Improving Argument Saliency Coverage in Small LLMs for Long Legal Opinion Summarization via Sequence-Level Distillation

Anonymous ACL submission

Abstract

We show that sequence-level distillation from a capable long-context teacher model is a simple, annotation-free, and data-efficient strategy for improving argument saliency coverage in long legal opinion summarization, where small LLMs often struggle to retain the most salient argumentative content. Across student model sizes, distillation consistently surpasses tuning on expert-written summaries. We further demonstrate that most gains are achieved with as few as ∼10 training summaries, highlighting the strong data efficiency of teacher-generated supervision. Finally, we find that summary distillation is sufficient for improvements: reasoning-chain distillation remains competitive with summary-only distillation, but provides marginal benefit when combined with summary supervision.

1 Introduction

Summarizing long legal opinions requires preserving salient argument roles that are sparse across long inputs yet critical to include in the final summary (Xu et al., 2021; Elaraby and Litman, 2022). Recent work shows that LLMs still struggle to prioritize the same salient information as humans in summarization tasks (Trienes et al., 2025), a challenge that becomes especially apparent in legal opinion summarization, where important argumentative content is dispersed across thousands of tokens (Elaraby and Litman, 2026). This problem is further amplified in smaller models, which limits the development of efficient and reliable summarization systems for high-stakes legal settings.

While tuning models on expert-written summaries is a plausible solution, obtaining domain-expert annotations remains costly and difficult to scale. Sequence-level distillation (Kim and Rush, 2016), which transfers generation behavior from a stronger teacher through teacher-generated sequences, offers an annotation-free alternative. Recent work has explored related forms of summarization distillation by training smaller models on LLM-generated summaries (Liu et al., 2024; Zhu et al., 2025) or prompted structured rationales (Jiang et al., 2024; Wang et al., 2025). However, prior work has largely focused on news-style benchmarks, such as `CNN/DailyMail` (Hermann et al., 2015) and `XSum` (Narayan et al., 2018), emphasizing faithfulness, fluency, or generic summary quality rather than saliency coverage in long, high-stakes domains. We therefore ask: *To what extent can sequence-level distillation improve argument saliency coverage in long legal opinion summarization?* We empirically study this question using small long-context LLMs ($<$ 10B parameters) from the `Qwen3` family. Motivated by recent advances in reasoning-capable LLMs, we examine whether *native reasoning traces* provide useful supervision beyond teacher-generated summaries. While intermediate reasoning supervision has been shown to improve smaller models on reasoning-intensive tasks (Guo et al., 2025), its value for long-context summarization remains unclear.

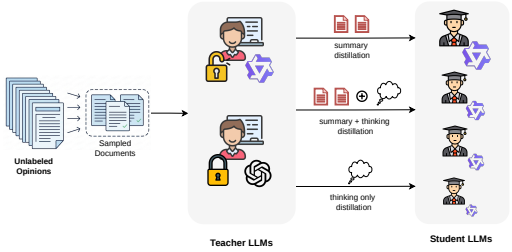

Figure 1: Distillation strategies from proprietary (🔒) and open-weight (🔓) teacher models to student models of various sizes.

We compare three supervision strategies (Figure 1): *(1) summary-only*, trained on teacher-generated summaries; *(2) reasoning-trace-only*, trained only on the teacher's native reasoning trace; and *(3) reasoning-trace + summary*, trained on

both signals. We evaluate proprietary and open-weight teachers, comparing distillation against expert-summary tuning and structured inference-time planning baselines. Finally, we vary the number of distillation examples to study data efficiency, a question largely unexplored in prior summarization distillation work, which typically relies on thousands of generated summaries.

Our findings are twofold. First, *sequence-level distillation is an effective annotation-free strategy for improving argument saliency coverage in long legal summarization*: it consistently surpasses both tuning on expert-written summaries and structured inference-time planning baselines. Second, *distillation is simple and data-efficient*: most gains emerge with as few as ~ 10 training documents, and teacher-generated summaries alone are sufficient, with reasoning-trace supervision providing only marginal additional benefit.

2 Related Work

Legal Opinion Summarization and Saliency Coverage. While prior work on LLM-based legal opinion summarization has identified faithfulness as a persistent issue (Heddaya et al., 2025; Zhong and Litman, 2025), recent work shows that coverage gaps of salient argumentative content can be even larger (Elaraby and Litman, 2026). Existing approaches to legal coverage rely on external argument labeling (Elaraby and Litman, 2022) or reranking pipelines (Elaraby et al., 2023); in news, coverage is improved through structured planning over entities (Adams et al., 2023) or events (Gantt et al., 2024; Walden et al., 2025), which we adapt as a baseline. *Unlike the news domain, which often involves short-context, entity-focused settings, we study sequence-level distillation for argument saliency coverage in long legal opinions, where salient content is sparse across thousands of tokens.*

Distillation from LLMs for Summarization. Sequence distillation has long been used to transfer generation behavior from larger models to smaller ones (Kim and Rush, 2016). Recent work applies this idea to summarization by training smaller models on LLM-generated summaries (Liu et al., 2024), as well as using richer intermediate supervision signals such as structured rationales (Jiang et al., 2024; Wang et al., 2025). However, prior work has largely focused on general-domain or news-style benchmarks, emphasizing fluency, faithfulness, or generic summary quality. Long legal opinions pose a different challenge: salient argumentative content is sparse and dispersed across thousands of input tokens, making content selection and omission errors central to summary quality. *We extend distillation work to empirically study its effect in improving argument saliency in small LLMs.*

3 Distillation Setup

Given a prompt P and a legal opinion L, a teacher model M generates a native reasoning trace τ followed by a summary S, denoted $(\tau, S) \sim M(P, L)$. We empirically study sequence distillation under three supervision settings differing only in the supervision sequence $\mathcal{G}$. Student models m are trained using supervised fine-tuning (SFT) (Ouyang et al., 2022):

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{|\mathcal{G}|} \sum_{t=1}^{|\mathcal{G}|} \log P_m(g_t \mid g_{<t}, P, L), \quad (1)$$

where $\mathcal{G}$ denotes the supervised target sequence. We use these settings to test whether the teacher's native reasoning trace provides supervision beyond the final summary. In particular, while standard rationale-distillation setups often supervise both reasoning and the final answer, our reasoning-only setting optimizes only over τ, masking the final summary from the loss. This tests whether the reasoning trace alone can transfer the teacher's intermediate saliency-selection process, rather than simply teaching the student to imitate summary outputs. This yields three settings: *(1) summary-only distillation*, $\mathcal{G} = S$; *(2) reasoning distillation*, $\mathcal{G} = \tau$; and *(3) reasoning + summary distillation*, $\mathcal{G} = \tau \oplus S$.

4 Experimental Setup

4.1 Dataset

We use the Canadian Legal Information Institute (`CANLII`) dataset, which contains long legal opinions paired with expert annotations of argumentative roles relevant for summarization (Xu et al., 2021). Argument roles are annotated using the **IRC** scheme (Xu et al., 2021), consisting of **Issues** (legal questions raised in a case), **Reasons** (legal principles and justifications underlying decisions), and **Conclusions** (final rulings resolving the issues).[1] The full dataset contains approximately 28k legal

[1] Appendix A shows examples of argument roles.

opinion–summary pairs. A subset of 1,049 CANLII opinions is annotated at the sentence level with argumentative roles for both source documents and summaries (Xu et al., 2021),[2] enabling evaluation of argument saliency coverage; we use these as our test split. For training, we sample 1,000 unannotated legal opinions that do not overlap with the annotated test set, truncating inputs to 4,096 tokens. This is motivated by prior work showing that supervised fine-tuning on shorter contexts can generalize to longer-context settings (Gao et al., 2025; Grattafiori et al., 2024). Training documents average 2,990 words, compared to 4,382 words in the test split.

4.2 Models

We use two teacher models representing complementary settings: the proprietary `GPT-5-mini` and the open-weight `Qwen3-14B`. Since proprietary models do not expose reasoning traces, reasoning-trace distillation is applied only with `Qwen3-14B`. As student models, we use the `Qwen3` family ranging from 0.6B to 8B parameters, allowing controlled comparison across model sizes under a shared architecture. Qwen3 additionally supports reasoning-enabled and reasoning-disabled inference[3] using the same underlying weights, enabling us to separate the effect of test-time reasoning from distillation itself. For each of the 1,000 training documents, teachers generate summaries of approximately 250 words, matching the average expert summary length in the test split (270 words), using temperature 0.7 and nucleus sampling ($p = 0.95$).[4]

4.3 Baselines and Evaluation

We compare distillation against three baselines. The **zero-shot** (**B**) baseline generates summaries from each student in thinking and non-thinking modes. The **expert tuning** (**E**) baseline fine-tunes each student on expert-written summaries from the same 1,000 documents, enabling direct comparison between teacher-generated and expert supervision. Motivated by prior structured planning work (Adams et al., 2023; Wang et al., 2025), the **Chain-of-Arguments** (**CoA**) baseline disables native thinking and prompts the student to generate an argument plan before the final summary,[5] isolating inference-time planning from fine-tuning.

[2]Data obtained under a license agreement with CANLII.
[3]via the `/no_think` prompt suffix.
[4]Training details and prompts are in Appendix B.
[5]Prompt in Appendix C.

Setup	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B
Baselines				
B	.408±.219	.522±.218	.631±.188	.658±.184
	.514±.222	.530±.209	.635±.187	.654±.186
E	.599±.176	.599±.176	.571±.206	.526±.235‡
	.600±.175	.618±.210	.613±.197	.510±.229‡
CoA	.578±.219	.624±.203	.679±.171	.698±.169
Distillation (GPT-5-mini Teacher)				
S	.553*±.213	.652*±.196	.712*±.175	.741*±.168
	.554*±.210	.658*±.193	.714*±.172	.735*±.170
Distillation (Qwen3-14B Teacher)				
S	.621*±.209	.717*±.177	.744*±.160	.781*±.150
	.628*±.204	.723*±.177	.741*±.165	.784*±.151
R	.622*±.200	.718*±.175	.742*±.163	.779*±.154
	.632*±.200	.716*±.179	.745*±.165	.774*±.161
SR	.630*±.210	.722*±.177	.751*±.162	.777*±.154
	.632*±.199	**.733***±.169	**.753***±.161	**.785***±.154

Table 1: $\text{ARC}_{\text{score}}$ (mean±std.). **B**=zero-shot; **E**=expert-tuning; **CoA**=Chain-of-Arguments prompting; **S**=summary-only; **R**=reasoning-only; **SR**=summary+reasoning. Shaded rows use thinking inference. */‡ denote significantly above / below **B** using Mann–Whitney U test ($p < 0.05$). **Bold = best.**

We evaluate argument saliency coverage using $\text{ARC}_{\text{score}}$ (Elaraby and Litman, 2026), which compares generated summaries S against the set of expert-annotated argumentative roles $\mathcal{A}$. The metric computes the proportion of arguments in $\mathcal{A}$ supported by S using an LLM-based atomic verifier, and has shown the strongest correlation with expert coverage judgments on a subset of the CANLII test set (Elaraby et al., 2024). Unlike decomposition-based metrics such as `FactScore` (Min et al., 2023) or claim-verification approaches such as `MiniCheck` (Tang et al., 2024), $\text{ARC}_{\text{score}}$ distinguishes between *Missing Errors* (**ME**), where salient arguments are omitted, and *Factual Errors* (**FE**), where generated summaries contradict one or more salient arguments, enabling finer-grained analysis of whether gains stem from improved coverage or factual consistency. [6]

5 Results and Discussion

5.1 Distillation Findings

Table 1 presents our main results using 1,000 training documents capped at 4,096 tokens.[7] (1) *Distillation significantly improves saliency coverage.* Across all student model sizes, distillation yields statistically significant improvements over the zero-

[6]$\text{ARC}_{\text{score}}$ details in Appendix D
[7]Increasing training context length yields only marginal additional gains; see Appendix E.

Setup	Qwen3-0.6B		Qwen3-1.7B		Qwen3-4B		Qwen3-8B	
	FE	ME	FE	ME	FE	ME	FE	ME
Baselines								
B	10.20	55.87	11.82	43.65	12.22	31.76	11.86	30.01
	10.85	47.28	11.84	42.78	12.37	31.49	12.17	29.84
Distillation (GPT-5-mini Teacher)								
S	14.40	36.31	13.77	28.33	12.15	22.48	10.87	21.12
	14.22	36.71	13.96	27.20	11.91	22.40	10.89	20.56
Distillation (Qwen3-14B Teacher)								
S	12.22	31.31	10.89	22.42	10.37	19.78	9.05	17.40
	12.59	29.87	10.82	22.36	10.39	20.24	8.38	17.74
R	12.02	31.20	10.20	22.78	10.42	19.88	8.28	18.57
	12.66	30.69	10.24	23.46	10.48	19.79	8.75	18.68
SR	12.23	30.31	11.06	22.36	9.72	19.71	8.31	17.75
	12.55	30.67	10.21	22.33	9.59	19.56	8.23	17.53

Table 2: Fact Errors (**FE**) and Missing Errors (**ME**) in % per distillation setup from $\text{ARC}_{\text{score}}$ diagnostic evaluation. Setup abbreviations and row shading follow Table 1.

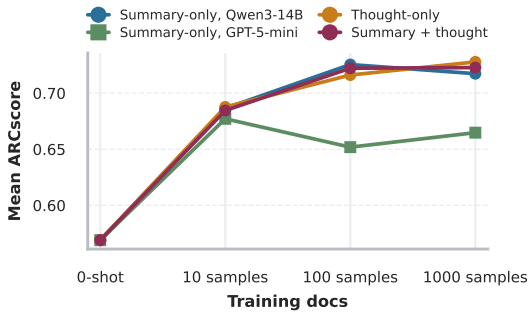

Figure 2: Few-shot distillation averaged across students.

shot baseline (**B**; $p < 0.05$). It also consistently outperforms both expert-summary tuning (**E**) and the Chain-of-Arguments (**CoA**) inference-time planning baseline, suggesting that sequence-level distillation is more effective for transferring argument saliency behavior than either limited expert supervision or structured prompting alone. (2) *Distillation performance is teacher-dependent.* While both `GPT-5-mini` and `Qwen3-14B` improve coverage, `Qwen3-14B` consistently achieves the strongest results. We hypothesize that this partly reflects closer distributional alignment between the open-weight teacher and the `Qwen3` student family.[8] By contrast, `GPT-5-mini` yields smaller gains and underperforms **CoA** for the smallest 0.6B student. (3) *Summary-only supervision is sufficient.* Native-reasoning distillation (**R**) remains competitive with summary-only distillation (**S**), while combining reasoning traces and summaries (**SR**) yields only marginal gains over **S** (+0.001–+0.011). This suggests that, for argument saliency coverage, teacher-generated summaries already encode much of the useful supervision signal. This contrasts with reasoning-intensive settings, where intermediate rationale or reasoning supervision can substantially improve smaller models (Hsieh et al., 2023; Guo et al., 2025; Muennighoff et al., 2025).

Table 2 provides a diagnostic view of how distillation affects different error types. We observe that *Missing Errors* (**ME**) — the dominant error type in the zero-shot baseline — decrease consistently across student sizes, whereas *Factual Errors* (**FE**) decrease only for larger students ($\geq 4B$). We focus this breakdown on distillation settings, as the central diagnostic question is whether distillation gains stem from improved coverage or improved faithfulness. For example, Qwen3-14B summary distillation reduces ME from 55.87% to 31.31% for the 0.6B student, while FE slightly increases from 10.20% to 12.22%. In contrast, the 8B student shows reductions in both ME (30.01% → 17.40%) and FE (11.86% → 9.05%).[9]

5.2 Effect of Scaling Distillation Samples

Figure 2 shows that argument coverage improves significantly with as few as 10 training documents across both proprietary and open-weight teachers (Mann-Whitney U test, $p < 0.05$). Scaling data further yields diminishing returns: `Qwen3-14B` shows only marginal gains beyond 100 documents and near-saturation at 1000.[10] *GPT-5-mini*, however, follows a less stable trajectory—coverage drops at 100 documents before partially recovering at 1,000—further confirming that scaling training data yields diminishing returns regardless of the teacher. [11]

6 Conclusion and Future Work

We show that sequence-level distillation is a simple, highly data-efficient strategy for improving argument saliency in long legal summarization, consistently outperforming expert-supervised tuning with as few as 10 training examples. Our results demonstrate that summary-only supervision is sufficient, as reasoning-chain distillation provides only marginal gains. Future work will explore alternative distillation objectives and extend this framework to additional long-context domains and evaluation metrics beyond argument saliency.

[8] Appendix F provides supporting perplexity analysis.

[9] Appendix G visualizes the relative change in FE and ME.

[10] Higher LoRA ranks do not help; see Appendix H.

[11] Detailed breakdowns are in Appendix I.

Limitations

Distillation and Learning Strategy. This work focuses on simple sequence-level distillation using standard supervised fine-tuning (SFT), motivated by isolating the effect of distillation itself. While effective, prior work suggests that alternative objectives such as contrastive learning or preference optimization may provide stronger supervision signals than standard SFT. Additionally, all experiments employ parameter-efficient LoRA tuning rather than full-parameter fine-tuning, largely due to GPU constraints. Exploring whether the observed trends persist under stronger optimization strategies remains an important direction for future work.

Limited Domain and Evaluation Scope. Our experiments focus on argument saliency coverage in long legal opinion summarization using the CANLII dataset. While argument coverage is particularly important in legal reasoning, coverage-related failures also arise in other domains. Finally, our setup assumes access to unlabeled in-domain legal opinions for teacher distillation, which may not always be available in specialized high-stakes settings. Future work should explore out-of-distribution adaptation.

Subjectivity of Saliency. Our evaluation follows the $\text{ARC}_{\text{score}}$ decomposition of arguments into atomic facts, treating all atomic facts as equally important. However, in practice, some legal arguments or supporting facts may carry greater saliency than others depending on the downstream use case. Future work should explore learning saliency weighting schemes from domain-expert feedback, although obtaining such annotations remains costly in domains such as law.

Ethical Statement

While our approach improves argument saliency coverage, generated summaries may still omit important details or contain unsupported claims, and therefore should not be used as a substitute for professional legal analysis. Our distillation approach may also inherit biases or systematic preferences from teacher models. Finally, our experiments rely on publicly available legal documents from the CANLII dataset and do not involve private or personally identifiable information. Our approach is not intended for legal decision-making or autonomous legal advice.

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.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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 and Diane Litman. 2026. ARC: Argument representation and coverage analysis for zero-shot long document summarization with instruction following LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3626–3643, Rabat, Morocco. Association for 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.

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.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399, Vienna, Austria. 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, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, and 1 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638.

Mourad Heddaya, Kyle MacMillan, Hongyuan Mei, Chenhao Tan, and Anup Malani. 2025. CaseSumm: A large-scale dataset for long-context summarization from U.S. Supreme Court opinions. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1917–1942, Albuquerque, New Mexico. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Pengcheng Jiang, Cao Xiao, Zifeng Wang, Parminder Bhatia, Jimeng Sun, and Jiawei Han. 2024. TriSum: Learning summarization ability from large language models with structured rationale. 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 2805–2819, Mexico City, Mexico. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of EMNLP*.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. 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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China. Association for Computational Linguistics.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/lora/.

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.

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.

Linyong Wang and 1 others. 2025. Distilling structured rationale from large language models to small language models for abstractive summarization. In *Proceedings of AAAI*.

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.

Yang Zhong and Diane Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document 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 2050–2073, Albuquerque, New Mexico. Association for Computational Linguistics.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2025. Factual dialogue summarization via learning from large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4474–4492.

A Argument Roles in CANLII

Figure 3 illustrates an example of argumentative roles within an annotated legal opinion. The original annotations from (Xu et al., 2021) include only role labels, while the argumentative relations shown in the figure are added for illustration purposes.

B Student Training and Summarization Prompting

B.1 Student Model Training

Training Setup. We fine-tune all student models using QLoRA (Dettmers et al., 2023) with 4-bit quantization, applying LoRA adapters to both attention and linear layers which shows comparable performance to full-finetuning across downstream applications (Schulman and Lab, 2025). We use LoRA rank $r = 32$ and scaling factor $\alpha = 64$, selected via grid search on `Qwen3-1.7B` over $r \in \{4, \ldots, 32\}$ and $\alpha \in \{8, \ldots, 64\}$. All models are trained for 3 epochs using the `AdamW` optimizer (Loshchilov and Hutter).

B.2 Reasoning-Trace-Only Distillation

In the reasoning-trace-only setting (**R**), the student is trained on teacher-generated reasoning traces without access to the final summary. To obtain these traces, we replace the standard summarization prompt (Table 4) with the prompt shown in Table 3.

"Think through how you would summarize the following text in {word_count} words. Output only your reasoning process, and do not provide the final summary."

Table 3: Prompt used for reasoning-trace-only distillation.

Naively training only on reasoning traces caused degeneration at inference time: when prompted to summarize, the model often continued generating reasoning traces instead of summaries. To avoid this behavior, we retain the standard summarization prompt during training while computing the SFT loss only over reasoning-trace tokens.

B.3 Summarization Prompt

Table 4 shows the prompt used for teacher summary generation and supervised fine-tuning of student models. The prompt is adapted from prior work on instruction-following summarization models (Ravaut et al., 2024; Elaraby and Litman, 2026).

Summarization Prompt
Read the following text and summarize it: {input}. Summarize the above text in {length} words. **Summary:**

Table 4: Prompt used for summary generation.

C Chain-of-Arguments Prompting

Table 5 shows the prompt used for generating a structured rationale grounded in argumentative definitions. Unlike distillation, this approach avoids fine-tuning by relying on inference-time argument planning. However, it requires domain-specific argument definitions and prior task knowledge, which may not generalize across domains. Additionally, inference-time planning introduces ad-

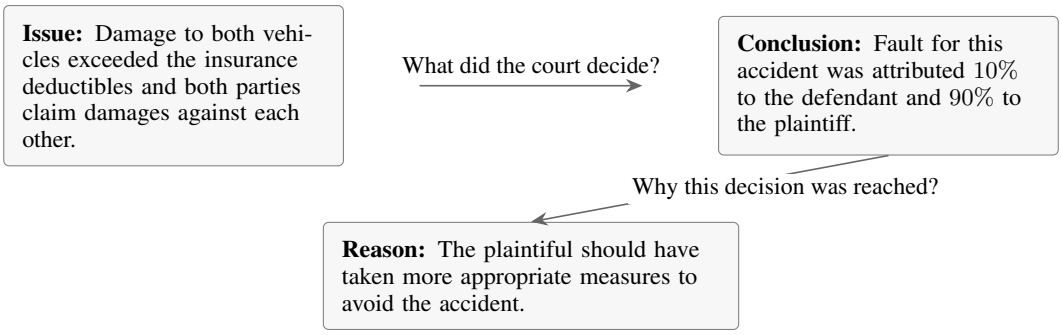

Figure 3: Example argument graph from a long legal opinion (CANLII): issue, conclusion, and supporting reason connected with labeled relations.

ditional token generation overhead that standard fine-tuning avoids.

D $\mathsf{ARC}_{\text{score}}$ Details

Given a generated summary S and a set of expert-annotated salient argument roles $\mathcal{R}$, $\mathsf{ARC}_{\text{score}}$ measures the extent to which S covers the argumentative content in $\mathcal{R}$ (Elaraby and Litman, 2026). Each role $r \in \mathcal{R}$ is decomposed into atomic facts $\mathcal{F}_r$. An LLM-based verifier assigns each fact a coverage label:

$$\delta(f_i, S) = \begin{cases} 1 & \text{if } f_i \text{ is correctly covered in } S, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Role-level coverage is computed as the proportion of facts in r covered by S:

$$\mathsf{ARC}_{\text{role}}(r, S) = \frac{1}{|\mathcal{F}_r|} \sum_{f_i \in \mathcal{F}_r} \delta(f_i, S). \tag{3}$$

The final score averages role-level coverage across all annotated roles:

$$\mathsf{ARC}_{\text{score}}(S) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathsf{ARC}_{\text{role}}(r, S). \tag{4}$$

In addition to the binary coverage label, the verifier assigns an error type $e \in \{\text{MISSING ERROR}, \text{FACTUAL ERROR}\}$ for uncovered facts. *Missing Errors* (**ME**) indicate salient content omitted from the summary, while *Factual Errors* (**FE**) indicate contradictions or unsupported renderings of salient content. This allows us to distinguish whether improvements come primarily from better saliency coverage or reduced factual inconsistency.

$\mathsf{ARC}_{\text{score}}$ ranges from 0 to 1, with higher values indicating greater coverage of salient argument roles. Prior work shows that $\mathsf{ARC}_{\text{score}}$ achieves the strongest correlation with expert coverage judgments among lexical, semantic, entailment-based, and decomposition-based metrics such as ROUGE, BERTScore, SummaC, and FactScore.

E Scaling Training Context Length

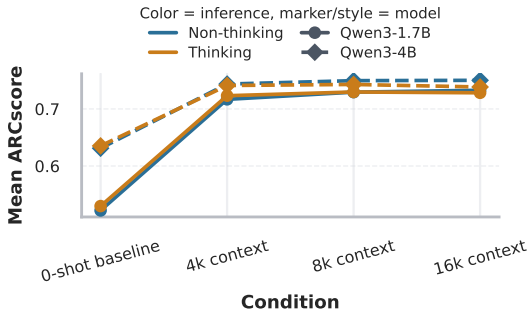

Figure 4: Context-length distillation across student models and inference strategies

We investigate whether exposing student models to longer training documents improves argument saliency coverage. Using the same 1,000 training documents, we resample inputs to include contexts up to 8k and 16k tokens while maintaining a balanced length distribution across groups using 2k-token bins. We evaluate summary-only distillation from `Qwen3-14B` on two representative student sizes: `Qwen3-1.7B` and `Qwen3-4B`. Figure 4 shows that the dominant improvement comes from distillation itself, with the jump from zero-shot to 4k training contexts substantially outweighing gains from further context scaling. Beyond 4k, longer training contexts provide only marginal additional improvements, particularly in non-thinking inference mode in the $4B$ student model. These findings

suggest that while longer contexts can modestly improve argument coverage, their effect is secondary to distillation itself, consistent with our broader data-efficiency results.

F Perplexity Analysis: Explaining Teacher Impact

To investigate why `Qwen3-14B` summaries yield stronger distillation signal than *GPT-5-mini* or expert-written summaries, we analyze the perplexity of each summary type under the base student models prior to fine-tuning. Since autoregressive language models are trained to minimize the negative log-likelihood of their training data (Brown et al., 2020), perplexity directly reflects how compatible a given text is with the student's learned distribution. Summaries with lower perplexity under the student are closer to its pretraining distribution, providing a more on-policy supervision signal (Agarwal et al., 2024) and reducing the distribution mismatch that standard off-policy fine-tuning must overcome (Ross et al., 2011).

Figure 5 shows that `Qwen3-14B` summaries consistently yield the lowest median perplexity across all student sizes (median: 9.7–19.3), followed by *GPT-5-mini* (median: 13.6–28.8), with expert-written summaries exhibiting substantially higher and more variable perplexity (median: 34.1–76.5, with a heavy tail extending to thousands). This ordering is consistent across all four student model sizes and directly mirrors the distillation performance ranking in Table 1: the teacher whose summaries are most compatible with the student's distribution produces the strongest coverage gains. Notably, as student model size increases, perplexity under `Qwen3-14B` summaries decreases (from 19.3 at 0.6B to 9.7 at 8B), suggesting that larger students within the same model family are increasingly well-aligned with the teacher's output distribution. This distributional alignment may partly explain why expert fine-tuning degrades most severely at 8B, where the gap between expert and teacher outputs is largest relative to the student's strong zero-shot prior.

We emphasize that this analysis is correlational and offered as a partial diagnostic of distributional alignment between teachers and students, not as a complete causal account of teacher impact. A rigorous explanation would additionally need to control for teacher quality on the summarization task itself and for surface-level features such as lexical and semantic similarity between teacher outputs and expert references. We leave a fuller analysis spanning multiple student–teacher pairs along these axes to future work.

G Relative Change in Factual and Missing Errors

Figure 6 reveals two patterns. Coverage gains are immediate: ME drops at every student size for both teachers, with the 1.7B student showing the largest relative reduction (-48% with Qwen3-14B). Factuality gains, in contrast, require a capacity threshold: FE *increases* at 0.6B for both teachers and only begins decreasing once the student is large enough to absorb the factuality signal – between 0.6B and 1.7B for Qwen3-14B, and between 1.7B and 4B for GPT-5-mini. Across every size, Qwen3-14B dominates GPT-5-mini on both metrics, with the FE gap widening as the student grows.

Table 6 shows the atomic facts decomposed from salient rhetorical roles and the per-fact verifier verdicts for summaries generated by the baseline and the two fine-tuned models. The $\mathtt{ARC}_{\text{score}}$ gains over the baseline come primarily from a reduction in Missing Errors (**ME**: $10 \rightarrow 5 \rightarrow 3$), while Factual Errors remain low and roughly constant (**FE**: $1 \rightarrow 2 \rightarrow 1$).

H Impact of LoRA Rank on Distillation Performance

To verify that the observed saturation in distillation performance is not an artifact of limited model capacity, we increase the LoRA rank beyond the default $r = 32$ to $r \in \{64, 128\}$, where $r = 128$ approaches the capacity of full-parameter tuning (Schulman and Lab, 2025). Figure 7 shows that increasing the rank yields no significant improvement ($p > 0.05$ in all cases), and in most settings performance remains flat or slightly degrades when thinking mode is enabled at inference. These results confirm that the saturation pattern reported in 5.2 reflects a genuine property of the distillation signal rather than a capacity bottleneck in the student model.

I Few-shot Distillation by Student Model

Figure 8 disaggregates few-shot distillation by student model size and inference mode. The main trend is consistent with Figure 2: most of the improvement occurs with only 10 training

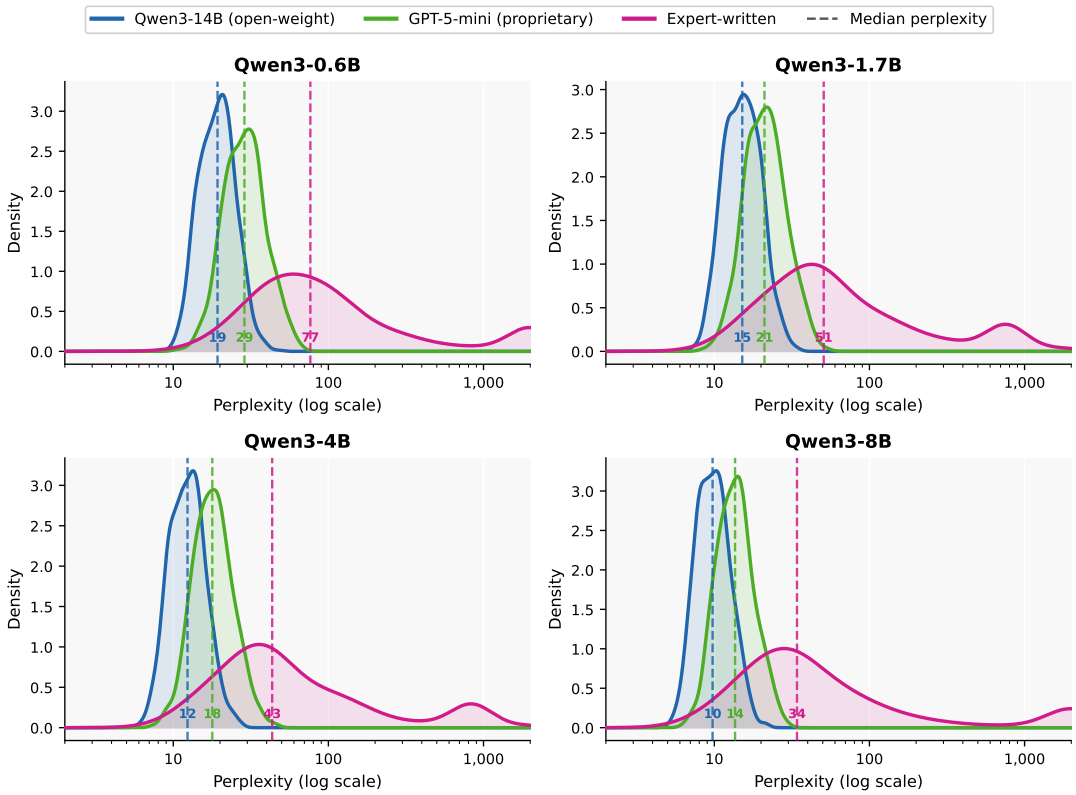

Figure 5: Perplexity distributions of teacher-generated and expert-written summaries under each base student model (before fine-tuning). Dashed lines indicate medians.

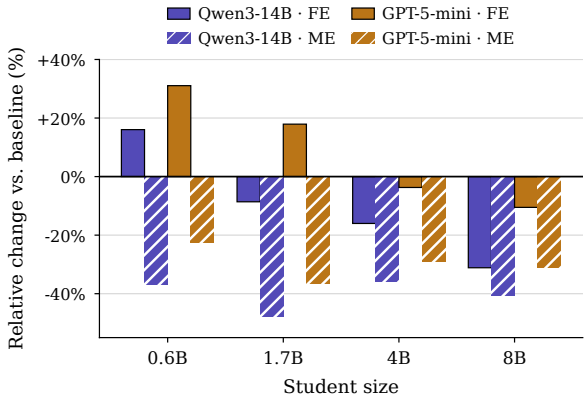

Figure 6: Relative change in Factual Errors (FE, solid) and Missing Errors (ME, hatched) under summary-only distillation relative to the zero-shot baseline, grouped by student size and teacher (■ Qwen3-14B, ■ GPT-5-mini). Results are shown under thinking-mode inference; negative values indicate fewer errors.

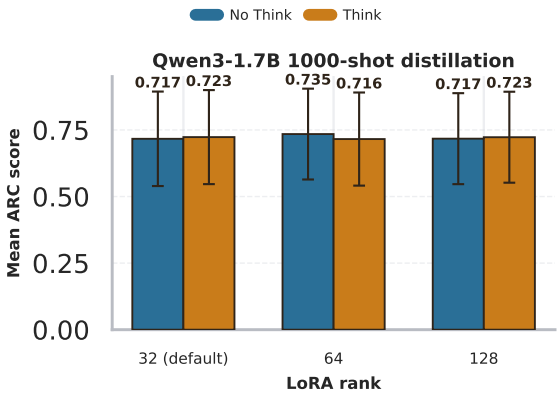

Figure 7: Impact of increasing LoRA rank on `Qwen3-1.7B` student.

documents, and scaling to 100 or 1,000 documents provides diminishing and sometimes inconsistent gains. This pattern holds across student sizes, suggesting that argument-saliency distillation is highly data-efficient rather than primarily data-scale driven. Moreover, reasoning-chain-only (**R**) and summary+reasoning-chain (**SR**) supervision closely follow summary-only distillation from `Qwen3-14B`, suggesting that reasoning traces do not provide a consistently stronger supervision signal than summaries alone. Finally, *GPT-5-mini* exhibits a distinct instability, with performance often decreasing at 100 documents before recovering at 1,000, although this effect is weaker for the `Qwen3-8B` student.

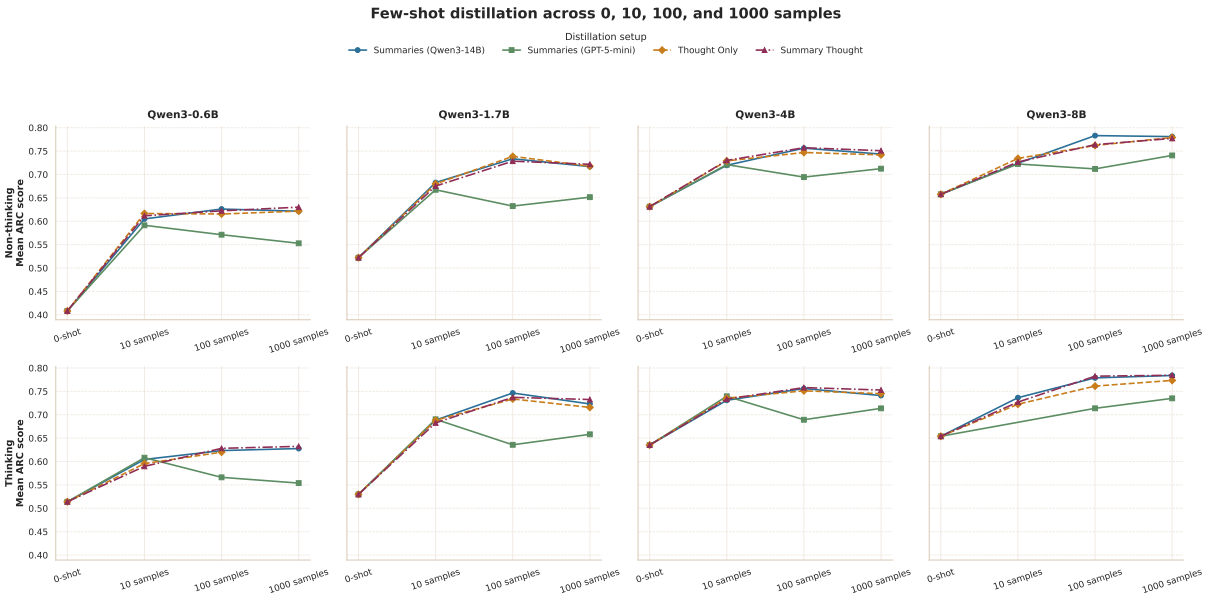

Figure 8: Few-shot distillation across models and inference strategies

Chain-of-Arguments (CoA) Prompt
`You will read a legal opinion and produce a coherent {length}-word summary` of its most salient legal arguments. DEFINITIONS A legal argument has three components: - Issue: a legal question the court addressed. - Reasoning: the court's basis for its conclusion — statutes cited, precedents applied, factual findings relied on. - Conclusion: how the court resolved that issue. PROCEDURE Step 1 — Extract. Identify every distinct argument in the opinion. Order them by how decisive they are to the outcome (most decisive first). Keep each component to a single line. Step 2 — Summarize in {length} words. The summary must: - cover every argument from Step 1, weighted by importance; - contain only information present in the opinion — do not introduce facts, parties, statutes, holdings, or inferences from outside the document; - name parties, statutes, and courts exactly as the document does. OPINION <opinion> {input} </opinion> OUTPUT FORMAT <arguments> 1. Issue: ... Reasoning: ... Conclusion: ... 2. Issue: ... Reasoning: ... Conclusion: ... </arguments> <summary> exactly {length} words </summary>

Table 5: Prompt used for the Chain-of-Arguments (CoA) baseline.

#	Atomic fact (rhetorical role)		(a)	(b)	(c)
1	Custodial arrangements for the parties' two youngest children were at issue.	*(Issue)*	–	–	–
2	The two youngest children are aged 15 and 13.	*(Issue)*	–	✗	–
3	The two youngest children currently lived with the father the majority of the time.	*(Issue)*	–	✗	–
4	Children are placed in the custody of the mother.	*(Conclusion)*	✓	✓	✓
5	Children are placed in the primary care of the mother.	*(Conclusion)*	✓	✓	✓
6	Until the separation, the mother had been the primary caretaker of the children.	*(Reason 1)*	–	✓	✓
7	The mother appeared to have a stronger emotional bond with the children.	*(Reason 1)*	–	✓	✓
8	The father had been physically, emotionally, and verbally abusive to the wife throughout the marriage.	*(Reason 1)*	–	–	✗
9	The father seemed incapable of putting the children's needs above his own.	*(Reason 1)*	–	✓	✓
10	The mother was much more likely than the father to encourage the children to have contact with the other parent.	*(Reason 1)*	–	✓	✓
11	The father was retired.	*(Reason 2)*	–	–	✓
12	The father might have more time to spend with the children.	*(Reason 2)*	✗	–	✓
13	Having more time to spend with the children did not outweigh all of the other factors.	*(Reason 2)*	–	–	✓
		✓ *Supported (of 13)*	2	6	**9**
		– *Missing (of 13)*	10	5	**3**
		✗ *Contradicted (of 13)*	1	2	1

(a) Qwen3-0.6B (baseline) *302 words*	**(b) + GPT-5-mini teacher** *321 words*	**(c) + Qwen-14B teacher** *327 words*
The case between Christie Chika Ezurike and Levi Chucks Ezurike, which resulted in a divorce, was brought to the Supreme Court of Nova Scotia (Family Division) in 2006. [...] The parents were unable to agree on custody arrangements or the division of assets or child support. [...] The marriage lasted over 20 years, and the Respondent worked in the federal government and maintained a taxi driver's license. [...] The court ruled that the Petitioner should have custody of the children and primary care[4,5], despite the Respondent's financial difficulties. The Petitioner emphasized her commitment to the children's well-being, while the Respondent was unable to provide adequate care. The case highlighted the importance of the children's best interests [...] and the need for the children to have a positive relationship with both parents.	This is a divorce judgment from the Nova Scotia Family Division, issued by Justice Leslie J. Dellapinna. Christie Ezurike, the Petitioner, sought divorce and custody of her two older children [...]; the Respondent, Levi Ezurike, opposed [her] application and sought custody of Nneka and access to Abuchi. [...] The Petitioner testified she is the primary caregiver to both children and that she has made effort to provide for them[5,6], though the Respondent contested that effort. [...] Christie also expressed concern about the Respondent's treatment of the children and his belief that they should not be left unsupervised for extended periods.[7,9,10] [...] The court concluded Christie has the better interests of the children and granted her divorce application, with custody of Abuchi and Nneka determined by the trial judge.[4]	In *Ezurike v. Ezurike*, the Supreme Court of Nova Scotia ruled on custody and access arrangements for four children of the divorcing couple. [...] The Respondent, now retired from federal service[11], resides in a reconstructed home with the children [...]. He has spent considerable time with the children[12] [...]. The Petitioner, a successful social worker, has focused on providing emotional and practical care for the children[6,7] [...]. [...] The Respondent's emotional detachment and lack of understanding of the children's challenges[9,10] [...] raised concerns. Ultimately, the court concluded that the Petitioner should have custody and primary care of the children[4,5], with the Respondent sharing reasonable access on alternate weekends[13]. [...]
ARCScore: 0.250	**ARCScore: 0.450**	**ARCScore: 0.700**

Table 6: **Legal opinion: *Ezurike v. Ezurike* (`d_2006nssc73.txt`).** Per-fact coverage (top) and generated summaries (bottom) for three systems: non-tuned Qwen3-0.6B (a), and the same model tuned with a GPT-5-mini teacher (b) or a Qwen-14B teacher (c). Each cell shows the verifier's verdict: ✓ supported, – missing (fact absent from the summary), ✗ contradicted (summary asserts something inconsistent with the fact). Highlighted spans were accepted by the verifier as supporting an atomic fact; the superscript indicates which fact (e.g. [4]). $\mathrm{ARC}_{\mathrm{score}}$ is the mean recall across the four rhetorical components; passages of the summaries not bearing on any atomic fact have been elided as [...].