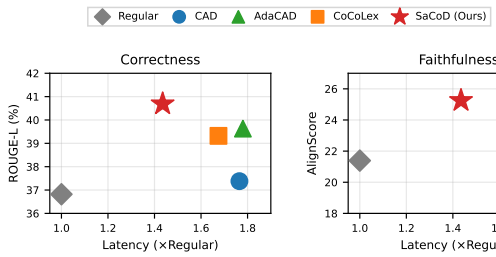
Copy Less, Ground More: Selective and Confidence-Aware Decoding for Grounded Legal Generation

Anonymous ACL submission

Abstract

While retrieval-augmented generation (RAG) grounds large language models (LLMs) in external knowledge, ensuring faithful evidence utilization in complex domains like law remains challenging. LLMs often ignore retrieved facts or indiscriminately copy from noisy contexts. Existing context-aware decoding methods attempt to mitigate this by amplifying the influence of retrieved content, but they treat the context as a monolithic block—amplifying irrelevant spans alongside evidence—and intervene at every decoding step, incurring substantial inference overhead. To address this, we propose SACoD (Selective and Confidence-aware Decoding), a training-free inference framework. SACoD leverages the LLM's internal attention signals to dynamically isolate salient evidence paragraphs at each step, restricting a customized copy mechanism strictly to relevant facts. Furthermore, an entropy-based gate activates this intervention only when the model is uncertain, bypassing copy operations for deterministic tokens to preserve linguistic fluency and cut computational overhead. Experiments on four legal question-answering benchmarks across three open-weight LLMs demonstrate that SACoD significantly improves faithfulness and response correctness over robust decoding baselines while operating substantially faster than prior context-aware methods.



Method	Evidence selection	Selective intervention	Extra passes	Blind-copy risk	Overhead
Vanilla	N/A	N/A	0	Low	Low
CAD	No	No	+1	Med	High
AdaCAD	No	No	+1	Med	High
CoCoLex	No	No	0	High	Med
SACoD	Self-guided	Yes (gate)	0	Low	Low

Figure 1: Comparison of context-aware decoding methods. SACoD achieves superior correctness (ROUGE-L) and faithfulness (AlignScore) while maintaining lower latency than CAD, AdaCAD, and CoCoLex. Results are averaged across four legal benchmarks on Mistral-7B. Latency is relative to Regular decoding (1.0×).

1 Introduction

Large language models are increasingly used to support high-impact decision workflows, including legal question answering (Italiani et al., 2025; Akarajaradwong et al., 2025), contract review (Wang et al., 2025a; Narendra et al., 2024), and compliance checking (Alecci et al., 2025; Hassani, 2024). In these settings, users need more than fluent text: outputs must be *auditable*, where each claim is traceable to an authoritative span in the underlying document (Chalkidis et al., 2022; Wiratunga et al., 2024; Hendrycks et al., 2021; Jain et al., 2025). Meeting this requirement is particularly challenging in the legal domain, where evidence is long, structurally heterogeneous, and often templated. As a result, even strong LLMs can produce plausible but unsupported statements, risking incorrect interpretation and eroding trust (T.y.s.s et al., 2025; Chitgopkar et al.).

Retrieval-Augmented Generation partially addresses this issue by conditioning generation on retrieved passages (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022). However, retrieval alone does not guarantee *faithful evidence use*. Under standard decoding, RAG models can still fail to selectively incorporate the retrieved context (Hagström et al., 2025), yielding outputs that drift from the supporting evidence—a particularly serious limitation in high-stakes legal applications (Magesh et al., 2025).

To address evidence misuse, a growing line of work enforces grounding at inference time via

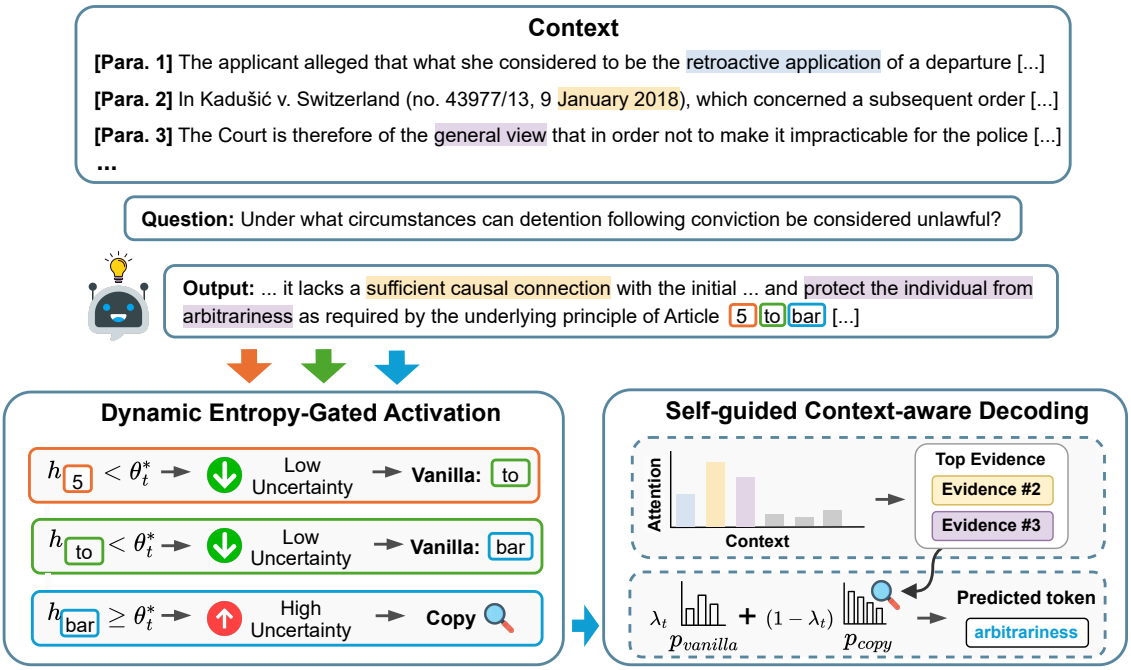


Figure 2: Overview of SACoD. The entropy-level gate bypasses context-aware decoding for confident tokens (e.g., "5", "to") and activates it for uncertain ones (e.g., "bar"). When activated, the self-guided module identifies salient evidence via attention and interpolates p_{vanilla} with p_{copy}.

context-aware decoding, which modifies the next-token distribution to prefer context-consistent tokens (Shi et al., 2024a; Wang et al., 2025b; T.y.s.s et al., 2025). Despite progress, existing methods face two practical limitations for long legal contexts (see the comparison table in Figure 1). **First**, none of them perform *evidence selection*: contrastive methods (CAD (Shi et al., 2024a), AdaCAD (Wang et al., 2025b)) amplify the influence of the entire retrieved context via an extra forward pass, which can over-weight irrelevant spans; copy-based methods (CoCoLex (T.y.s.s et al., 2025)) build a datastore from the full context, leading to even higher *blind-copy risk* as tokens from non-evidence paragraphs directly enter the generation distribution. **Second**, all prior methods apply *intervention at every decoding step*—contrastive approaches further require an additional forward pass per token (+1 *extra pass*), compounding the overhead, while CoCoLex interpolates at every step without selectivity, adding unnecessary computation even when the model is already confident.

We propose SACoD (Selective and Confidence-aware Decoding), a training-free decoding framework designed to improve faithfulness while keeping inference overhead low (Figure 2). SACoD is guided by two observations: (i) in long legal documents, only a small subset of spans is relevant for the current decision, so enforcing adherence to the entire retrieved context is unnecessary and can be harmful; and (ii) the need for contextual guidance varies across tokens—many tokens are nearly deterministic (boilerplate and syntactic connectors), while ambiguous tokens are where hallucinations and evidence misuse concentrate.

Based on these insights, SACoD makes two lightweight decisions during generation. First, it performs *self-guided evidence identification* by using the model's internal attention signals to dynamically focus on high-salience evidence spans rather than the full retrieved text, reducing blind-copy risk. Second, it employs *entropy-gated selective intervention*, which activates context-aware interpolation only when the model is uncertain, and otherwise falls back to vanilla decoding—requiring zero extra forward passes and keeping overhead low. This design targets evidence neglect while reducing blind copying and avoiding per-token overhead when guidance is unnecessary. In the experiments, we evaluate SACoD on four legal grounded-generation benchmarks using multiple open-weight LLMs. Across settings, SACoD improves contextual faithfulness and answer quality over prior context-aware decoding strategies, while

maintaining inference efficiency close to vanilla decoding (Figure 1). In summary, our contributions are as follows:

- We characterize key failure modes of existing context-aware decoding methods in legal RAG, including evidence over-amplification (leading to blind copying) and excessive inference time overhead.
- We introduce SACoD, a training-free decoding method that combines self-guided evidence identification with entropy-gated, adaptive context-aware interpolation.
- We demonstrate empirically that SACoD improves faithfulness and correctness while remaining computationally efficient across diverse legal benchmarks and LLM backbones.

2 Preliminary

Problem Setup. We consider grounded text generation in a retrieval-augmented setting. Given an input query x and a retrieved context c, a language model with parameters Θ generates a response $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ of length n. Generation proceeds autoregressively: at each step t, the model samples the next token from the conditional distribution $y_t \sim p_\Theta(y_t \mid c, x, y_{<t})$.

Notation. We adopt standard Transformer notation (Vaswani et al., 2017). Let L denote the number of layers and H the number of attention heads per layer. At decoding step t, the attention distribution at layer l and head h is denoted $\mathbf{a}_t^{(l,h)} \in \mathbb{R}^{|c|+|x|+t-1}$, representing how the model attends over all preceding tokens. We define the head-averaged attention at layer l as

$$\mathbf{a}_t^{(l)} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{a}_t^{(l,h)}. \tag{1}$$

Intuitively, $\mathbf{a}_t^{(l)}$ captures which input positions layer l attends to when predicting token y_t.

3 SaCoD

We present SACoD, a training-free decoding framework for grounded legal generation that improves evidence utilization without incurring per-token computational overhead. As illustrated in Figure 2, SACoD integrates two complementary components: (i) *Self-guided Context-aware Decoding*, which leverages the model's internal attention signals to identify and focus on a compact set of salient evidence paragraphs within the retrieved context, and (ii) an *Dynamic Entropy-gated Activation* module, which triggers context-aware interpolation only when the model exhibits high uncertainty and otherwise defaults to standard decoding. We describe each component in detail below.

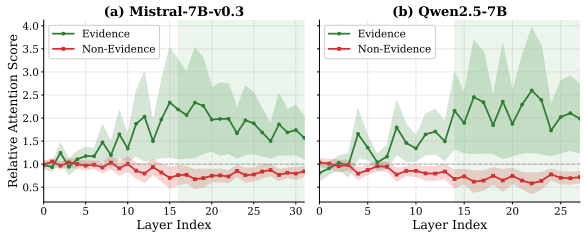


Figure 3: Attention-Evidence Alignment Across Transformer Layers. Relative attention scores for evidence paragraphs (green) vs. non-evidence paragraphs (red) across layers for (a) Mistral-7B-v0.3 and (b) Qwen2.5-7B. Shaded regions indicate standard deviation. Deeper layers consistently assign higher attention to human-annotated evidence paragraphs.

3.1 Self-guided Context-aware Decoding

Motivating Observations. To ground our framework empirically, we construct an expert-annotated dataset of 80 legal RAG samples (20 per benchmark), with two annotators achieving high inter-annotator agreement (Cohen's Kappa of 0.8720). For an input sequence with context c, we divide it into m paragraphs $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ and track which spans actively support the generated outputs. Two critical insights guide our method:

Observation 1: Legal evidence is highly sparse. Human evaluation confirms that a mere 8–36% of the retrieved text serves as operative evidence (Figure 4). Consequently, uniform amplification of the full context injects noise and drives blind copying of irrelevant statutes. This makes extracting a compact relevance subset $\mathcal{P}_E \subset \mathcal{P}$ computationally and qualitatively vital.

Observation 2: Attention correctly proxies evidence salience. To isolate $\mathcal{P}_E$ efficiently, we measure the head-averaged attention $\bar{a}_{t,i}^{(l)}$ paid to paragraph $\mathbf{p}_i$ by layer l at decoding step t. Computing the step-aggregated score $e_i^{(l)} = \frac{1}{n}\sum_{t=1}^{n} \bar{a}_{t,i}^{(l)}$ shows that deeper Transformer layers naturally concentrate mass on human-verified evidence paragraphs (Figure 3). Thus, intrinsic self-attention mechanisms offer a training-free proxy for adaptive evidence selection without requiring additional

forward passes or auxiliary models. Details of observations are provided in Appendix C.1.

Self-Guided Evidence Selection. Based on these observations, SACoD dynamically identifies evidence paragraphs by aggregating attention over the deepest 50% of layers, denoted $\mathcal{L}_{\text{top}}$ (ablation in Appendix C.5). At decoding step t, the integrated evidence score for paragraph $\mathbf{p}_i$ is $e_{t,i} = \frac{1}{|\mathcal{L}_{\text{top}}|}\sum_{l\in\mathcal{L}_{\text{top}}} \bar{a}_{t,i}^{(l)}$. Controlled by a precision-recall tolerance parameter $\alpha \in [0,1]$, we filter for the active supporting paragraphs:

$$\mathcal{P}_E^{(t)} = \left\{ \mathbf{p}_i \in \mathcal{P} \mid e_{t,i} \geq \alpha \cdot \max_j(e_{t,j}) \right\}. \quad (2)$$

Crucially, this selection is executed *only* when the entropy gate activates (Section D.1), bypassing overhead at computationally deterministic steps.

Adaptive Copy-Based Decoding. To faithfully reproduce factual entities without rote copying, SACoD refines confidence-guided copy decoding (T.y.s.s et al., 2025) by restricting its reference datastore exclusively to the dynamically isolated evidence tokens $c_E^{(t)} \in \mathcal{P}_E^{(t)}$. At each intervention step t, the pre-computed hidden states $\mathbf{k}_i$ for evidence tokens $w_i \in c_E^{(t)}$ form a step-specific datastore $\mathcal{D}_E^{(t)}$. Token-wise similarities $s_t(i) = \exp(-\|\mathbf{h}_t - \mathbf{k}_i\|_2)$ map the current hidden state $\mathbf{h}_t$ to a localized vocabulary copy distribution:

$$p_{\text{copy}}(v \mid \mathcal{D}_E^{(t)}) \propto \sum_{\{i \mid w_i = v\}} s_t(i) \quad (3)$$

This targeted copy distribution is then softly interpolated with the base parametric prior p_Θ, modulated by a step-specific confidence weight $\lambda_t \in [0,1]$:

$$p(y_t \mid x, c_E^{(t)}, y_{<t}) = \lambda_t\, p_\Theta(y_t) + (1-\lambda_t) p_{\text{copy}}(y_t \mid \mathcal{D}_E^{(t)}). \quad (4)$$

As detailed next, when the model is confident, λ_t prioritizes p_Θ; when uncertain, mass shifts to the selective copy mechanism. By isolating $\mathcal{D}_E^{(t)}$, this interpolation ensures precise evidence recall while actively guarding against noise from unrelated boilerplate.

3.2 Dynamic Entropy-Gated Activation

A key challenge in context-aware decoding is deciding *when* contextual intervention is actually necessary. During generation, tokens fall into two broad categories: deterministically predictable words (e.g., function words, boilerplate) and highly uncertain tokens where external evidence is required (e.g., named entities, statutory references).

Existing methods typically intervene at *every* decoding step. This is both inefficient and counterproductive, as injecting contextual bias at deterministic steps amplifies irrelevant text and degrades fluency. This motivates a lightweight gate that activates contextual guidance only when the model's confidence wanes.

At decoding step t, we compute the normalized entropy $h_t \in [0,1]$ of the next-token distribution to quantify uncertainty (Eq. 8). This metric directly maps to the soft interpolation weight $\lambda_t = \exp(-h_t)$, dictating how aggressively the context-aware copy mechanism is applied. When confident (low h_t), $\lambda_t \approx 1$ and the base generation proceeds normally; when uncertain, λ_t dampens, heavily weighting the evidence-filtered copy distribution.

To strictly avoid unnecessary computation, we introduce a dynamic hard-gate whose threshold θ^* is automatically determined via Otsu's method (Eq. 10). Before generating tokens, Otsu's thresholding partitions the prefill entropy buffer $\mathcal{H}_0 = \{h_i^{\text{pfx}}\}$ to separate historically confident tokens from uncertain ones without manual tuning.

During autoregression, the gate evaluates the step entropy h_t:

$$\texttt{Gate}(t) = \begin{cases} \texttt{Vanilla}, & h_t < \theta^*, \\ \texttt{Copy}, & h_t \geq \theta^*, \end{cases} \quad (5)$$

where `Vanilla` bypasses intervention, and `Copy` triggers the entropy-weighted interpolation. As new entropy values accrue, θ^* is periodically refreshed over the expanding buffer every $K = 10$ steps, keeping the boundary calibrated. This two-level design—adaptive hard gating paired with confidence-driven soft interpolation—concentrates computation purely on the uncertain steps where evidence grounding is vital. The complete SACoD algorithmic formulation is provided in Appendix D.

4 Experiments

We evaluate the effectiveness and generality of the proposed SACoD method across diverse legal text generation benchmarks and multiple state-of-the-art model architectures. Our evaluation is designed to answer two primary research questions: (i) whether SACoD improves generation faithfulness and correctness compared to existing context-

Method	CUAD		OALQA		ObliQA		AQuAECHR		Average	
	RL	BERT-P	RL	BERT-P	RL	BERT-P	RL	BERT-P	RL	BERT-P
Mistral-7B-Instruct-v0.3										
Regular	32.83	83.53	45.95	89.48	42.20	86.98	26.30	84.83	36.82	86.21
CAD	33.59	84.98	47.06	89.83	42.85	87.45	26.02	85.72	37.38	87.00
AdaCAD	33.90	85.60	49.00	90.65	44.55	89.24	27.05	85.88	38.63	87.84
CoCoLex	34.13	86.09	49.27	90.13	45.72	89.12	29.15	86.30	39.57	87.91
SACoD	**35.16***	**87.89***	**50.95***	**91.32***	**46.84***	**89.45***	**29.80***	**86.81***	**40.69**	**88.87**
Qwen2.5-7B-Instruct										
Regular	32.09	83.85	42.73	87.49	42.89	86.78	25.79	81.19	35.88	84.83
CAD	33.54	84.35	43.00	87.89	40.98	87.01	24.04	80.62	35.39	84.97
AdaCAD	33.85	84.56	44.58	88.59	43.38	87.82	25.26	80.77	36.77	85.44
CoCoLex	35.45	86.00	45.51	89.62	44.95	89.70	28.54	86.95	38.61	88.07
SACoD	**36.43***	**87.44***	**48.47***	**90.61***	**45.32***	**90.56***	**29.24**	**87.33***	**39.87**	**88.99**
Meta-Llama-3-8B-Instruct										
Regular	34.63	84.97	42.30	86.43	43.53	86.68	26.23	84.14	36.67	85.56
CAD	33.41	83.43	41.91	85.64	40.24	86.61	26.96	84.56	35.63	85.06
AdaCAD	35.08	85.28	43.57	87.30	45.44	88.03	27.31	85.45	37.85	86.52
CoCoLex	35.32	86.45	45.76	88.29	45.57	88.54	29.12	86.80	38.94	87.52
SACoD	**36.62***	**87.34***	**47.73**	**89.57***	**46.75***	**90.96***	**29.54**	**87.08***	**40.16**	**88.74**

Table 1: Response correctness comparison of different decoding-based methods across four legal text generation datasets and three LLMs. Entries marked with * are statistically significantly higher than the second-best performing baseline at the 95% confidence level, according to the Wilcoxon signed-rank test.

aware decoding strategies, and (ii) whether it can accelerate inference relative to prior context-aware decoding methods.

4.1 Metrics

Following established protocols for grounded generation, we evaluate performance across three dimensions. To assess **Correctness**, we measure the overlap between the generated response and the reference answer using ROUGE-L for lexical overlap and BERTScore-Precision (Zhang* et al., 2020) for semantic alignment. We quantify **Faithfulness** by computing the AlignScore (Zha et al., 2023) and Factual Consistency of UniEval (Zhong et al., 2022) between the generated response and the provided input context, ensuring the generation is grounded in the source text. We also evaluate linguistic properties such as **Fluency** using UniEval (Zhong et al., 2022), which assesses the quality of individual sentences and whether generated text forms a cohesive narrative.

4.2 Benchmarks and Baselines

To ensure a comprehensive evaluation of generation abilities, we utilize four challenging legal benchmarks. We employ **CUAD** (Hendrycks et al., 2021) to test the model's ability to locate specific clauses within dense legal contracts. We also utilize **AQuAECHR** (Weidinger et al., 2025), a dataset derived from European Court of Human Rights judgments, which requires the model to synthesize answers based on long case law documents. Next, we employ **OALQA** (Butler, 2023), a question answering dataset built on the Open Australian Legal Corpus, designed to evaluate hallucination mitigation by requiring models to produce concise answers grounded in reference cases and judicial decisions. Finally, we include **ObliQA** (Gokhan et al., 2024), which focuses on regulatory QA over Abu Dhabi Global Market financial regulations, requiring extractive answers grounded in the full document corpus.

We compare SACoD against four distinct baselines using three representative open-weights LLMs: Mistral-7B-Instruct-v0.3 (Mis-7B) (Jiang et al., 2023a), Qwen-2.5-7B (Q2.5-7B) (Ahmed et al., 2025), and Llama-3-8B-Instruct (L3-8B) (Dubey et al., 2024). The baselines include **Vanilla Decoding** using standard greedy decoding, and **CAD** (Context-Aware Decoding) (Shi et al., 2024a), which amplifies output probabilities conditioned on the context by contrasting them with a context-free output. We also compare against **AdaCAD** (Wang et al., 2025b), an adaptive version of CAD that dynamically adjusts the contrastive weight based on the divergence between contextual and parametric distributions. Finally, we benchmark against **CoCoLex** (T.y.s.s et al., 2025), a re-

Method	CUAD			OALQA			ObliQA			AQuAECHR			Average		
	AS	FC	Flu	AS	FC	Flu	AS	FC	Flu	AS	FC	Flu	AS	FC	Flu
Mistral-7B-Instruct-v0.3															
Regular	59.51	63.83	82.12	66.93	74.24	84.43	84.02	64.23	**77.25***	73.45	79.27	82.76	70.98	70.39	81.89
CAD	61.35	64.60	79.22	68.46	76.55	85.38	84.63	64.01	76.83	74.02	79.77	79.07	72.12	71.23	80.13
AdaCAD	60.87	63.54	80.49	67.78	75.13	86.04	83.73	64.27	76.39	75.88	82.93	80.75	72.07	71.47	80.92
CoCoLex	62.10	65.69	80.54	69.68	76.77	84.81	85.44	66.93	76.51	76.31	81.44	81.48	73.38	72.71	80.84
SACoD	**64.35***	**67.84***	**82.81**	**70.88***	**77.42***	**85.94***	**87.83***	**68.12***	76.35	**77.09***	**83.90***	**83.19***	**75.04**	**74.32**	**82.32**
Qwen2.5-7B-Instruct															
Regular	59.15	63.18	85.56	68.59	80.06	**87.72***	88.90	67.62	78.12	74.09	81.42	**81.18***	72.68	73.07	**83.15**
CAD	60.95	65.77	81.73	68.16	79.37	79.87	86.69	65.35	71.92	74.90	84.23	77.25	72.68	73.68	77.69
AdaCAD	59.25	64.53	83.95	67.80	78.56	83.01	89.39	66.05	72.65	78.55	84.62	77.86	73.75	73.44	79.37
CoCoLex	62.37	66.79	84.58	71.59	81.37	85.39	90.46	68.37	78.56	78.45	84.66	80.25	75.72	75.30	82.20
SACoD	**64.37***	**68.48***	**85.63**	**72.33***	**83.61***	86.87	**92.09***	**70.09***	**78.84**	**80.01***	**85.25***	80.77	**77.20**	**76.86**	83.03
Meta-Llama-3-8B-Instruct															
Regular	62.17	66.71	84.12	67.56	78.56	**87.92**	88.02	68.03	76.36	76.48	74.31	**78.49***	73.56	71.90	81.72
CAD	63.63	67.26	80.93	67.11	78.69	84.52	84.32	64.61	71.11	76.15	75.95	72.18	72.80	71.63	77.19
AdaCAD	63.39	67.73	82.77	67.53	78.82	85.40	88.09	68.37	76.20	76.73	75.57	79.80	73.94	72.62	81.04
CoCoLex	66.43	66.30	83.66	69.01	80.14	86.05	90.27	70.70	78.29	78.90	79.81	76.88	76.15	74.24	81.22
SACoD	**67.07***	**69.28***	**84.72***	**72.83***	**81.47***	87.05	**91.61***	**72.28***	**78.53**	**79.71***	**81.24***	77.02	**77.81**	**76.07**	**81.83**

Table 2: Contextual faithfulness comparison of different decoding-based methods across four legal text generation datasets and three LLMs. AS, FC, and Flu indicate AlignScore, Factual Consistency, and Fluency. Entries marked with * are statistically significantly higher than the second-best performing baseline at the 95% confidence level, according to the Wilcoxon signed-rank test.

cent method specifically designed for legal text that interpolates the model's logits with a copy-based distribution derived from the hidden states of the retrieved context. Implementation details are provided in Appendix B.

4.3 Implementation Details

For the baselines, we adhere to the optimal hyperparameters reported in their respective literature. For CAD, we set the static weight $\alpha = 0.5$. For AdaCAD, we clamp the adaptive weight $\alpha_t \in [0.3, 1.0]$. For CoCoLex, we use the last-layer hidden states for similarity computation with a smoothing factor of $\lambda = 0.5$. For our proposed SACoD, we set the attention layer $\mathcal{L}_{\text{top}}$ to the last 50% of transformer layers, the attention evidence threshold to $\alpha = 0.7$, and the Otsu threshold update interval to $K = 10$ for all experiments.

5 Results and Analysis

5.1 Main Results

We report response quality (ROUGE-L and BERT-Precision) in Table 1 and contextual faithfulness (AlignScore, Factual Consistency, and Fluency) in Table 2 across four legal QA benchmarks and three LLMs. Overall, the two families of contrastive baselines exhibit importantly different cross-model behaviors. CAD is the more brittle method: while it yields marginal correctness gains on Mistral-7B and Qwen2.5-7B, it degrades both response quality and faithfulness on Llama-3-8B, and consistently impairs fluency across all three models, confirming that always-on, uniformly weighted context amplification destabilizes generation. AdaCAD's adaptive weighting makes it a meaningfully stronger baseline—it improves over Regular on all three LLMs, including Llama-3-8B, yet it still consistently underperforms SACoD, particularly on long-form extractive tasks such as OALQA and ObliQA.

CoCoLex introduces confidence-guided copying and generally serves as a strong baseline, consistently outperforming contrastive methods. However, it still falls short of SACoD's performance. Because CoCoLex interpolates output distributions using the entire retrieved context, it is susceptible to incorporating irrelevant or noisy text passages, which limits its ceiling for factual fidelity and correctness.

Our proposed SACoD consistently outperforms prior approaches, improving average contextual fidelity while achieving the highest response quality across all settings. The correctness improvements are statistically significant (Wilcoxon signed-rank test (Wilcoxon, 1945), $p < 0.05$) over the second-best baseline, as indicated by the * markers in Table 1. Importantly, these gains co-occur with improved contextual faithfulness without incurring the fluency regressions observed in contrastive decoding methods. The improvements are particularly pronounced on long-form extractive tasks such as OALQA and ObliQA. By narrowing the copy mechanism exclusively to self-guided, attention-identified relevant spans, SACoD reduces the blind-copy behavior and noise injection of prior

approaches. Concurrently, the dynamically optimized entropy gate prevents unnecessary context-forcing on confident steps, ensuring the generative prior is preserved when the model is certain. Overall, our approach synergistically enhances faithfulness and correctness without compromising fluency, delivering robust improvements across a diverse set of tasks and models.

Human Evaluation. We conduct an expert human evaluation on 60 generated responses sampled from the AQuAECHR dataset. Two legal experts assessed each response on a 5-point Likert scale for correctness, faithfulness, fluency, and coherence. As shown in Table 3, SACoD consistently achieved the highest ratings across all four criteria, confirming its effectiveness in producing highly faithful and accurate legal text compared to the baselines. Detailed human evaluations are provided in Appendix C.2, and qualitative case studies are available in Appendix E.

	Corr	Faith	Flu	Coh
Regular	$3.50_{\pm 0.52}$	$4.35_{\pm 0.38}$	$4.80_{\pm 0.14}$	$4.45_{\pm 0.32}$
CoCoLex	$3.90_{\pm 0.50}$	$4.40_{\pm 0.33}$	$4.75_{\pm 0.23}$	$4.65_{\pm 0.30}$
SACoD	$\mathbf{4.15}_{\pm \mathbf{0.35}}$	$\mathbf{4.60}_{\pm \mathbf{0.28}}$	$\mathbf{4.93}_{\pm \mathbf{0.05}}$	$\mathbf{4.75}_{\pm \mathbf{0.31}}$

Table 3: Human evaluation results on 60 randomly sampled questions from the AQuAECHR dataset. Corr, Faith, Flu, and Coh indicate Correctness, Faithfulness, Fluency, and Coherence, respectively.

5.2 Efficiency Measurement

We empirically evaluate the average decoding latency of SACoD against the baselines on 200 samples from the ObliQA and AQuAECHR datasets. Our findings show that contrastive decoding methods (CAD and AdaCAD) nearly double inference time, as they require two forward passes per generated token, corresponding to the prior and context distributions. CoCoLex, although it does not require an additional forward pass, still incurs substantial overhead latency due to interpolating token probabilities at every timestep. In contrast, SACoD with contextual guidance from the entropy-level gate activates context-aware interpolation only when necessary, resulting in improved efficiency compared to prior context-aware decoding approaches.

Method	# Forwards	Average Latency	
		ObliQA	AQuAECHR
Regular	1	1.00×	1.00×
CAD	2	1.82×	1.71×
AdaCAD	2	1.84×	1.72×
CoCoLex	1	1.73×	1.62×
SACoD	1	**1.48**×	**1.39**×

Table 4: Inference time comparison of different methods on Mistral-7B, scaled to regular decoding.

Dataset	Model		
	Mistral	Qwen	Llama
Metric: AUROC			
CUAD	75.78	84.21	77.98
OALQA	69.76	78.67	77.19
ObliQA	84.63	91.22	89.09
AQuAECHR	67.25	73.57	72.12
Metric: NDCG			
CUAD	72.81	79.22	73.05
OALQA	64.87	72.00	70.52
ObliQA	71.88	76.05	76.10
AQuAECHR	86.20	87.70	87.41

Table 5: Evidence selection quality on our legal benchmarks and backbone models.

5.3 Validating Self-Guided Evidence Selection

To verify that SACoD faithfully utilizes relevant context, we evaluate whether its attention-based evidence scores favor ground-truth spans (Section 3.1). Treating continuous importance scores as a soft-ranking task, we evaluate via AUROC and NDCG@all. As shown in Table 5, SACoD accurately localizes supporting evidence across diverse models and datasets, typically scoring between 70–91. Performance is lower on OALQA due to the multi-hop narrative complexity of its legal queries. Because step-wise self-attention concentrates strictly on the immediate reasoning path for the current token, other candidate evidence is temporarily down-weighted. Despite this step-wise sparsity, dynamic localization effectively guides the full generative trajectory without overwhelming the model with the entire document.

5.4 Ablation Study

We assess the individual contribution of core components to the overall performance on ObliQA and AQuAECHR using Mistral-7B, reported in Table 6. **Entropy-gated activation** prevents unnecessary context-aware intervention at low-uncertainty steps. Removing the entropy gate (*w/o Dynamic Entropy Gate*) notably degrades fluency, as the copy mechanism is applied indiscriminately at every step, injecting contextual bias even when the model is

Model Variant	ObliQA		
	RL	AS	Flu
SACoD	**46.84**	**87.82**	76.35
w/o Dynamic Entropy Gate	44.17	87.09	76.31
w/o Adaptive Evidence	38.12	82.82	71.38
w/o Self-guided Evidence	42.67	86.32	76.51
Regular	42.20	84.02	**77.25**
	AQuAECHR		
	RL	AS	Flu
SACoD	**29.80**	77.09	**83.19**
w/o Dynamic Entropy Gate	29.00	**78.71**	76.58
w/o Adaptive Evidence	27.02	74.01	76.41
w/o Self-guided Evidence	26.86	76.65	75.93
Regular	26.30	73.45	82.76

Table 6: Performance comparison under different ablations of SACoD decoding method. Each variant removes or alters a core component to assess its individual contribution to the performance.

already confident. **Dynamic evidence selection** is the most critical component. The variant *w/o Adaptive Evidence* identifies evidence using attention signals from only the first decoding step and fixes that set for all subsequent steps. This static strategy exhibits the largest degradation across all metrics—particularly on long-form answers, where different parts naturally draw on different evidence paragraphs. **Self-guided evidence identification** via attention signals provides meaningful gains over copying from the full context. The *w/o Self-guided Evidence* variant constructs the copy datastore from the entire retrieved context, introducing noise from irrelevant paragraphs that dilutes the signal from genuinely supporting spans.

Other Analyses. The appendix provides further experiments and details: sensitivity of α over $[0.1, 1.0]$ (Appendix C.3), entropy gate activation rate (Appendix C.4), layer-selection ablation (Appendix C.5), context-size scaling (Table 10), generalization evaluation on summarization and medical domains (Table 11), a latency and memory profile (Appendix C.8), and a selection granularity comparison (Appendix C.9).

6 Related Work

LLMs in the Legal Domain. LLMs are increasingly integrated into legal workflows for tasks such as contract review and legal research (T.y.s.s et al., 2025). However, their deployment introduces significant reliability challenges (Avery et al., 2023; Harasta et al., 2024; Chitgopkar et al.). While Retrieval-Augmented Generation (RAG) helps reduce hallucinations (Dahl et al., 2024), unfaithful and unsupported outputs remain a critical issue (Magesh et al., 2025). Recent approaches like CoCoLex (T.y.s.s et al., 2025) have introduced confidence-guided, copy-based decoding to improve alignment with the context. Building on this, our work proposes a more fine-grained strategy that weights token contributions based on the model's attention to salient evidence, improving faithfulness to the most relevant context while mitigating indiscriminate copying.

Retrieval-Augmented Generation grounds language models in external knowledge, incorporating retrieved text at the input, intermediate, or output layers (Guu et al., 2020; Ram et al., 2023; Wu et al., 2022; Khandelwal et al., 2019). Existing pipelines range from training-based retriever-generator optimizations (Karpukhin et al., 2020; Izacard et al., 2023) to training-free techniques that augment prompt context at inference time (Jiang et al., 2023b). Because simple context augmentation can still yield inconsistencies, prior works have explored contrastive decoding (Shi et al., 2024b; Li et al., 2023) and copy-based decoding (T.y.s.s et al., 2025) to steer generation. However, these methods often either lack explicit enforcement of source adherence or mistakenly copy from irrelevant segments of the context. To address these limitations, our method leverages a selective decoding mechanism that dynamically focuses only on the most relevant contextual information.

7 Conclusion

We presented SACoD, an efficient, training-free decoding framework that enhances the factual faithfulness of retrieval-augmented generation in the legal domain. Rather than intervening at every step and uniformly copying from the entire retrieved context, SACoD selectively applies a localized copy mechanism only when the model is uncertain, strictly confining it to attention-identified relevant paragraphs. Extensive experiments confirm that this dual strategy not only yields state-of-the-art correctness and fluency but also substantially reduces decoding latency compared to prior context-aware methods. Future work will explore adapting SACoD to ultra-long context models (e.g., 100K+ tokens) via sparse attention or hierarchical evidence selection to maintain these efficiency gains over massive document corpora.

Limitations

While SACoD significantly improves faithfulness and correctness in grounded legal text generation, several limitations remain. First, in real-world applications, the ultimate quality and accuracy of the generated text inherently depend on the quality of the initial retrieval. Addressing this retrieval bottleneck remains a primary challenge, as even the most advanced context-aware decoding cannot compensate for missing or fundamentally incorrect source evidence. Second, our framework requires access to the model's internal signals, specifically attention weights and next-token probability distributions, to guide the decoding process. Consequently, SACoD cannot currently be assessed on proprietary LLMs (e.g., GPT-4) where these internal states are inaccessible via standard APIs. Finally, while our experiments demonstrate that the method is generally robust to different values for the attention threshold α, these parameters can still be fine-tuned on specific tasks to achieve optimal performance. An interesting future research direction would be to develop mechanisms that dynamically select the optimal thresholds for each decoding step or domain, further reducing the need for manual parameter tuning.

Ethics Statement

All datasets utilized in this study are publicly available and were used in strict compliance with their respective licenses and data usage agreements. We acknowledge that datasets containing case law judgments, such as OALQA and AQuAECHR, consist of public legal records that are inherently unanonymized. Our work processes these documents exactly as published for academic benchmarking and does not attempt to profile, track, or cross-reference individuals beyond the scope of the public record.

Furthermore, while SACoD demonstrably improves the faithfulness and grounding of language model outputs in legal contexts, we explicitly caution against fully automated deployment in high-stakes domains. This method is designed to serve as a decision-support tool for legal practitioners by enhancing information traceability, and it must not replace qualified legal expertise in consequential or binding decision-making processes.

References

Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Naseef Ur Rahman Chowdhury, and Ahshanul Haque. 2025. Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potentioal to surpass all competitors. *Authorea Preprints*.

Pawitsapak Akarajaradwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, Keerakiat Pratai, and Sarana Nutanong. 2025. Nitibench: Benchmarking llm frameworks on thai legal question answering capabilities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34292–34315.

Marco Alecci, Nicolas Sannier, Marcello Ceci, Sallam Abualhaija, Jordan Samhi, Domenico Bianculli, Tegawendé Bissyandé, and Jacques Klein. 2025. Toward llm-driven gdpr compliance checking for android apps. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, pages 606–610.

Joseph J Avery, Patricia Sanchez Abril, and Alissa del Riego. 2023. Chatgpt, esq.: Recasting unauthorized practice of law in the era of generative ai. *Yale JL & Tech.*, 26:64.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Umar Butler. 2023. Open australian legal qa.

Adam Byerly and Daniel Khashabi. 2024. How effective is self-consistency for long-context problems? *arXiv e-prints*, pages arXiv–2411.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Sudhan Chitgopkar, Noah Dohrmann, Stephanie Monson, Jimmy Mendez, Finale Doshi-Velez, and Weiwei Pan. Accuracy isn't everything: Understanding the desiderata of ai tools in legal-financial settings. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Michiel De Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, and William Cohen. 2021. Mention memory: incorporating textual knowledge into transformers through entity mention attention. *arXiv preprint arXiv:2110.06176*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. *arXiv preprint arXiv:2004.07202*.

Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2024. Rirag: Regulatory information retrieval and answer generation. *arXiv preprint arXiv:2409.05677*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2025. A reality check on context utilisation for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19691–19730.

Jakub Harasta, Tereza Novotná, and Jaromir Savelka. 2024. It cannot be right if it was written by ai: on lawyers' preferences of documents perceived as authored by an llm vs a human. *arXiv preprint arXiv:2407.06798*.

Shabnam Hassani. 2024. Enhancing legal compliance and regulation analysis with large language models. In *2024 IEEE 32nd International Requirements Engineering Conference (RE)*, pages 507–511. IEEE.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated NLP dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Paolo Italiani, Gianluca Moro, and Luca Ragazzi. 2025. Enhancing legal question answering with data generation and knowledge distillation from large language models: P. italiani et al. *Artificial Intelligence and Law*, pages 1–26.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Jivitesh Jain, Nivedhitha Dhanasekaran, and Mona T. Diab. 2025. From complexity to clarity: AI/NLP's role in regulatory compliance. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26629–26641, Vienna, Austria. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. *arXiv preprint arXiv:2408.01084*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Taiji Li, Zhi Li, and Yin Zhang. 2024. Improving faithfulness of large language models in summarization via sliding generation and self-consistency. *arXiv preprint arXiv:2407.21443*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding:

Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Savinay Narendra, Kaushal Shetty, and Adwait Ratnaparkhi. 2024. Enhancing contract negotiations with llm-based legal document comparison. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 143–153.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Ohad Rubin and Jonathan Berant. 2024. Retrieval-pretrained transformer: Long-range language modeling with self-retrieval. *Transactions of the Association for Computational Linguistics*, 12:1197–1213.

TYSS Santosh, Hassan Sarwat, Ahmed Mohamed Abdelaal Abdou, and Matthias Grabmair. 2024. Mind your neighbours: Leveraging analogous instances for rhetorical role labeling for legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11296–11306.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th international conference on content-based multimedia indexing*, pages 1–7.

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

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024a. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Santosh T.y.s.s, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, and Xiaomo Liu. 2025. CoCoLex: Confidence-guided copy-based decoding for grounded legal text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19002–19018, Vienna, Austria. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chi-Hung Wang, Xiang-Shun Yang, Jun-Yi Liu, and Yao-Jun Liu. 2025a. Contract pre-review assistance system based on rag and llm. In *International Congress on Information and Communication Technology*, pages 369–380. Springer.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. 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 11636–11652, Albuquerque, New Mexico. Association for Computational Linguistics.

Korbinian Q. Weidinger, Santosh T.y.s.s, Oana Ichim, and Matthias Grabmair. 2025. AQuAECHR: Attributed question answering for European court of human rights. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1418–1447, Vienna, Austria. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering. In *Case-Based Reasoning Research*

and Development, pages 445–460, Cham. Springer Nature Switzerland.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. *arXiv preprint arXiv:2203.08913*.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. *arXiv preprint arXiv:2405.02750*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuyan Zhou, Uri Alon, Frank F Xu, Zhengbao Jiang, and Graham Neubig. 2022. Docprompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

A More Discussion on Related Work

LLM in Legal AI. LLMs are increasingly adopted in legal practice for tasks such as contract drafting, document review, and legal research, offering notable efficiency gains and becoming an integral part of routine legal workflows (T.y.s.s et al., 2025). However, their deployment introduces significant ethical and practical challenges (Avery et al., 2023; Harasta et al., 2024; Chitgopkar et al.; T.y.s.s et al., 2025). Recent studies have systematically evaluated the reliability of LLMs in legal settings: Dahl et al. (2024) analyzes hallucination behaviors across legal tasks, while Magesh et al. (2025) demonstrates that although retrieval-augmented generation reduces hallucinations in legal question answering, unfaithful and unsupported outputs remain a persistent challenge. CoCoLex (T.y.s.s et al., 2025) introduces a confidence-guided, copy-based decoding strategy that dynamically interpolates the model's vocabulary distribution with a distribution derived from copying tokens directly from the retrieved context. In this work, we propose a fine-grained decoding strategy within an RAG framework. Specifically, rather than uniformly leveraging the entire retrieved context or directly copying from it, we assign greater influence to tokens originating from high-attention sentences, which are identified as being most relevant to the query. These weighted token contributions are then incorporated into the token generation process, enabling the model to maintain interpretive precision in the legal domain by improving faithfulness to the most salient portions of the retrieved evidence.

Retrieval-Augmented Generation. enhances large language models by incorporating relevant information retrieved from external knowledge sources, with the goal of improving reasoning performance and reducing hallucinated responses. Retrieval can be conducted at varying levels (Guu et al., 2020; Lewis et al., 2020; Khandelwal et al., 2019; Yogatama et al., 2021; De Jong et al., 2021; Févry et al., 2020). The retrieved evidence may then be integrated into the language model at different stages, namely the input layer (Ram et al., 2023; Izacard et al., 2023), intermediate layers (Wu et al., 2022; Borgeaud et al., 2022), or the output layer (Khandelwal et al., 2019; Santosh et al., 2024). Existing RAG approaches can be broadly categorized into training-based methods, which independently fine-tune the retriever and generator (Karpukhin et al., 2020; Zhou et al., 2022; Borgeaud et al., 2022; Yoran et al., 2023; Lin et al., 2023; Sarto et al., 2022; Rubin and Berant, 2024; Izacard et al., 2023), and training-free methods, which incorporate retrieved knowledge at inference time by augmenting the input prompt (Jiang et al., 2023b; Khattab et al., 2022; Trivedi et al., 2023). Simple augmentation of the input with retrieved context causes

inconsistency between the generated output and the supporting evidence. To mitigate this issue, prior work has proposed decoding-based strategies (Shi et al., 2024b; Kim et al., 2024; Zhao et al., 2024) and prompt-based techniques (Zhou et al., 2023; Byerly and Khashabi, 2024; Li et al., 2024) to steer generation toward greater faithfulness. Contrastive decoding (Li et al., 2023) amplifies the influence of retrieved context by modifying token logits, but they do not explicitly enforce adherence to the source content. Copy-based decoding (T.y.s.s et al., 2025), which manipulates token production distribution by integrating distribution derived based on duplicating tokens from the context, incorporates information from the non-evidence part in the context, therefore reducing the performance. To overcome this limitation, we propose a novel decoding method that utilizes relevant information within the context via a copy-based decoding mechanism.

B Experiment Details

B.1 Dataset

Table 7 summarizes the statistics of the four legal benchmarks used in our experiments. The datasets vary significantly in scale, document length, and answer complexity. **CUAD** focuses on clause extraction from dense contracts with relatively short answers. **OALQA** contains Australian legal QA pairs requiring concise responses grounded in judicial decisions. **ObliQA** presents the longest documents (averaging 48K tokens) with regulatory content from financial regulations. **AQuAECHR** involves multi-document reasoning over European Court of Human Rights judgments, with the longest average answer length among all datasets.

Dataset	#Instances	#Docs	Docs Len.	Answer Len.
CUAD	4,182	1.00	7,125.60	31.11
OALQA	2,024	1.00	5,973.99	90.07
ObliQA	2,786	1.32	48,138.35	117.55
AQuAECHR	1,116	3.11	14,041.33	193.58

Table 7: Dataset statistics for the four legal benchmarks used in our experiments.

B.2 Implementation Details

For retrieval, we use BM25 to fetch the top-k relevant chunks from the document corpus, setting $k = 3$ for all benchmarks. For the baselines, we adhere to the optimal hyperparameters reported in their respective literature. For CAD, we set the static weight $\alpha = 0.5$. For AdaCAD, we clamp the adaptive weight $\alpha_t \in [0.3, 1.0]$. For CoCoLex, we use the last-layer hidden states for similarity computation with a smoothing factor of $\lambda = 0.5$. For our proposed SACoD, we set the attention layer $\mathcal{L}_{\text{top}}$ to the last 50% of transformer layers, the attention evidence threshold to $\alpha = 0.7$, and the Otsu threshold update interval to $K = 10$ for all experiments. We implement SACoD using PyTorch, leveraging `pytorch` version 2.8.0 and the Hugging Face `transformers`[1] library (version 4.40.2).

B.3 Models and Hardware

We conduct our experiments primarily with the following models:

- **Mistral-7B-Instruct-v0.3**[2] is an instruct fine-tuned version of the Mistral-7B-v0.3 with extended vocabulary to 32,768 tokens and supporting v3 tokenizer.
- **Meta-Llama-3-8B-Instruct**[3] are optimized for dialogue use cases and several question answering tasks. It was trained on a massive text of 15 trillion tokens, with a context length of 8,000 tokens.
- **Qwen-2.5-7B-Instruct**[4] achieve improvements in instruction following, generating long texts (over 8K tokens) and long-context support up to 128K tokens.

All checkpoints are obtained from official repositories or Hugging Face without any fine-tuning or modification. All models are evaluated using 16-bit (FP16) precision. The majority of our experiments are performed on a single NVIDIA A6000 GPU with 48GB of memory.

C Additional Experiment Results

C.1 Details on Motivation Observations

Motivating Observations. We begin by examining how evidence is used in legal grounded generation, deriving two observations that motivate our design. Since existing legal RAG benchmarks lack evidence annotations for generated outputs, we construct a dataset of 80 expert-annotated samples (20 per benchmark) from our evaluation datasets

[1] https://github.com/huggingface/transformers
[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

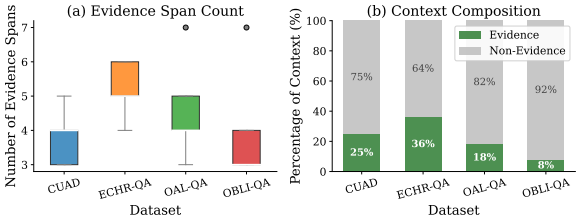


Figure 4: Evidence Sparsity Analysis. (a) Number of evidence paragraphs per sample. (b) Evidence vs. non-evidence context proportion. Only 8–36% of retrieved context supports the generated answers.

(Hendrycks et al., 2021; Weidinger et al., 2025; Butler, 2023; Gokhan et al., 2024). Two annotators with legal training independently labeled each sample, achieving high inter-annotator reliability with a Cohen's Kappa of 0.8720.

Each instance consists of an input query x, a retrieved context c, and a model-generated answer $\mathbf{y} = \{y_1, \ldots, y_n\}$. Annotators identified which paragraph(s) in c directly support the claims in the output, following detailed guidelines that define supporting evidence as spans containing facts, legal provisions, or reasoning explicitly used in the answer. We segment the context into m paragraphs $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m\}$, where each $\mathbf{p}_i$ spans token indices $(t_i^{\text{start}}, t_i^{\text{end}})$.

Observation 1: Evidence is sparse. Despite the length and heterogeneity of legal contexts, expert annotations reveal that only 8–36% of the retrieved content directly supports the generated answers, while the remainder provides no substantive contribution (Figure 4). This high concentration of relevance indicates that treating the entire retrieved context as uniformly informative is both unnecessary and potentially detrimental—it increases exposure to irrelevant spans and may encourage blind copying of templated content. This observation motivates SACoD's evidence selection mechanism, which identifies a compact evidence set $\mathcal{P}_E \subset \mathcal{P}$ and grounds decoding on $\mathcal{P}_E$ rather than the full context.

Observation 2: Attention correlates with human-selected evidence. We observe that the model's attention provides a useful signal for identifying supporting evidence. At decoding step t, let $\mathbf{a}_t^{(l)} \in \mathbb{R}^{|c|}$ denote the head-averaged attention from layer l over context tokens. We define the paragraph-level attention for $\mathbf{p}_i$ as

$$\bar{a}_{t,i}^{(l)} = \frac{1}{t_i^{\text{end}} - t_i^{\text{start}} + 1} \sum_{j=t_i^{\text{start}}}^{t_i^{\text{end}}} \mathbf{a}_{t,j}^{(l)}. \quad (6)$$

To capture evidence relevance across the entire answer, we aggregate over all decoding steps:

$$e_i^{(l)} = \frac{1}{n} \sum_{t=1}^{n} \bar{a}_{t,i}^{(l)}. \quad (7)$$

Figure 3 shows that paragraphs with higher $e_i^{(l)}$ are substantially more likely to be expert-annotated as evidence. Notably, deeper layers assign significantly higher attention to evidence paragraphs than to non-evidence ones—a pattern that holds consistently across model families and datasets. This finding suggests that attention provides a lightweight, training-free proxy for paragraph-level evidence relevance, computable during generation without additional forward passes or auxiliary models. By aggregating attention within paragraphs and across decoding steps, we obtain a stable signal that aligns well with human judgments, motivating its use for self-guided evidence identification.

Human-annotated evidence For our expert-annotated samples, we achieved a high level of inter-annotator reliability, yielding an overall Cohen's Kappa of 0.8720 and an Agreement Rate of 93.61%.

Annotation Guidelines To ensure consistency and legal rigour, our two expert annotators (both with legal training) followed the guidelines below.
Objective. For each instance, identify every text span in the retrieved context c that *directly* supports one or more claims made in the model-generated answer y. A text span is the atomic annotation unit, defined as any contiguous sequence of one or more sentences within c that constitutes a coherent semantic unit. Annotators may freely set span boundaries at the sentence level; they are not constrained to pre-defined structural units such as paragraphs or sections.
Task Setup. Each annotation instance comprises three elements: (i) an Input Query x; (ii) a Retrieved Context c with sentences individually numbered to facilitate precise span boundary specification; and (iii) a Model-Generated Answer y. Annotators record each selected span as a start–end sentence index pair (e.g., `[s14-s17]`). Annotators

are required to read all three elements in full before making any selection.

Taxonomy of Supporting Evidence. Each text span in c receives one of two mutually exclusive labels:

[E] Evidence. The span directly supplies the textual basis for at least one claim in y. This encompasses spans containing operative legal provisions applied in the answer, case-specific facts or procedural details relied upon to reach a conclusion, judicial or administrative reasoning reproduced or paraphrased in the answer, or precedents and authorities *actively invoked* (not merely cited in passing). The test is strictly textual: annotators must be able to point to a specific sentence within the span that corresponds to a specific claim in y. If that correspondence cannot be located, the span is **[N]**. Span boundaries should be drawn as tightly as the correspondence permits; annotators must not inflate spans to include surrounding sentences that do not satisfy the test.

[N] Not Evidence. The span does not meet the **[E]** threshold. This includes background and procedural context that does not contribute any specific claim, boilerplate and formulaic recitals whose content is not substantively used, precedents that the answer explicitly distinguishes or overrides, and any span whose connection to y requires inferential steps beyond explicit textual correspondence.

Every sentence in c belongs to exactly one span and receives exactly one label. The default label is **[N]**; annotators assign **[E]** only when the explicit correspondence test is satisfied.

Annotation Process.

1. **Independent reading.** Each annotator reads the query x and answer y in full before inspecting the context c.

2. **Claim decomposition.** Annotators decompose y into atomic claims (factual assertions, legal conclusions, citations) and record each claim before beginning span selection, to avoid anchoring bias.

3. **Span selection.** For each claim, identify the minimal contiguous span(s) in c that supply its direct textual basis. A span may be selected for multiple claims; overlapping spans from different claims are merged into a single **[E]** span. If no span in c supports a given claim, it is recorded as unsupported (this does not disqualify the instance).

4. **Confidence scoring.** Each selected span is assigned a binary confidence label: CERTAIN (the span unambiguously supports the claim) or BORDERLINE (the connection requires interpretive judgment). Borderline selections are flagged for adjudication.

5. **Conflict resolution.** After independent review, annotators compare selections. Disagreements on span boundaries or CERTAIN labels are resolved by discussion to reach consensus. Persistent disagreements on BORDERLINE items are resolved by a third legal expert acting as adjudicator, whose decision is final.

Edge Cases.

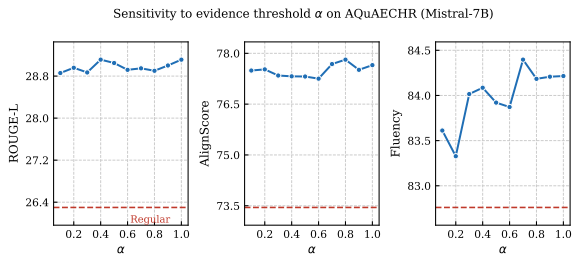
- **Multi-hop reasoning.** If an answer synthesises facts from two or more non-adjacent spans that are individually insufficient to support the claim, all contributing spans are marked **[E]** independently, even if they are not contiguous.

- **Minimal span principle.** Annotators should select the smallest span that satisfies the correspondence test. Where a single sentence suffices, the span should not be extended to its surrounding context unless that context is itself evidentially necessary.

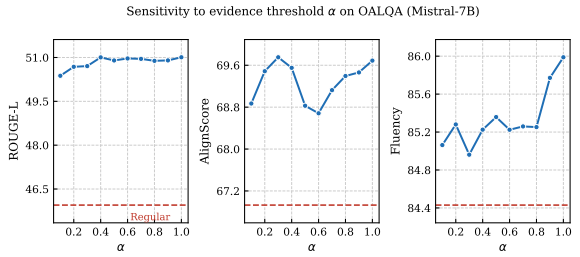
- **Holding vs. dicta.** In judgments, only spans containing the court's operative holding or the direct reasoning leading to it qualify. Obiter dicta spans are excluded unless the answer explicitly relies on them.

C.2 Details on Human Evaluation

We randomly sample 20 questions from the AQuAECHR dataset and generate responses using three methods: Regular, CoCoLex, and SACoD, applied to the Mistral-7B model, obtaining a total of 60 responses. Each response is assessed by a legal expert across four criteria: correctness (relevance to the question and alignment with the reference answer), faithfulness (adherence to the provided context), fluency, and coherence. Evaluations are conducted on a 5-point Likert scale, where 1 represents the lowest quality and 5 the



(a) AQuAECHR



(b) OALQA

Figure 5: Sensitivity to evidence threshold α on AQuAECHR (a) and OALQA (b) using Mistral-7B. Dashed lines denote Regular decoding. In both cases, SACoD consistently outperforms the baseline across metrics, showing high robustness (flat profile) to this hyperparameter. Gains are more pronounced on OALQA due to its longer answer format.

highest. Table 3 reports the average scores for each criterion across methods. Notably, legal experts consistently ranked SACoD higher than Regular and CoCoLex, reinforcing its superiority in generating legally faithful responses. Qualitative examples illustrating these differences are provided in Appendix E.

C.3 Sensitivity to Evidence Threshold α

Figures 5 show the effect of varying $\alpha \in [0.1, 1.0]$ on AQuAECHR and OALQA, respectively (both Mistral-7B). Across all ten values and both datasets, SACoD consistently outperforms Regular decoding on ROUGE-L and AlignScore by a clear margin. The near-flat profile on AQuAECHR confirms that $\alpha = 0.7$ is a robust default rather than a carefully tuned optimum. The OALQA profile is similarly stable, though the absolute gains are larger given the dataset's longer, multi-sentence answers, where selective evidence focus yields more benefit. Together, these results confirm that the method's gains do not depend on precise threshold selection across diverse task types.

C.4 Entropy Gate Activation Rate

A key factor behind SACoD's latency advantage is that the entropy gate activates the copy mechanism on only a small fraction of decoding steps. Table 8 reports the gate statistics on AQuAECHR using Mistral-7B-Instruct-v0.3. Across 50 samples, the gate fires on 1,275 out of 8,180 total steps, yielding an overall activation rate of 15.6% and a per-sequence mean of 15.2%. On average, only 25.5 out of 163.6 steps per sequence invoke the evidence-selective copy mechanism; the remaining 84% of tokens are generated with standard vanilla decoding at zero additional overhead. This directly explains the latency profile in Table 4: the computational cost of attention extraction, evidence selection, and datastore construction is incurred only at the small subset of steps where the model is genuinely uncertain, rather than at every token as in CoCoLex.

Statistic	Value
Samples evaluated	50
Total decoding steps	8,180
Mean steps per sequence	163.6
Gate-on steps (total)	1,275
Mean gate-on steps per seq.	25.5
Overall activation rate	15.6%
Mean activation rate	15.2%

Table 8: Entropy gate activation statistics on AQuAECHR (Mistral-7B-Instruct-v0.3). The gate triggers on only ∼15% of decoding steps, confirming that vanilla decoding is preserved for the large majority of tokens.

C.5 Layer Selection Analysis

A key design choice in SACoD is which transformer layers to use for computing evidence scores. As demonstrated in Figure 3, deeper layers exhibit stronger alignment between attention patterns and human-annotated evidence paragraphs. Here, we systematically evaluate how different layer selection strategies affect downstream generation quality.

We evaluate three layer selection strategies on the Mistral-7B model across all four legal benchmarks: (1) *All layers* (0%–100%), which aggregates attention from the entire transformer stack; (2) *Early layers* (0%–50%), which uses only the first half of layers; and (3) *Late layers* (50%–100%), which uses only the deeper half, corresponding to our default configuration $\mathcal{L}_{\text{top}} = \{l \mid$

Layer Span	RL	AS	Flu	Avg
0%–100% (All)	29.27	20.09	**83.26**	44.21
0%–50% (Early)	29.22	19.77	82.88	43.96
50%–100% (Late)	**29.80**	**20.46**	83.19	**44.48**

Table 9: Effect of layer selection on generation quality. Late layers (50%–100%) achieve the best balance across correctness (RL), faithfulness (AS), and fluency. Results averaged across four legal benchmarks on Mistral-7B.

$l > L/2\}$. Table 9 presents the results. The late-layer configuration (50%–100%) achieves the highest average score (44.48), outperforming both the all-layer baseline (44.21) and the early-layer variant (43.96). Specifically, late layers yield improvements of +0.53 in ROUGE-L and +0.37 in AlignScore compared to using all layers, while maintaining comparable fluency (83.19 vs. 83.26). These findings align with our motivating observation (Section 3): deeper transformer layers encode more semantically meaningful attention patterns that better distinguish relevant evidence from noise. Early layers, in contrast, tend to capture shallow lexical and syntactic patterns that are less discriminative for evidence identification. Including early layers (0%–100%) dilutes the signal with less informative attention, slightly degrading correctness and faithfulness.

C.6 Analysis of Context-size Robustness

Table 10 illustrates the impact of varying the number of provided passages on model performance using the AQuAECHR dataset. Across all methods, increasing the context size generally improves Rouge-L, AlignScore, and Fluency, suggesting that providing more context helps the model generate more accurate, faithful, and comprehensive responses. SACoD consistently outperforms both Regular decoding and CoCoLex across different context lengths in all metrics. In particular, SACoD pushes performance to the highest level overall with 5 passages, achieving **32.10** ROUGE-L, **77.90** AlignScore, and **87.14** Fluency. SACoD dynamically selects salient evidence spans and activates context-aware copying only when the model is uncertain. This allows it to effectively use the extra passages when they contain useful legal evidence, while filtering out distracting text and avoiding unnecessary intervention for confident tokens.

Method	#Psg	RL	AS	Flu
Regular	1	25.74	59.07	77.85
	3	26.30	73.45	82.76
	5	29.59	76.49	83.37
CoCoLex	1	29.10	55.25	75.70
	3	29.15	74.20	81.48
	5	29.84	76.92	82.57
SACoD	1	29.21	63.59	79.76
	3	29.80	77.09	83.19
	5	**32.10**	**77.90**	**87.14**

Table 10: Performance comparison on the AQuAECHR dataset using Mistral-7B-Instruct-v0.3 with varying context sizes (#Psg). RL, AS, and Flu denote Rouge-L, Faithfulness-AlignScore, and Fluency, respectively.

C.7 Evaluation on Non-Legal Datasets

To examine whether SACoD generalizes beyond the legal domain, we conduct additional experiments on news summarization using the CNN/DailyMail (CNN-DM) dataset (See et al., 2017), a widely adopted benchmark for abstractive summarization that pairs news articles with multi-sentence summaries. We randomly sample 200 examples from the original 11.5K test set and evaluate all decoding methods under the same metrics used in our main experiments.

As shown in Table 11, SACoD consistently achieves the highest ROUGE-L and AlignScore across both datasets and backbone models, demonstrating that its benefits generalize beyond the legal domain. On CNN-DM with Mistral-7B, SACoD improves ROUGE-L by +2.96 and AlignScore by +7.16 over Regular decoding, while simultaneously improving fluency (89.84 vs. 87.68). A similar trend is observed on Qwen2.5-7B, where SACoD pushes performance to the highest levels across all three metrics. Compared to CoCoLex, which also leverages contextual copying, SACoD achieves higher correctness and faithfulness with notably better fluency (+1.39 on Mistral-7B and +1.99 on Qwen2.5-7B for CNN-DM), confirming that entropy-gated, evidence-selective copying averts the fluency degradation often caused by indiscriminate whole-context aggregation. Furthermore, evaluation on PubMedQA demonstrates that SACoD generalizes effectively to the complex biomedical domain. Across both models on PubMedQA, SACoD robustly outperforms Regular decoding, AdaCAD, and CoCoLex in generation quality (ROUGE-L) and faithfulness (AlignScore)

with minimal to no impact on fluency. These results robustly validate that SACoD's self-guided framework is broadly effective for context-grounded generation tasks across diverse semantic domains and model architectures.

C.8 Efficiency Framing

To better understand the operational overhead of SACoD, we conduct a comprehensive efficiency profiling covering both latency and memory consumption using the Mistral-7B-Instruct-v0.3 model.

Latency Breakdown: We analyze the proportional time spent across different stages of the SACoD pipeline at each decoding step. The primary bottleneck is *Input Extraction* (48.64%), representing the initial context processing and representation mapping. *Base LLM Inference* requires the next largest share at 27.74%. The operations introduced specifically by SACoD's dynamic copying mechanism add minimal latency overhead: *Evidence Selection* (8.33%), *Attention Extraction* (8.08%), and *Datastore Construction* (6.11%). Finally, the actual *Copying / Sample Next Token* stage accounts for just 1.10% of the total latency. This breakdown highlights that SACoD's evidence-gated interventions introduce only a negligible computational delay over standard context processing.

Memory Analysis: Table 12 details the peak memory consumption of different decoding s trategies. SACoD incurs the highest peak memory usage among all methods (31.10 GB), exceeding CoCoLex (2 9.06 GB) by ∼7% and AdaCAD (28.51 GB) by ∼9%. This overhead is directly attributable to caching the l ayer-wise attention distributions required for self-guided evidence identification and maintaining the dynamic, t oken-level datastore. While SACoD does not offer a memory advantage over other context-aware methods, th is cost is offset by its latency profile: unlike contrastive methods (CAD and AdaCAD), it requires no additional forward passes, yielding substantially lower decoding latency while achieving superior faithfulness.

C.9 Finer-Grained Evidence Selection

To systematically assess the computational tradeoffs and performance impacts of finer-grained evidence selection, we evaluated SACoD at the sentence and word levels. The empirical results demonstrate that while finer granularity is computationally feasible, it does not yield a favorable performance trade-off compared to our default paragraph-level approach.

- **Word-level:** Individual words in the retrieved context are segmented using `spacy` and assigned attention scores based on the average attention of the tokens they contain.
- **Sentence-level:** Context is segmented using spacy; attention is averaged over sentence spans to select the top-α percentile.
- **Paragraph-level (Default):** Context is split by double newlines and header patterns; attention is averaged over paragraph spans.

Results and Analysis: Table 13 presents the evaluation of word- and sentence-level selection against our default paragraph-level approach. Transitioning to finer granularities introduces minimal latency overhead (maintaining a tight bound of 1.38×–1.49×), as token-level attention weights are inherently computed during the model's standard forward pass without necessitating additional generation cycles.

Despite this computational efficiency, paragraph-level selection offers the most favorable and robust trade-off between ROUGE-L, faithfulness (AlignScore), and generation fluency across the majority of conditions. For instance, on the complex ObliQA benchmark using Qwen-2.5-7B, paragraph-level segmentation achieves the highest average performance (72.03), substantially outperforming the word-level variant (70.98), largely owing to a notable degradation in faithfulness where AlignScore drops from 92.09 (paragraph) to 90.35 (word). A similar trend is observed with Mistral-7B on ObliQA, where shrinking the selection boundary to single words leads to diminished faithfulness (84.12 vs. 87.83) as well as impaired fluency (75.98 vs. 77.12).

While word-level selection exhibits a marginal improvement over the paragraph baseline on AQuAECHR utilizing Mistral-7B (achieving an average of 63.66 versus 63.36), it predominantly demonstrates inferior performance in the broader cross-section of models and tasks.

We attribute this prevalent structural degradation to semantic fragmentation. Isolating discrete words or disjointed sentences strips away the encompassing discourse essential for coherent, multi-step reasoning. By operating at the paragraph structural

Model	Method	CNN-DM			PubMedQA		
		RL	AS	Flu	RL	AS	Flu
Mistral-7B	Regular	22.96	83.79	87.68	16.28	54.65	**93.89**
	AdaCAD	23.21	85.39	84.32	16.77	57.91	93.33
	CoCoLex	23.52	88.96	88.45	17.84	58.57	89.16
	SACoD	**25.92**	90.95	**89.84**	**19.41**	**60.24**	93.75
Qwen2.5-7B	Regular	21.64	**85.57**	85.78	15.65	56.57	82.34
	AdaCAD	**22.21**	88.34	86.46	16.37	59.62	82.49
	CoCoLex	21.53	90.25	84.92	17.86	58.22	78.14
	SACoD	**23.89**	**91.91**	**86.91**	**18.91**	**60.49**	**82.76**

Table 11: Cross-domain generalization: Performance of different decoding methods on news summarization (CNN-DM) and biomedical QA (PubMedQA) datasets.

Method	Used Mem. (GB)	Relative
Vanilla	14.40	1.00x
AdaCAD	28.51	1.98x
CoCoLex	29.06	2.02x
SACoD	31.10	2.16x

Table 12: Peak memory footprint comparison across decoding methods using Mistral-7B-Instruct-v0.3.

level, SACoD preserves complete semantic units and rhetorical context, thereby empowering the model to draw upon cohesive evidence blocks that seamlessly integrate into factually grounded, fluent responses.

D Details on SACoD

D.1 Dynamic Entropy-Gated Activation

A key challenge in context-aware decoding is deciding *when* contextual intervention is actually necessary. During generation, tokens fall into two broad categories. Many tokens are highly predictable, such as function words, punctuation, and legal boilerplate, where the model exhibits strong confidence and little ambiguity. Other tokens are inherently uncertain, including named entities, dates, statutory references, and decision outcomes, where multiple alternatives are plausible and external evidence is often required.

Existing context-aware methods (Shi et al., 2024a; Wang et al., 2025b; T.y.s.s et al., 2025) typically intervene at *every* decoding step, treating these two cases uniformly. This design is inefficient and can be counterproductive: injecting contextual bias at deterministic steps adds unnecessary computation and may amplify irrelevant or templated spans. In contrast, the uncertain steps are precisely where hallucinations and evidence misuse tend to occur, and where targeted contextual guidance is most beneficial.

This motivates a lightweight gate that activates contextual guidance only at *intervention points*. We operationalize this idea using the model's own uncertainty signal. At decoding step t, we compute the normalized entropy $h_t \in [0, 1]$ of the next-token distribution $p_t(v) = p_\Theta(y_t = v \mid c, x, y_{<t})$ to quantify uncertainty:

$$h_t = -\frac{1}{\log |\mathcal{V}|} \sum_{v \in \mathcal{V}} p_t(v) \log p_t(v), \qquad (8)$$

This entropy measure dictates the soft interpolation weight $\lambda_t = \exp(-h_t)$, ensuring that as uncertainty grows (high h_t), the parametric prior yields more mass to the context-aware copy mechanism.

Using this entropy measure, we now formally define the interpolation weight λ_t introduced in Eq. 4: $\lambda_t = \exp(-h_t)$ which maps normalized entropy to a confidence-based weight. When the model is confident (low h_t), λ_t approaches 1, causing the base distribution to dominate. When the model is uncertain (high h_t), λ_t decreases toward $\exp(-1) \approx 0.37$, allowing the copy mechanism to exert greater influence.

To strictly avoid unnecessary computation, we introduce a dynamic gate whose threshold θ^* is determined automatically from the model's own entropy signal in two phases.

Prefill initialization. Before any tokens are generated, the model processes the full input $(c \oplus x)$ in its standard forward pass. We record the normalized entropy h_i^{pfx} at each of the N input positions, forming the initial entropy buffer $\mathcal{H}_0 = \{h_i^{\text{pfx}}\}_{i=1}^N$. Applying Otsu's method (Eq. 10) to $\mathcal{H}_0$ yields the initial threshold θ_0^*, which separates the input's confident tokens (e.g., syntactic connectors, boil-

Model	Level	ROUGE-L	AlignScore	Fluency	Average	Latency
			ObliQA			
Mistral-7B	Paragraph	46.84	87.83	77.12	**70.60**	1.48×
	Sentence	46.43	86.51	76.41	69.78	1.48×
	Word	47.62	84.12	75.98	69.24	1.46×
Qwen2.5-7B	Paragraph	45.32	92.09	78.84	**72.03**	1.42×
	Sentence	44.62	91.18	79.16	71.65	1.40×
	Word	43.4	90.35	79.18	70.98	1.45×
			AQuAECHR			
Mistral-7B	Paragraph	29.80	77.09	83.19	63.36	1.39×
	Sentence	30.11	76.42	83.37	63.30	1.44×
	Word	30.32	76.78	83.88	**63.66**	1.43×
Qwen2.5-7B	Paragraph	29.24	80.01	80.77	**63.34**	1.38×
	Sentence	29.23	78.51	80.24	62.66	1.42×
	Word	29.51	78.69	80.20	62.80	1.49×

Table 13: Performance comparison of SACoD across different segmentation level of evidence selection.

erplate) from uncertain ones without any manual tuning.

Generation-time gating and updates. At each decoding step t, the gate compares h_t against the current threshold:

$$\texttt{Gate}(t) = \begin{cases} \texttt{Vanilla}, & h_t < \theta^*, \\ \texttt{Copy}, & h_t \geq \theta^*, \end{cases} \tag{9}$$

where `Vanilla` uses the base model distribution p_Θ directly, and `Copy` triggers the entropy-weighted interpolation from Eq. 4. As generation proceeds, each new entropy value h_t is appended to the buffer: $\mathcal{H} = \mathcal{H}_0 \cup \{h_1, \ldots, h_{t-1}\}$. To keep the threshold adapted to the current output distribution without per-step overhead, θ^* is refreshed by re-running Otsu on the full buffer $\mathcal{H}$ every $K = 10$ steps; the threshold remains fixed between updates.

This two-level design, which combines adaptive hard gating with soft interpolation, avoids unnecessary intervention on predictable tokens while concentrating computation on uncertain, high-impact steps where evidence grounding matters most. The complete SACoD algorithm is provided in Algorithm 1.

D.2 Overview of SACoD Procedure

Algorithm 1 presents the complete SACoD procedure. At each decoding step, the method first computes the normalized entropy to determine model uncertainty. For low-uncertainty steps, vanilla decoding is used directly. For high-uncertainty steps, the algorithm dynamically identifies evidence paragraphs using attention signals at the current step, builds a step-specific datastore from the selected evidence, and applies entropy-weighted interpolation between the base model and copy distributions.

D.3 Dynamic Entropy-Guided Algorithm

To find the appropriate threshold θ that distinguishes between high-confidence (predictable) and low-confidence (uncertain) tokens, we apply a 1D variant of Otsu's method (Algorithm 2), a nonparametric thresholding technique originally developed for image segmentation. Instead of treating θ as a static, manually tuned hyperparameter, this approach dynamically discovers a natural boundary directly from the model's token distribution.

Given a set of normalized token entropy values $\mathcal{H}$ gathered from the input prefix (and periodically augmented during autoregressive generation), the algorithm sorts and evaluates every unique value as a candidate threshold. For each candidate θ, it partitions the entropy distribution into two disjoint classes: a low-entropy class $\mathcal{H}_0 = \{h \in \mathcal{H} \mid h < \theta\}$ representing confident tokens, and a high-entropy class $\mathcal{H}_1 = \{h \in \mathcal{H} \mid h \geq \theta\}$ representing uncertain tokens.

The optimal natural threshold θ^* is defined as the value that maximizes the inter-class variance σ^2:

$$\theta^* = \arg\max_{\theta} \Big[w_0(\theta) \cdot w_1(\theta) \cdot \big(\mu_0(\theta) - \mu_1(\theta)\big)^2 \Big] \tag{10}$$

where $w_0(\theta)$ and $w_1(\theta)$ are the proportions of tokens falling into each class ($|\mathcal{H}_0|/|\mathcal{H}|$ and

$|\mathcal{H}_1|/|\mathcal{H}|$, respectively), and $\mu_0(\theta)$ and $\mu_1(\theta)$ are their respective mean entropy values.

By maximizing the inter-class variance, the algorithm naturally identifies the most distinct cutoff separating standard language modeling priors from highly uncertain states. This allows the SACoD entropy gate to selectively trigger the computationally intensive copy-mechanism only precisely when needed, bypassing intervention on deterministic segments.

E Case Study

E.1 Static vs. Shifting Relevance in Evidence Selection

To further illustrate the importance of dynamically updating the evidence datastore, we analyze the generation process of SACoD on Case Study 2 in Table 15 compared to a variant without dynamic evidence ("w/o Dynamic Evidence"). The static approach identifies evidence using attention signals exclusively from the very first decoding step and locks that set in place for the entire generation process. This causes severe degradation—especially in long-form answers—because different parts of a complex answer naturally require different pieces of evidence as the generation progresses.

Fixed Evidence (Step 1): When the evidence is identified at the start of generation (Step 1), the model casts a wide net. It selects a broad array of 10 different snippets spanning paragraphs 71, 70, and 64. This block includes procedural history about Polish authorities, foreign birth certificates, and the biological father "Mr S", alongside broader definitions of private life:

- *"001-214296 71. In that connection, the present case must be clearly distinguished from Mennesson and Labassee."*
- *"In the present case it is true that the Polish authorities refused to give effect to the foreign birth certificates establishing the legal parent-child relationship between the applicants and their biological father, Mr S. However, this link is recognised in the country where the family resides."*
- *"001-214296 70. Furthermore, in the circumstances of this case the Court notes that the applicants can benefit, in the State where they live, from the legal parent - child relationship with their biological father..."*
- *"001-214296 64. The Court notes that the concept of 'private life' is a broad term not susceptible to exhaustive definition."*
- *"It covers the physical and psychological integrity of a person."*
- *"It can therefore embrace multiple aspects of the person's physical and social identity which includes the legal parent-child relationship...; an essential aspect of the identity of individuals is at stake where the legal parent-child relationship is concerned..."*

If the model is forced to use only this fixed block of text for the entire response, it struggles to maintain precision as it generates later sentences, leading to noise and less faithful outputs.

Dynamically Updated Evidence (Step 110): When the dynamic evidence selection is enabled, the model updates its focus at specific intervention points (when the confidence gate triggers). At Step 110, the model is trying to generate a specific legal standard. Instead of relying on the broad, 10-snippet context from Step 1, its attention hones in on just two highly specific target sentences (indices 10 and 11 from the context pool):

- **Index 10:** *"It covers the physical and psychological integrity of a person."*
- **Index 11:** *"It can therefore embrace multiple aspects of the person's physical and social identity which includes the legal parent-child relationship...; an essential aspect of the identity of individuals is at stake where the legal parent-child relationship is concerned (see Mennesson v. France, no."*

The Impact on Output: Because the model dynamically updated its evidence to focus intensely on sentences 10 and 11 at step 110, it was able to perfectly reproduce that exact legal reasoning in its output: *"However, the Court has emphasized that the concept of 'private life' is a broad term that covers multiple aspects of a person's identity, including the legal parent-child relationship [001-214296, 64]."*

In conclusion, dynamically updating evidence allows the model to filter out irrelevant paragraphs and pinpoint the exact legal clauses needed for the specific sentence it is currently writing. Fixing the evidence at Step 1 strips the model of this precision, forcing it to navigate a stagnant pool of con-

text, which directly causes the severe performance penalty reported in our ablations.

E.2 Qualitative Examples

We present qualitative examples comparing SACoD with baseline methods and the reference answer on two legal QA benchmarks, illustrating how different decoding strategies affect faithfulness, completeness, and coherence. We provide Case Study 1 in Table 14 and responses generated from three decoding methods: Regular, CoCoLex, and SACoD, to a question derived from OALQA dataset. The question at hand revolves around the court's decision regarding the motion to restrain a film from action in the case of Moore v Scenic Tour Pty Ltd [2025].

Among the methods, SACoD stands as the most faithful to the source document by directly incorporating key legal terminology and phrases such as "not in the interest of justice to deprive the plaintiff of their choice of solicitor.". In addition to correctly stating the dismissal of the motion and associated procedural orders, it explicitly identifies the court's core reasoning: that it was not in the interest of justice to deprive the plaintiff of their choice of solicitor. This reasoning is central to the decision and is highlighted in the case catchwords and reference answer, making SACoD the most complete response in terms of correctness.

On the other hand, Regular decoding output correctly identifies the outcome, that the motion to restrain the firm from acting was dismissed. This aligns with both the context and the reference text at a high level. However, the response remains purely outcome-oriented and does not engage with the legal basis for the decision. The question asks what the court's decision was, which in a legal QA setting typically encompasses not only the order made but also the reason why the court exercised its discretion in that manner. By omitting the court's reasoning—specifically, the principle that it was not in the interest of justice to deprive the plaintiff of their choice of solicitor—the Regular output oversimplifies the answer and offers a legally shallow response. The Regular answer is well-organized and easy to follow, presenting the Court's decision. Still, it lacks completeness and fails to reflect how legal decisions are conventionally summarized or evaluated.

CoCoLex similarly identifies the dismissal of the motion and correctly names the parties and legal representatives, maintaining surface-level factual consistency with the context. However, its response exhibits redundancy without analytical gain, reiterating facts about representation while still failing to articulate why the court declined to restrain the firm. CoCoLex focuses on facts that are irrelevant to the given question, while still failing to articulate *why* the court dismissed the restraining order against the firm.

We present Case Study 2 in Table 15, sourced from the AQuECHR dataset, which explores whether a State violates the Convention when it adopts a procedure to recognise a child's legal relationship with an intended mother that cannot be implemented promptly or effectively, resulting in prolonged uncertainty and vulnerability for the child.

Regular decoding produces a structurally sound and fluent response that correctly identifies the procedural question and cites relevant paragraphs. However, it arrives at an overly definitive conclusion ("may not be held in violation") without fully engaging with the nuanced legal reasoning required by the question. While the response maintains reasonable faithfulness to the context by referencing the Court's definition of "private life," it misses the opportunity to synthesize the broader legal framework connecting procedural delay to potential Convention violations. The output is grammatically correct and well-organized but lacks the analytical depth expected in legal QA.

CoCoLex exhibits two concrete blind-copy failures that trace directly to its full-context datastore. First, it substitutes *"intended father"* for the query's *"intended mother"*. This is not a reasoning error but a copying artifact: paragraph 71 of the context discusses a precedent case involving the *intended father* (the biological father), and because CoCoLex's datastore covers all retrieved paragraphs indiscriminately, copying pressure draws that token from the wrong paragraph rather than preserving the term supplied by the query. Second, *"inchild-child relationship"* is a garbled fragment produced when the copy distribution simultaneously pulls from multiple context paragraphs with overlapping but inconsistent token sequences, yielding a token that exists in neither source. The absence of entropy gating compounds these errors: the copy mechanism fires at every decoding step, including high-confidence steps (e.g., generating citations) where it adds malformed reference strings instead of deferring to the model's own generation.

SACoD avoids both failure modes through its

Table 14: Case Study 1: Moore v Scenic Tours Pty Ltd (OALQA). Only SACoD captures the key legal reasoning that it was "not in the interest of justice to deprive the plaintiff of their choice of solicitor."

Question	In the case of Moore v Scenic Tours Pty Ltd [2015] NSWSC 237, what was the court's decision regarding the motion to restrain a firm from acting?
Context	Moore v Scenic Tours Pty Ltd [2015] NSWSC 237. Supreme Court, New South Wales. Decision: (a) Amended Notice of Motion dated 20 February 2015 is dismissed. (b) Defendant to pay the plaintiff's costs. Catchwords: PRACTICE AND PROCEDURE – civil – representative proceedings – whether court should exercise jurisdiction to restrain a firm from acting – plaintiff consents to firm continuing to act – plaintiff entitled to a solicitor of their choice – not in interest of justice to deprive plaintiff of their choice of solicitor – motion to restrain firm from acting dismissed. [. . .] On 29 July 2014, David Moore commenced representative proceedings pursuant to Part 10 of the Civil Procedure Act 2005 against Scenic Tours Pty Ltd. The plaintiff has retained and is represented by Benjamin Hemsworth, of Somerville Legal Pty Ltd. [. . .]
Reference	In the case of Moore v Scenic Tours Pty Ltd [2015] NSWSC 237, the court decided to dismiss the motion to restrain a firm from acting. The court found that the plaintiff was entitled to a solicitor of their choice and it was not in the interest of justice to deprive the plaintiff of their choice of solicitor.
Regular	The court's decision in the case of Moore v Scenic Tours Pty Ltd [2015] NSWSC 237 was to dismiss the motion to restrain the firm from acting. The plaintiff, David Moore, was allowed to continue to be represented by the firm, Somerville Legal Pty Ltd.
CoCoLex	The court's decision in the case of Moore v Scenic Tours Pty Ltd [2015] NSWSC 237 was to dismiss the motion to restrain the firm from acting. The plaintiff, David Moore, was represented by Benjamin Hemsworth of Somerville Legal Pty Ltd. The defendant had filed a Notice of Motion seeking to have the firm restrained from acting for the plaintiff as his legal advisers. However, the Amended Notice of Motion was dismissed with costs.
SACoD	The court dismissed the motion to restrain Somerville Legal Pty Ltd from acting for the plaintiff, David Moore, in the representative proceedings. The court's decision was made on the grounds that it was not in the interest of justice to deprive the plaintiff of their choice of solicitor. The plaintiff was represented by Benjamin Hemsworth of Somerville Legal Pty Ltd. The Amended Notice of Motion dated 20 February 2015 was dismissed, and the defendant was ordered to pay the plaintiff's costs.

two core mechanisms. The entropy gate suppresses copying at low-uncertainty steps, which is why it correctly preserves *"intended mother"* verbatim from the query: the model is already confident in that term and the gate routes those steps to vanilla decoding, preventing the context's *"intended father"* from being substituted. At the uncertain, evidence-dependent steps, the attention-based evidence selection identifies paragraph 64—which defines *"private life"* and establishes the *"legal parent-child relationship"* as an aspect of personal identity—as the operative evidence, directing the copy mechanism to that paragraph rather than paragraph 71. The result is a response that accurately reflects the query's subject, faithfully reproduces the applicable legal standard, and presents a coherent argument leading to a legally sound conclusion.

Algorithm 1 SACoD: Selective and Confidence-aware Decoding

Require: Model parameters Θ, context c, query x, update interval $K = 10$.

$\mathcal{H} \leftarrow \emptyset$ {Initialize entropy buffer}
$\text{tokens}_{\text{pfx}} \leftarrow \text{tokenize}(c \oplus x)$
for each position i in $\text{tokens}_{\text{pfx}}$ **do**
  Compute step entropy h_i^{pfx} using Eq. 8
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{h_i^{\text{pfx}}\}$
end for
$\theta \leftarrow$ `OtsuThreshold`$(\mathcal{H})$ {Obtain initial dynamic threshold}
$y_0 \leftarrow \emptyset$, $t \leftarrow 1$
while $y_{t-1} \neq$ EOS **do**
  Compute next-token probabilities $p_\Theta(y_t \mid c, x, y_{<t})$
  Compute current step entropy h_t using Eq. 8
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{h_t\}$
  if $h_t < \theta$ **then**
    `Gate` $\leftarrow$ `Vanilla`
    $y_t \sim p_\Theta(y_t \mid c, x, y_{<t})$
  else
    `Gate` $\leftarrow$ `Copy`
    Extract relevant evidence $\mathcal{P}_E^{(t)}$ using Eq. 2
    Sample y_t from interpolated distribution $p(y_t \mid x, c_E^{(t)}, y_{<t})$ (Eq. 4)
  end if
  if $t \bmod K = 0$ **then**
    $\theta \leftarrow$ `OtsuThreshold`$(\mathcal{H})$ {Periodically refine threshold}
  end if
  $t \leftarrow t + 1$
end while
return Generated sequence $\mathbf{y}$

Algorithm 2 Otsu's Method for Dynamic Entropy Thresholding (`OtsuThreshold`)

Require: Set of entropy values $\mathcal{H}$

$\sigma^2_{\text{best}} \leftarrow 0$
$\theta^* \leftarrow \text{median}(\mathcal{H})$ {Fallback threshold}
$\mathcal{U} \leftarrow$ Unique, sorted values of $\mathcal{H}$
for each candidate $\theta \in \mathcal{U}$ **do**
  $\mathcal{H}_0 \leftarrow \{h \in \mathcal{H} \mid h < \theta\}$ {Confident tokens}
  $\mathcal{H}_1 \leftarrow \{h \in \mathcal{H} \mid h \geq \theta\}$ {Uncertain tokens}
  if $|\mathcal{H}_0| = 0$ **or** $|\mathcal{H}_1| = 0$ **then**
    continue
  end if
  $w_0 \leftarrow |\mathcal{H}_0|/|\mathcal{H}|, \quad w_1 \leftarrow |\mathcal{H}_1|/|\mathcal{H}|$
  $\mu_0 \leftarrow \text{mean}(\mathcal{H}_0), \quad \mu_1 \leftarrow \text{mean}(\mathcal{H}_1)$
  $\sigma^2 \leftarrow w_0 \cdot w_1 \cdot (\mu_0 - \mu_1)^2$ {Inter-class variance}
  if $\sigma^2 > \sigma^2_{\text{best}}$ **then**
    $\sigma^2_{\text{best}} \leftarrow \sigma^2$
    $\theta^* \leftarrow \theta$
  end if
end for
return θ^*

Table 15: Case Study 2: Surrogacy and Parent-Child Recognition (AQuAECHR). CoCoLex produces garbled text (red) and excessive citations, while SACoD maintains coherent legal reasoning.

Question	In a case where a State has chosen a particular means to achieve recognition of the legal relationship between a child and an intended mother, can the State be held in violation of the Convention if the procedure laid down to establish recognition of the relationship is not capable of being implemented promptly and efficiently, causing prolonged uncertainty for the child and leaving the child in a vulnerable position?
Context	001-214296 71. In that connection, the present case must be clearly distinguished from Mennesson and Labassee. The Court reiterates that in the above-mentioned case it expressly held that a lack of possibility of recognition of the legal relationship between a child born via surrogacy abroad and the intended father, where he was the biological father, entailed a violation of the child's right to respect for his or her private life (see Mennesson , §§ 100-01, and Labassee , § 79, both cited above). In the present case it is true that the Polish authorities refused to give effect to the foreign birth certificates establishing the legal parent-child relationship between the applicants and their biological father, Mr S. However, this link is recognised in the country where the family resides. 001-214296 70. Furthermore, in the circumstances of this case the Court notes that the applicants can benefit, in the State where they live, from the legal parent - child relationship with their biological father where the recognition of that relationship is not put into doubt. Moreover, it cannot be said that the decisions of the Polish authorities left them in a legal vacuum both as to their citizenship and as to the recognition of the legal parent - child relationship with their biological father (see, conversely, Orlandi and Others v. Italy , nos. 26431/12 and 3 others, § 209, 14 December 2017). 001-214296 64. The Court notes that the concept of "private life" is a broad term not susceptible to exhaustive definition. It covers the physical and psychological integrity of a person. It can therefore embrace multiple aspects of the person's physical and social identity which includes the legal parent-child relationship...; an essential aspect of the identity of individuals is at stake where the legal parent-child relationship is concerned (see Mennesson v. France , no. 65192/11, §§ 46,96, ECHR 2014 (extracts), and Labassee v. France , no. 65941/11, §§ 38,75, 26 June 2014).
Reference	The Court has stressed that it is in the child's interests in such a situation for the uncertainty surrounding the legal relationship with his or her intended mother to be as short-lived as possible, noting that until that relationship is recognised, the child is in a vulnerable position. However, once the relationship between the child and the intended mother has become a practical reality, the procedure laid down to establish recognition of the relationship in domestic law must be capable of being implemented promptly and efficiently, in accordance with the child's best interests.
Regular	In a case where a State has chosen a particular means to achieve recognition of the legal relationship between a child and an intended mother, it may not be held in violation of the Convention if the procedure laid down to establish recognition of the relationship is not capable of being implemented promptly and efficiently, causing prolonged uncertainty for the child and leaving the child in a vulnerable position [001-214296, 70]. However, the Court has previously held that the concept of "private life" is a broad term not susceptible to exhaustive definition, and an essential aspect of the identity of individuals is at stake where the legal parent-child relationship is concerned [001-214296, 64]. Therefore, while the State may not be in violation of the Convention for the delay in implementing the procedure, it is important to consider the impact of such delay on the child's private life and well-being.
CoCoLex	In the context of the Convention, the Court has not expressly held that a State is in violation for choosing a particular means to achieve recognition of the legal relationship between a child and an intended father, but the prolonged uncertainty and vulnerable position of the child due to the inchild-child relationship being left in a legal vacuum is a matter of concern. (001-214296, §70). The Court notes that the legal parent-child relationship is an essential aspect of the identity of individuals (001-214296, §64, Mennesson, [. . .])
SACoD	In a case where a State has chosen a particular means to achieve recognition of the legal relationship between a child and an intended mother, it may not necessarily be held in violation of the Convention if the procedure laid down to establish recognition of the relationship is not capable of being implemented promptly and efficiently, causing prolonged uncertainty for the child and leaving the child in a vulnerable position. However, the Court has previously emphasized that the concept of "private life" is a broad term that covers multiple aspects of an individual's identity, including the legal parent-child relationship [001-214296, 64]. The Court has also recognized that an essential aspect of an individual's identity is at stake where the legal parent-child relationship is concerned [001-214296, 64]. Therefore, while the State's failure to implement the procedure promptly and efficiently may not directly violate the Convention, it could potentially lead to a violation of the child's right to respect for his or her private life if the prolonged uncertainty and vulnerability caused significant harm to the child's well-being and identity.