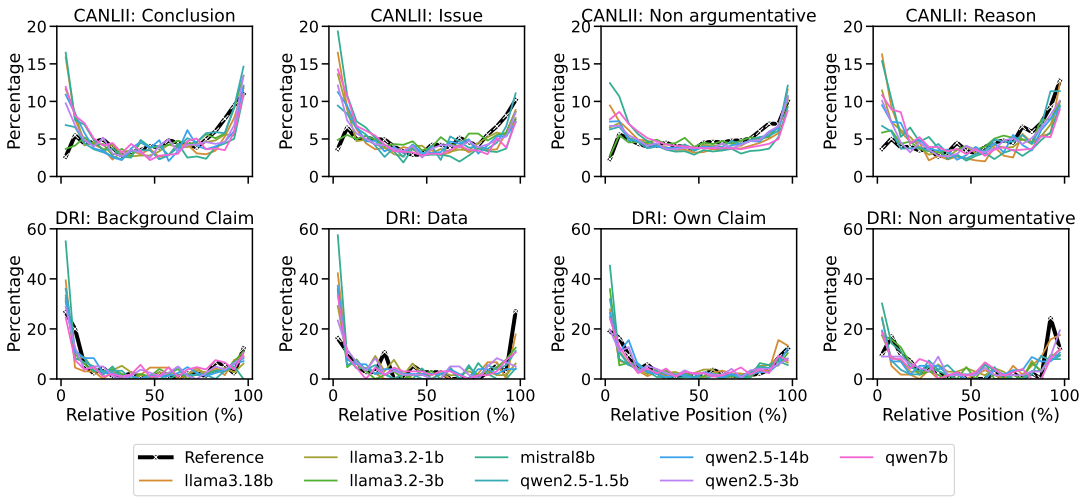

Figure 2: Source sentences relative position in the LLM context window across all models and various argument roles for both CANLII and DRI corpora.

erence (0.838/0.813) under both judges. By contrast, on CANLII_{100} and the full CANLII, the best-performing models plateau at 0.677/0.707 and 0.725, far below the human reference (0.986/0.977 and 0.982, respectively). This disparity underscores the difficulty of preserving salient argumentative content in legal texts, where arguments (both salient and non-salient) are sparsely distributed across lengthy contexts.

Fact-level Analysis. We analyze fact-level decisions (missing versus non-factual) to determine whether coverage errors arise primarily from omissions of key information or from factual inaccuracies. For each model, we compute the proportion in % of missing facts and factual errors across all generated summaries. As shown in Table 6, missing facts are the dominant source of error across both datasets, judges, and models, indicating that salient information is more often omitted than misrepresented. While factual inconsistencies do occur, they are consistently less frequent. These findings suggest that, beyond hallucination, the central challenge in summarization is achieving comprehensive coverage of salient content.

5.3 RQ2: Do argument positions in the source affect their coverage in summaries?

Following Ravaut et al. (2024), 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., 2024; Adams et al., 2023) by iteratively adding sentences in the source that maximizes ROUGE-1 score until there is no further improvement. [10] We analyze the source sentence indices by their argument role annotations. Figure 2 shows a clear U-shaped context window bias across all models, most pronounced in CANLII. Argument role analysis indicates that source positions are heavily shaped by this pattern—problematic for CANLII, where reference summaries do not follow fixed positional trends. In contrast, DRI references align more closely with the LLM bias distribution.

	CANLII_{100}		CANLII		DRI	
Model	ρ	ρ	ρ	ρ	ρ
Qwen-2.5-14B	-0.144	-0.131	-0.146	0.119	-0.044
Qwen-2.5-7B	-0.301	-0.237	-0.214	-0.223	0.007
Qwen-2.5-3B	-0.232	-0.112	-0.136	0.006	0.039
Qwen-2.5-1.5B	-0.230	-0.170	-0.163	0.074	-0.010
Mistral-8B	-0.369	-0.244	-0.193	-0.055	0.012
LLaMA-3.1-8B	-0.230	-0.088	-0.153	0.129	0.166
LLaMA-3.2-3B	-0.216	-0.113	-0.190	-0.091	-0.031
LLaMA-3.2-1B	-0.171	-0.087	-0.198	-0.020	-0.031

Table 7: Pearson correlation (ρ) between mean relative position of salient arguments and $\text{ARC}_{\text{score}}$. Gray columns = DeepSeek–R1, non-shaded = GPT-4o. Values in gray are not statistically significant (p>0.05).

We analyze how the position of salient arguments affects coverage as measured by $\text{ARC}_{\text{score}}$. For CANLII and CANLII_{100}, we apply greedy sentence selection while restricting the pool to annotated arguments that appear in both the reference summary and the input document. This reduces computational cost and ensures that only arguments included in the reference summary are mapped

[10] Appendix H contains the full algorithm.

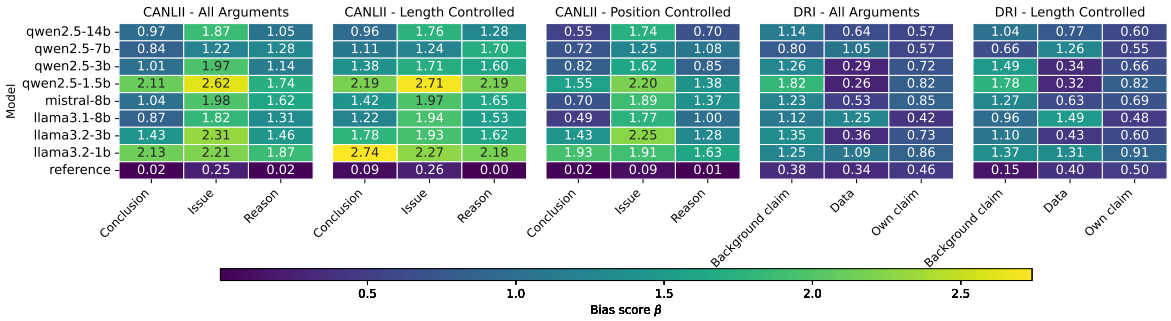

Figure 3: Bias score β across multiple argument roles for both $\texttt{CANLII}_{100}$ and DRI corpora with GPT4-o.

back to the input. For DRI, we directly identify the positions of arguments in the input document with a relevance score of 5. Following Ravaut et al. (2024), we compute the mean relative position of salient arguments and measure the Pearson correlation between this mean position and $\text{ARC}_{\text{score}}$.

Table 7 shows a consistent negative correlation in CANLII (both $\texttt{CANLII}_{100}$ with GPT4-o and DeepSeek–R1, as well as the full CANLII set with DeepSeek–R1), with several models reaching significance ($p < 0.05$). This confirms that LLM context windows systematically bias coverage in long legal cases. By contrast, correlations in DRI are weaker and largely non-significant, with some models (e.g., LLaMA-3.1-8B) even showing slight positive trends under DeepSeek–R1. These results align with Figure 2: in CANLII, reference summaries diverge from the strong U-shaped bias present in model outputs, whereas in DRI, reference arguments more closely mirror the positional distribution induced by the LLM context window.

5.4 RQ3: Are argument roles disproportionately covered?

We propose β as a role-specific bias score, designed so that higher values indicate stronger bias (i.e., poorer representation of a role). The intuition is that if a role r is perfectly represented in the summary, then $\text{ARC}_{\text{role}_r} = 1$, and the bias should vanish ($\beta_r = 0$). Conversely, lower coverage implies a larger gap from perfect representation, reflected in a higher β_r. Formally, we define:

$$\beta_r = (1 - \text{ARC}_{\text{role}_r}) \cdot \frac{1}{\log\left(1 + \frac{|r|_D}{|\text{args}|_D}\right)},$$

where $|r|_D$ is the frequency of role r in the source document D, and $|\text{args}|_D$ is the total number of arguments in D. The normalization term down-weights roles that are overrepresented in the source, ensuring that β_r captures true disparities in coverage rather than frequency effects.

To further reduce confounds from argument length and position—especially in CANLII,[11] where longer roles and arguments positioned in the middle of the document can bias coverage—we compute the bias score β within length-controlled groups, allowing a $\pm 20\%$ variation in word count. In addition, to control for positional bias effect on coverage, we restrict the positional analysis to cases where at least 80% of arguments fall within either the first or the last 20% of the document. Positions are computed using the relative positions of the indices of argumentative sentences from the beginning of the document. These constraints help ensure that the observed effects on role coverage are not confounded by argument length or positional bias.

Figure 3[12] shows that, for $\texttt{CANLII}_{100}$, β is consistently lowest for *conclusions*, indicating stronger coverage of this role across both the all-argument analysis and the position-controlled setting. Controlling for role length explains part of the variation across roles, but the relative advantage of *conclusions* persists. In DRI, by contrast, bias scores are generally lower than in $\texttt{CANLII}_{100}$, consistent with our earlier finding that LLMs capture more salient arguments in scientific summaries. Nonetheless, *background claims* tend to be covered less effectively, even when length is controlled. Across both domains, model size further influences role-level disparities: for instance, models with < 3B parameters consistently exhibit the highest β, reflecting weaker and less balanced argument coverage.

[11] This analysis uses $\texttt{CANLII}_{100}$ with the GPT4-o judge, as results in RQ1 and RQ2 show consistent trends across judges and dataset splits.

[12] Appendix I reports results without frequency normalization to rule out denominator inflation effects.

6 Conclusion and Future Work

We introduced ARC, a hierarchical evaluation framework for assessing how well LLM-generated summaries preserve salient argumentative content. ARC provides a principled and interpretable diagnostic tool for evaluating structured argument coverage in long-context summarization. Its hierarchical structure not only enables interpretability—revealing which roles are preserved and whether errors arise from omissions or factual inaccuracies—but also achieves higher correlation in holistic evaluation compared to lexical, semantic, entailment, and decomposition-based baselines. Applying ARC to legal and scientific domains uncovered two consistent limitations of current LLMs: *positional bias*, where the characteristic U-shaped context window negatively affects coverage, and *role bias*, where *conclusions* are favored over other roles such as *issues* and *reasons*. Future work can extend this framework by incorporating explicit argument structures into training or prompting, and by leveraging ARC's interpretable outputs to guide targeted improvements in summarization—particularly in high-stakes domains.

Limitations

While the ARC framework enables a comprehensive, multi-level evaluation of argument coverage, several limitations remain that suggest promising directions for future work.

Limited Benchmarks. Evaluation is constrained by the scarcity of benchmarks explicitly designed for argument coverage. Existing datasets provide limited annotation granularity, especially below the argument-unit level. To our knowledge, only the two benchmarks used here include both salience annotations and explicit argument roles. The small size of DRI (40 documents) also limits generalizability, motivating larger, rigorously annotated corpora—e.g., debates or financial texts—where arguments are central.

Decomposition Sensitivity. Our decomposition into atomic facts relies on the Yang et al. (2024b) algorithm, LLM-based prompting, and entailment filtering, which may introduce errors. While this decomposition proved more effective than that in FactScore (Min et al., 2023), it is worth noting that FactScore relied on human-authored examples from biographical texts, which might have contributed to their lower correlations. Future extensions of ARC should consider incorporating expert-written atomic facts to further improve consistency and balance.

Balancing Precision and Recall. ARC currently emphasizes *recall*—whether salient argument-role facts are preserved—similar to FactScore. While critical for high-stakes domains, this ignores *precision*, i.e., over-inclusion of non-salient content. Future work should jointly assess both dimensions, for instance via a harmonic mean, for a fuller view of coverage quality. [13]

Dependence on Gold Annotations. Our analysis assumes access to gold salience labels. Though this allows conclusive evaluation, future research should explore end-to-end systems that integrate salience detection, fact decomposition, and coverage assessment within a unified framework.

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.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2040490 and by Amazon. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. This research was partially supported by the CS50 fellowship awarded by the department of computer science at the University of Pittsburgh. We want to thank the members of the Pitt PETAL group, Pitt NLP group, and anonymous reviewers for their valuable comments in improving this work.

[13] We additionally conduct a precision–recall analysis on a subset of CANLII summaries; details and results are reported in Appendix J.

References

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.

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, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088.

William Gantt, Alexander Martin, Pavlo Kuchmiichuk, and Aaron Steven White. 2024. Event-keyed summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7333–7345, Miami, Florida, USA. Association for Computational Linguistics.

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*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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*.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Fables: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. 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.

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. 2023. G-eval:

Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

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.

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

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23428–23454, Vienna, Austria. Association for Computational Linguistics.

William Walden, Pavlo Kuchmiichuk, Alexander Martin, Chihsheng Jin, Angela Cao, Claire Sun, Curisia Allen, and Aaron Steven White. 2025. Cross-document event-keyed summarization. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 218–241, Vienna, Austria. Association for Computational Linguistics.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8791–8810, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024b. 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.

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 8 presents an excerpt from a legal opinion in the CANLII dataset, with arguments highlighted in both the input and the reference summary. Table 9 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.

A.2 Distribution of arguments across the input

Figure 4 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,

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 8: 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.

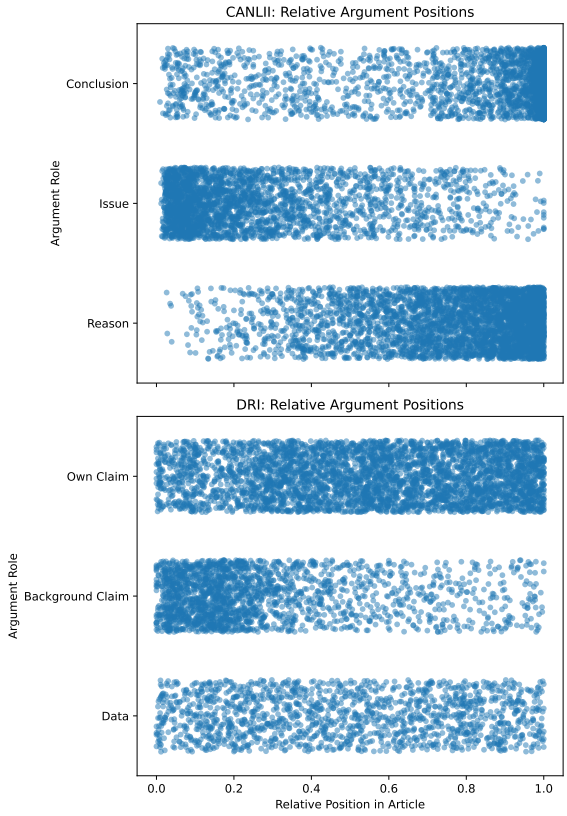

Figure 4: Distribution of argument roles in the input in both CANLII and DRI.

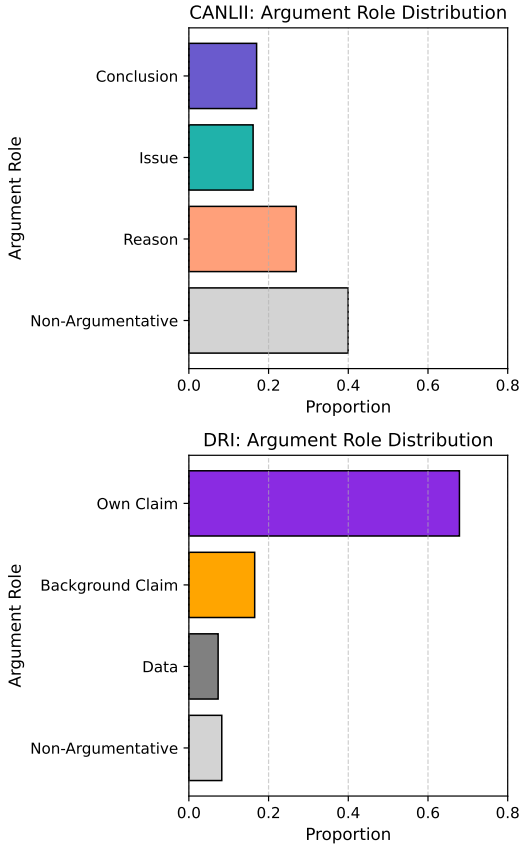

Figure 5: Argument role distributions in summaries for CANLII and DRI (for sentences with relevance score is 5). In CANLII, arguments are less densely represented compared to DRI, where own claims dominate.

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 5 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

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 9: 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.

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.

A.4 Rhetorical roles per relevance to summary Likert score

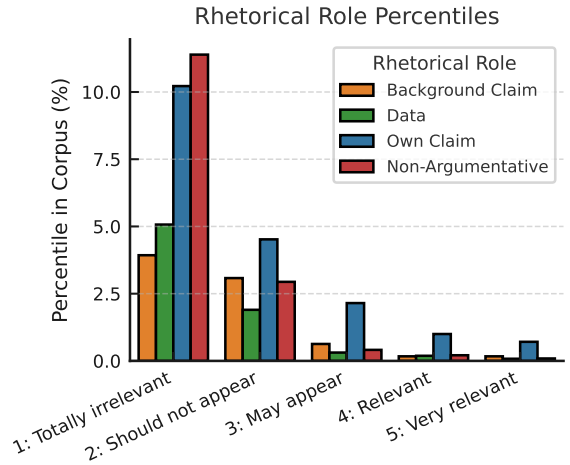

Figure 6: Argument roles per each relevance score to summary from 1 to 5.

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 6, 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 argument roles.

B Fact decomposition algorithm

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

Table 10 presents the prompt used to extract atomic facts. Table 11 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.

C Atomic fact error examples

Table 12 presents representative examples from the DRI dataset, illustrating model decisions on atomic facts as evaluated by GPT-4o.

Algorithm 1: Argument Decomposition and Entailment Filtering

Input: Argument a_i, Entailment Model $\mathcal{M}$, Entailment Threshold τ

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 prompt

Evaluation prompt for fact-level coverage $\texttt{ARC}_{\text{score}}$ is described in Table 13. We ask the LLM to generate a rationale before assigning a decision.

E Likert scores based on human evaluation

Table 14 shows the Likert scale from 1 to 4 definitions.

F Full expert correlation

Table 15 reports correlations with individual experts and their averaged ratings across both judges. $\texttt{ARC}_{\text{score}}$ (with `DeepSeek-R1-Distill-Qwen14B`) achieves the strongest overall correlation with expert judgments, surpassing all baselines. $\texttt{ARC}_{\text{score}}$ (with `GPT4-o`) yields the second-highest overall correlation, outperforming all metrics except $\texttt{SummaC}_{\text{ZS}}$ (sentence-level) on Expert 2. These results suggest that the decomposition in $\texttt{ARC}_{\text{score}}$ not only enables fine-grained and interpretable analyses across error types and argument-role distinctions, but also preserves—and even improves upon—holistic coverage evaluation compared to strong automatic baselines.

G Zero-shot summaries with argument-aware prompts

We perform a zero-shot ablation using four representative models from our set of eight: `Llama-3.2-1B`, `Llama-3.2-3B`, `Qwen-2.5-3B`, and `Qwen-2.5-7B`, varying both model families and sizes. In this setting, models are instructed to generate summaries of the same length as the human references, with an explicit emphasis on the key arguments in the input text. This experiment evaluates whether explicitly encoding argument salience in the prompt improves the inclusion of salient arguments in model-generated summaries. The zeroshot prompt used is described in Table 16.

Given the consistency across judges and datasets we report results on the smaller `CANLII` set ($\texttt{CANLII}_{100}$) and the full `DRI` set using `GPT-4-o` as a judge.

Table 17 shows that explicitly instructing models to focus on key arguments did not consistently improve performance and, in some cases, even degraded coverage of salient roles. Only `Qwen-2.5-3B,7B` achieved a modest improvement on `CANLII`. These findings suggest that simply prompting models to emphasize arguments is insufficient for enhancing saliency coverage, underscoring the need for deeper alignment or fine-tuning strategies that more effectively encode argumentative relevance.

H Source sentences identification

Algorithm 2 presents the lexical greedy procedure adopted from Ravaut et al. (2024) for identifying source sentences given a generated summary. We extend the original algorithm to return both the selected sentence indices and their corresponding argument role types, based on previously annotated sentence-level roles.

I Bias analysis for argument coverage without argument role normalization

While normalization in computing β corrects for frequency skew, it may also understate coverage for dominant roles with inherently high raw $\texttt{ARC}_{\text{role}_r}$ scores. To provide a fuller picture, we additionally compute role-specific bias directly as $1 - \texttt{ARC}_{\text{role}_r}$. As shown in Figure 7, results on $\texttt{CANLII}_{100}$ confirm prior findings: LLMs disproportionately prioritize *conclusions* over *issues* and *reasons*, both under length-controlled and non-controlled settings. For `DRI`, removing frequency normalization explains partially the higher β $\texttt{ARC}_{\text{role}_{\text{background claim}}}$ scores reported in Section 5.4.

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

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

J Precision–recall trade-offs in ARC

ARC is designed to prioritize *recall*, aligning with prior coverage-based factuality metrics (e.g., FactScore) that emphasize capturing as many salient facts as possible in the summary. Nevertheless, we believe that incorporating precision-based analyses provides a more complete view of different models behavior. To this end, we conducted an additional study on 100 summaries from the CANLII dataset, evenly sampled across summary length buckets. Using the same atomic fact decomposition and entailment filtering setup as in the main experiments, we compute: **Precision**: proportion of covered salient facts divided by the number of facts generated in the summary. **Recall**: proportion of reference salient facts covered by the summary. **F1**: harmonic mean of precision and recall. Both atomic fact decomposition and evaluation were conducted using GPT4-o.

Table 18 reports precision, recall, and F1 scores across models.

To further understand how evaluation choice affects model comparison, Table 19 shows model rankings induced by each metric.

Finally, Table 20 reports rank correlation coefficients between rankings induced by different metrics.

Overall, while precision introduces moderate reordering among models, the overall relative ranking of models remains largely stable. The positive correlations between recall and precision indicate that ARC's recall-oriented design captures a general performance trend that largely persists even under

Example 1 – Non-factual

Summary: The paper presents an efficient fluid simulation method using Laplacian eigenfunctions as a velocity basis, which avoids the high computational complexity of traditional Eulerian and Lagrangian approaches. This method is particularly advantageous for incompressible fluid dynamics, offering a principled way to reduce the dimensionality of the Navier–Stokes equations. The authors argue that Laplacian eigenfunctions provide a natural correspondence with spatial scales of vorticity, allowing precise control over turbulent energy at different scales. They show that for simple domains, these eigenfunctions have closed-form expressions, enabling fast and exact evaluation of the velocity field. The method is formulated using Galerkin projection onto Laplacian eigenfunctions, which are divergence-free and satisfy free-slip boundary conditions. For complex domains, a discrete exterior calculus (DEC) approach is used to compute the basis fields. The paper highlights the benefits of this method, including its ability to simulate fluid motion interactively, its low memory and precomputation requirements, and its flexibility in controlling viscosity and external forces. However, it also acknowledges limitations, such as stair-step artifacts on irregular boundaries and the need for additional methods to handle moving obstacles. The authors demonstrate that the method can be applied to various domains, including simple rectangles and complex 3D cavities, and discuss its potential for expressive control and optimization in fluid dynamics. The paper concludes by noting that while the method is particularly suited for gaseous phenomena, it may require further development for liquid simulations.

Fact: Our basis has a natural correspondence with spatial scales of vorticity that is lacking in [Treuille et al. 2006].
Decision: Non-factual
Explanation: The summary correctly states that the basis corresponds to vorticity scales but omits the comparative qualifier "lacking in [Treuille et al. 2006]," misleadingly implying a novel rather than comparative contribution.

Example 2 – Missing Fact

Summary: Pose space deformation (PSD) is a technique that unifies and improves upon both shape interpolation and skeleton-driven deformation methods. It represents deformations as mappings from a pose space to displacements in the object's local coordinate frames. PSD allows direct manipulation of desired shapes and maintains the performance benefits of traditional methods. Key applications include facial and body deformation in entertainment, telepresence, and gaming. The limitations of shape interpolation and skeleton subspace deformation (SSD) motivate the development of PSD. Shape interpolation lacks smoothness and requires frequent sculpting of new key shapes. SSD, while versatile, suffers from the "collapsing elbow" problem and is difficult to control. PSD addresses these issues by interpolating deformations in a pose space, enabling smooth and controllable deformations. PSD requires scattered data interpolation in high-dimensional spaces. Gaussian radial basis functions are used due to their well-behaved nature and ease of implementation. The algorithm is simple, general, and allows real-time synthesis, making it suitable for high-resolution models. Facial animation can benefit from PSD by allowing sculpting of intermediate expressions and multidimensional pose spaces. PSD can handle various deformation scenarios uniformly, from simple joints to complex secondary animations. The setup and synthesis costs are minimal, ensuring real-time performance. While anatomically based models offer higher quality, PSD provides a practical alternative for real-time applications and fanciful creature designs.

Fact: The synthesis cost is only slightly more than that of shape interpolation.
Decision: Missing Fact
Explanation: The summary omits the comparative detail that PSD's synthesis cost is slightly higher than that of shape interpolation, overstating the efficiency of the proposed method.

Table 12: Full examples of $\text{ARC}_{\text{atomic}}$ factuality and coverage decisions. Red rows denote **non-factual** content, and yellow rows denote **missing facts**. Each example includes the full model-generated summary followed by the evaluated fact, decision, and an explanatory rationale.

precision-aware evaluation.

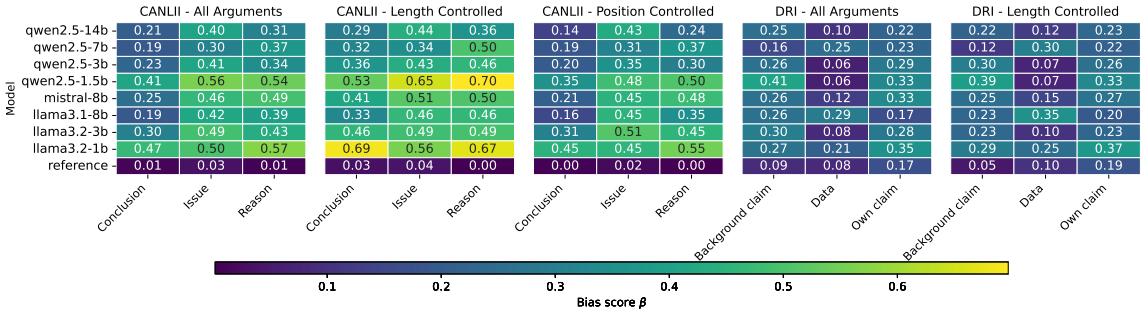

Figure 7: Bias score without any frequency normalization (β=1-ARC$_{\text{role}}$) across multiple argument roles (controlled length and non-controlled length) for both CANLII$_{100}$ and DRI corpus.

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.

Algorithm 2: Lexical Greedy Source Sentence Identification with Argument Roles

Input: Source document sentences $\mathcal{S} = \{s_1, \ldots, s_n\}$, summary y, argument role annotations $\mathcal{A}$

Output: Selected sentence indices $\mathcal{I}$ and corresponding argument roles $\mathcal{R}$

$\mathcal{I} \leftarrow \emptyset$ $R_{\text{best}} \leftarrow 0$
repeat
  until ;
  $R_{\text{prev}} \leftarrow R_{\text{best}}$ $i^* \leftarrow$ None
  foreach $s_i \in \mathcal{S} \setminus \mathcal{I}$ **do**
    Compute $R_i \leftarrow \text{ROUGE-1}(y, \text{concat}(\mathcal{I} \cup \{i\}))$
    if $R_i > R_{best}$ **then**
      $R_{\text{best}} \leftarrow R_i$ $i^* \leftarrow i$
  if $i^* \neq$ None **then**
    $\mathcal{I} \leftarrow \mathcal{I} \cup \{i^*\}$ $R_{\text{best}} \leq R_{\text{prev}}$
$\mathcal{R} \leftarrow \{\mathcal{A}(i) \mid i \in \mathcal{I}\}$
return $\mathcal{I}, \mathcal{R}$

Rating	Explanation of the Generated Summary Rating Scale
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 14: Likert scale exact meaning for each score based on definitions obtained from Elaraby et al. (2024).

Metric	Expert 1		Expert 2		Average	
	τ	ρ	τ	ρ	τ	ρ
ROUGE-1	0.378	0.520	0.347	0.456	0.391	0.539
ROUGE-2	0.363	0.441	0.276	0.419	0.336	0.475
ROUGE-L	0.310	0.417	0.337	0.450	0.345	0.479
BERTScore	0.319	0.409	0.292	0.398	0.354	0.517
SummaC$_{ZS}$ (sent)	0.310	0.517	*0.427*	*0.517*	0.387	0.537
SummaC$_{ZS}$ (doc)	0.476	0.531	0.262	0.344	0.375	0.476
SummaC$_{conv}$ (sent)	0.329	0.371	0.341	0.460	0.345	0.459
SummaC$_{conv}$ (doc)	0.408	0.300	0.269	0.264	0.352	0.311
FactScore (GPT4-o)	0.365	0.511	0.362	0.460	0.405	0.549
ARC$_{score}$ (GPT-4o)	*0.499*	*0.607*	0.401	0.482	*0.465*	*0.593*
ARC$_{score}$ (DeepSeek-R1)	**0.545**	**0.653**	**0.441**	**0.519**	**0.509**	**0.638**

Table 15: Metrics correlations (Kendall's τ, Pearson's ρ) with expert judgments using 87 articles ($\kappa = 0.605$). All rows: $p < 0.05$. Color legend: ARC$_{score}$ (GPT-4o), ARC$_{score}$ (DeepSeek-R1-Qwen-14B). **Bolded** is best overall and *italicized* is second best.

Argument-Aware Zero-Shot Prompt

Input:

Read the following text carefully:

`{input}`

Task Description:

Write an **abstractive summary** in about **{length}** words.

Instructions:

- Focus primarily on the **key arguments** presented in the text.
- Ensure the summary is **coherent, fluent, and written in continuous prose** (not bullet points or key phrases).

Output:

Summary:

Table 16: Argument-aware zero-shot prompt used for abstractive summarization.

Model	CANLII		DRI	
	Vanilla	Arg-Prompt	Vanilla	Arg-Prompt
LLaMA-3.2-1B	**0.451**	0.426	**0.674**	0.645
LLaMA-3.2-3B	**0.573**	0.570	**0.705**	0.667
Qwen-2.5-3B	0.648	**0.654**	**0.730**	0.718
Qwen-2.5-7B	0.676	**0.703**	**0.799**	0.754

Table 17: Effect of argument-aware prompting on ARC$_{score}$ with GPT-4o as the evaluator. Vanilla results are from Table 5. **Bolded** indicates highest across both vanilla and arg-prompted settings.

Model	Precision	Recall	F1
Qwen-2.5-14B	0.8766	0.6418	0.7149
Qwen-2.5-7B	0.6337	0.6374	0.5737
Qwen-2.5-3B	0.5027	0.5874	0.5305
Qwen-2.5-1.5B	0.7733	0.5608	0.6227
Mistral-8B	0.7695	0.5188	0.5724
LLaMA-3.1-8B	0.7734	0.6174	0.6473
LLaMA-3.2-3B	0.5622	0.3983	0.4354
LLaMA-3.2-1B	0.4618	0.4069	0.4080

Table 18: Precision, recall, and F1 scores computed using ARC$_{score}$ on 100 CANLII summaries (rows ordered by model family and size).

Model	Precision Rank	Recall Rank	F1 Rank
Qwen-2.5-14B	1	1	1
Qwen-2.5-7B	5	3	5
Qwen-2.5-3B	6	5	4
Qwen-2.5-1.5B	3	4	3
Mistral-8B	4	6	6
LLaMA-3.1-8B	2	2	2
LLaMA-3.2-3B	7	8	7
LLaMA-3.2-1B	8	7	8

Table 19: Model rankings induced by precision, recall, and F1 (rows ordered by model family and size).

Metric Pair	Kendall τ	Pearson r
Precision vs. Recall	0.50	0.62
Precision vs. F1	0.86	0.95
Recall vs. F1	0.64	0.81

Table 20: Rank correlations between different evaluation metrics ($p < 0.05$).