CoCoA: Confidence- and Context-Aware Adaptive Decoding for Resolving Knowledge Conflicts in Large Language Models

Anant Khandelwal, Manish Gupta, Puneet Agrawal
Microsoft, India
{anantk,gmanish,punagr}@microsoft.com

Abstract

Faithful generation in large language models (LLMs) is challenged by knowledge conflicts between parametric memory and external context. Existing contrastive decoding methods tuned specifically to handle conflict often lack adaptability and can degrade performance in low conflict settings. We introduce CoCoA (Confidence- and Context-Aware Adaptive Decoding), a novel token-level algorithm for principled conflict resolution and enhanced faithfulness. CoCoA resolves conflict by utilizing confidence-aware measures (entropy gap and contextual peakedness) and the generalized divergence between the parametric and contextual distributions. Crucially, CoCoA maintains strong performance even in low conflict settings. Extensive experiments across multiple LLMs on diverse Question Answering (QA), Summarization, and Long-Form Question Answering (LFQA) benchmarks demonstrate CoCoA's state-of-the-art performance over strong baselines like AdaCAD. It yields significant gains in QA accuracy, up to 9.2 points on average compared to the strong baseline AdaCAD, and improves factuality in summarization and LFQA by up to 2.5 points on average across key benchmarks. Additionally, it demonstrates superior sensitivity to conflict variations. CoCoA enables more informed, context-aware, and ultimately more faithful token generation.

1 Introduction

Large language models (LLMs) have achieved strong performance across Natural Language Processing (NLP) tasks such as question answering, summarization, and fact verification by leveraging vast parametric knowledge acquired during pretraining (Petroni et al., 2019; Roberts et al., 2020; Brown et al., 2020). However, this knowledge is static and limited by the training data, making it prone to becoming outdated or incomplete (Lazaridou et al., 2021; Sun et al., 2024; Hernandez et al., 2024). To address this, *context-aware generation* augments LLMs with auxiliary inputs (such as retrieved documents or tool outputs) at inference time, enabling incorporation of up-to-date, task-specific knowledge without retraining (Guu et al., 2020; Nakano et al., 2021; Schick et al., 2023). Yet, this introduces the risk of *knowledge conflict* between the external context $\boldsymbol{c}$ and the model's internal knowledge (Chen et al., 2022; Xie et al., 2024; Li et al., 2023a), especially when the two sources contradict each other.

Standard decoding methods over the context-conditioned distribution p_θ^{ctx} often fail to resolve such conflicts, defaulting to parametric priors even when contradicted by contextual evidence (Longpre et al., 2021; Zhou et al., 2023). This *model stubbornness* undermines performance in knowledge-sensitive settings (Fig. 1(top)). *Context-aware decoding* methods attempt to mitigate this by contrasting p_θ^{ctx} with the unconditional distribution p_θ, promoting tokens grounded in the context and penalizing those favored only by internal memory. For instance, **Context-aware Decoding (CAD)** (Shi et al., 2024) applies a fixed contrastive parameter α to reweight token probabilities. While CAD improves outcomes under strong conflict, its static nature can lead to *over-correction* when context and model agree, and *under-correction* in subtle conflicts, harming output quality in low-conflict cases (Wang et al., 2023).

AdaCAD (Wang et al., 2025) extends CAD by dynamically adjusting contrastive weights at each decoding step using the Jensen-Shannon Divergence (JSD) between p_θ^{ctx} and p_θ, removing the need for a fixed hyperparameter. However, it inherits two key limitations. First, JSD uniformly penalizes all distributional shifts, failing to distinguish meaningful context signals from noise. Second, it saturates on peaked or heavy-tailed distributions typical in autoregressive decoding, limiting sensitivity to subtle conflicts, especially early in

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6835–6855
November 4-9, 2025 ©2025 Association for Computational Linguistics

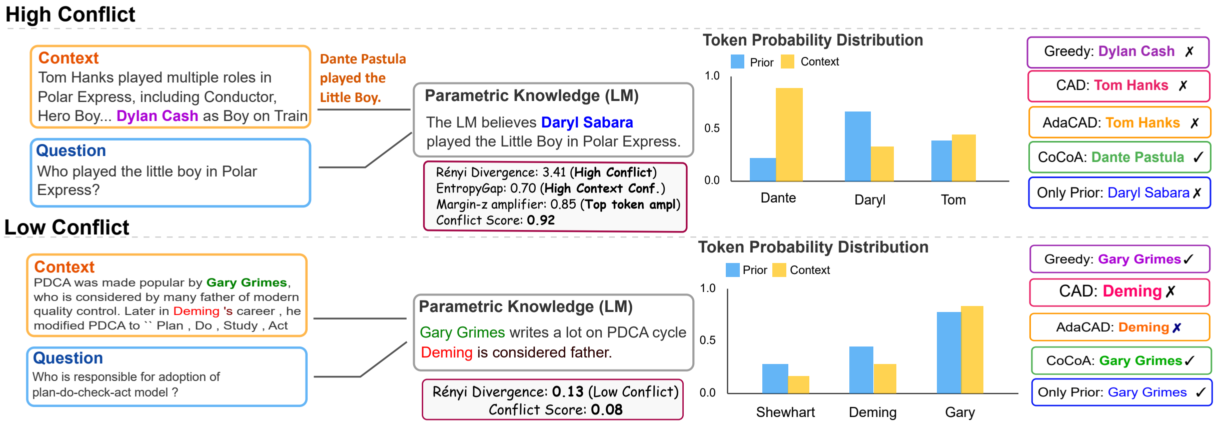

Figure 1: *Comparison of decoding methods on high- and low-conflict questions.* CoCoA *accurately resolves high-conflict answers by leveraging confidence-aware conflict resolution and preserves correct predictions in low-conflict cases, outperforming greedy, CAD, and AdaCAD.*

generation. AdaCAD mitigates this with heuristic "warm-up" inflation, introducing new hyperparameters and reducing adaptability.

To address these issues, we propose CoCoA (Confidence- and Context-Aware Adaptive Decoding), a principled framework (in Fig. 1) for resolving knowledge conflicts in context-aware generation at test time. Our key contributions are as follows: (a) **Tail-Sensitive Conflict Detection**: We replace JSD with Rényi divergence, enabling tunable sensitivity across entropy regimes. This enhances the model's ability to detect tail-heavy shifts between prior and contextual distributions, crucial for capturing subtle but meaningful conflicts. (b) **Contextual Confidence Estimation**: We introduce a novel confidence signal that combines the entropy gap between prior and context distributions with the contextual peakedness, capturing both divergence and certainty in a unified measure. (c) **Adaptive Gating Mechanism**: We propose a dynamic gating strategy that determines how much to trust the context at each step, leveraging a stable blend of conflict signals and contextual confidence. Experiments on diverse context-rich tasks (including QA, long-form generation, and fact verification) show that CoCoA achieves state-of-the-art accuracy compared to other test-time baselines while maintaining fluency and resolving conflicts effectively. Code: `https://github.com/infusion-zero-edit/CoCoA/`.

2 Related Work

Context-aware Decoding and Knowledge Integration. Recent studies tackle misalignment between retrieved context and parametric knowledge, especially in entity-centric QA (Longpre et al., 2021) and retrieval-augmented generation (Tan et al., 2024). While external knowledge can reduce hallucinations (Shuster et al., 2021), balancing it with model priors remains challenging. Some approaches rely on prompt engineering (Zhou et al., 2023) or train auxiliary discriminators (Zhang et al., 2023), whereas our method is training-free and prompt-agnostic.

Contrastive Decoding Methods. Contrastive decoding enhances contextual grounding (Shi et al., 2024), diversity (Li et al., 2016), and controllability (Liu et al., 2021). CAD (Shi et al., 2024) contrasts contextualized vs. unconditional outputs, while ConfCD (Zhao et al., 2024) adjusts generation confidence with noisy inputs. AdaCAD (Wang et al., 2025) introduces a continuous conflict spectrum using JSD, avoiding discrete binning.

Faithfulness in Long-Form QA (LFQA). LFQA systems often hallucinate beyond retrieved evidence (Fan et al., 2019; Han et al., 2024). Techniques like RECOMP (Xu et al., 2024a) and SelfRAG (Asai et al., 2024) aim to improve relevance via filtering or reflection during training. Contrastive methods (Li et al., 2023b; Shi et al., 2024; Wang et al., 2025) enhance context salience, but often rely on post-hoc steps. Our method improves contextual alignment during generation without extra decoding passes.

Knowledge Conflict in LLMs. Conflicts between contextual and parametric knowledge, i.e., context-memory conflicts (Chen et al., 2022; Xie et al., 2024) cause models to ignore retrieved facts. Xu et al. (2024b) categorize conflict types; we target dynamic resolution of context-memory conflict. Unlike revision-based approaches (Huang et al., 2025; Choi et al., 2023), our method integrates conflict signals directly into decoding for test-time correction.

3 Methodology

3.1 Task Setup and Notation

We consider autoregressive generation tasks (e.g., QA, instruction following), where a model generates a sequence $\boldsymbol{y} = (y_1, \ldots, y_T)$ given a query $\boldsymbol{x}$ and optional context $\boldsymbol{c}$. The model defines two token distributions: $p_\theta(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})$ (prior) and $p_\theta^{\text{ctx}}(y_t \mid \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t})$ (contextualized). Standard decoding treats context as static input, ignoring conflicts between p_θ and p_θ^{ctx}. Our goal is to resolve such conflicts during generation by amplifying trustworthy context while avoiding overreliance on either source.

3.2 Background: AdaCAD

AdaCAD (Wang et al., 2025) adapts contrastive decoding by replacing the fixed weight α in CAD (Shi et al., 2024) with a dynamic, stepwise signal. At each token t, it computes the JSD between the context-aware and prior distributions:

$$\alpha_t^{\text{JSD}} = \text{JSD}\left(p_\theta^{\text{ctx}} \,\|\, p_\theta\right) \tag{1}$$

and samples from a blended distribution:

$$\tilde{p}_\theta \propto p_\theta^{\text{ctx}} \cdot \left(\frac{p_\theta^{\text{ctx}}}{p_\theta}\right)^{\alpha_t^{\text{JSD}}} \tag{2}$$

This lets AdaCAD increase reliance on context when conflicts are high and defer to the model prior when they agree, offering token-level adaptability absent in CAD.

3.3 CoCoA: Confidence- and Context-aware Adaptive Decoding)

We propose CoCoA (Confidence- and Context-aware Adaptive Decoding), a token-level decoding algorithm that dynamically resolves conflicts between a model's parametric knowledge and external context. CoCoA resolves conflict by utilizing confidence-aware measures (entropy gap and contextual peakedness) and the generalized (Rényi) divergence between the prior and contextual distributions. Harnessing Rényi divergence and entropy gap in CoCoA enables it to better capture subtle shifts between the two distributions, especially in low-confidence regimes (Fig.1). At each decoding step t, CoCoA considers two distributions over the next token: the base model's prior $p_\theta(y_t)$ and the context-aware version $p_\theta^{\text{ctx}}(y_t)$. Rather than favoring one arbitrarily, CoCoA blends them via a conflict-aware weight $\lambda_t \in [0, 1]$, producing:

$$q(y_t) \;\propto\; p_\theta(y_t)^{1-\lambda_t} \cdot p_\theta^{\text{ctx}}(y_t)^{\lambda_t}, \tag{3}$$

which corresponds to logit interpolation:

$$\log q(y_t) = (1 - \lambda_t) \log p_\theta(y_t) + \lambda_t \log p_\theta^{\text{ctx}}(y_t), \tag{4}$$

followed by renormalization. The interpolation weight λ_t is dynamically computed based on the measured conflict between the two distributions, with Rényi divergence providing tunable sensitivity across entropy regimes.

3.3.1 Conflict Detection in CoCoA

To detect conflicts between the prior distribution $p_\theta(y_t)$ and the context-aware distribution $p_\theta^{\text{ctx}}(y_t)$, CoCoA leverages two complementary signals: Rényi divergence and entropy gap.

We first compute the *Rényi divergence* of order α, a generalization of KL divergence that is sensitive to discrepancies in the tail of distributions. For $\alpha \neq 1$, the divergence is defined as:

$$D_t^\alpha(p_\theta \| p_\theta^{\text{ctx}}) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{|V|} p_\theta(y_t^{(i)})^\alpha \, p_\theta^{\text{ctx}}(y_t^{(i)})^{1-\alpha} \tag{5}$$

where V is vocabulary. Choosing $\alpha < 1$ emphasizes low-probability events, allowing the model to surface sharp contextual shifts even when overall token distributions seem similar, unlike symmetric measures such as JSD. To further capture context-induced changes in uncertainty, we compute the *entropy gap*:

$$\Delta H_t = H(p_\theta(y_t)) - H(p_\theta^{\text{ctx}}(y_t)), \tag{6}$$

where entropy is given by:

$$H(P) = -\sum_y P(y) \log P(y). \tag{7}$$

A large positive ΔH_t indicates that context has increased certainty by concentrating probability mass, while a small or negative gap suggests minimal contextual influence. This helps differentiate between noisy and confident context, e.g., even when divergence is moderate, a high entropy gap signals that the context provides a strong directional cue. Together, Rényi divergence and the entropy gap allow CoCoA to detect not only distributional disagreement but also changes in certainty, enabling finer-grained adaptation to conflict during decoding.

3.3.2 Contextual Peakedness in CoCoA

To refine the model's sensitivity to confident contextual cues, we introduce a contextual peakedness measure based on the sharpness of the context distribution. Let m_t denote the margin between the top two token probabilities under the context-aware distribution:

$$m_t = p_\theta^{\text{ctx}}(y_t^{(1)}) - p_\theta^{\text{ctx}}(y_t^{(2)}), \tag{8}$$

where $y_t^{(1)}$ and $y_t^{(2)}$ are the highest-ranked and second-ranked candidates, respectively. A large margin suggests high certainty in the context's preferred token. This factor ensures that confident contextual signals have stronger influence during decoding.

3.3.3 Adaptive Gating in CoCoA

To determine how much to trust the context at each step, we compute a conflict score s_t based on divergence and entropy signals:

$$s_t = \sigma\left(D_t^\alpha + \gamma \Delta H_t + \delta\right), \tag{9}$$

where D_t^α is the Rényi divergence, ΔH_t is the entropy gap, σ is the sigmoid function, γ is the mixing weight between ΔH_t and D_t^α, and δ is a small constant for numerical stability. This score reflects the severity of knowledge conflict between the model and the context.

We then integrate the contextual peakedness m_t with the conflict score to form blending weight:

$$\lambda_t = \sigma\left(z \log m_t + \log \frac{1 - s_t}{s_t}\right), \quad \text{with } z > 1, \tag{10}$$

which sharpens the gating decision: confident context (large m_t) and high conflict (large s_t) push λ_t closer to 1; low conflict or weak context drive it toward 0. Finally, we compute the blended output distribution using power interpolation:

$$q(y_t) \propto p_\theta^{\text{ctx}}(y_t)^{\lambda_t} \cdot p_\theta(y_t)^{1-\lambda_t}, \tag{11}$$

The *logit-space normalization term* (unlike heuristic "warm-up" in CAD) stabilizes blending and preserves generation quality. This formulation allows CoCoA to adaptively modulate reliance on context versus prior, ensuring more trustworthy and calibrated generation across varying levels of knowledge conflict.

4 Experiments and Results

4.1 Experimental Setup

Datasets and Metrics. We evaluate our approach across a diverse set of benchmarks. For foundational question answering (QA), we use Natural Questions (NQ; (Kwiatkowski et al., 2019a)), TriviaQA (Joshi et al., 2017a), PopQA (Mallen et al., 2023a), and HotpotQA (Yang et al., 2018a). To assess robustness to conflicting information, we include NQ-SWAP (Longpre et al., 2021), a synthetic conflict variant of NQ. Structured reasoning is tested with the tabular QA dataset TabMWP (Lu et al., 2023a). Performance on these QA benchmarks is measured by exact match accuracy.

For long-form generation, we use CNN-DM (See et al., 2017a), XSum (Narayan et al., 2018a), and the topic-focused dialogue summarization dataset TofuEval (Tang et al., 2024b), which emphasizes marginal topics. Summarization quality on CNN-DM and XSum is evaluated via ROUGE-L (Lin, 2004) and BERT-P (Zhang et al., 2020). Since TofuEval lacks reference summaries, we measure factual consistency with AlignScore (Zha et al., 2023) for both main and marginal topics.

Our evaluation also covers diverse long-form QA (LFQA) datasets spanning various query types and complexities: CLAPNQ (Rosenthal et al., 2025a) (real web queries with gold documents and multi-source simulation), ExpertQA (Malaviya et al., 2024) (expert questions with verified answers), HAGRID (Kamalloo et al., 2023) (information-seeking questions with LLM-generated, evaluated answers), ELI5-WebGPT (Nakano et al., 2021) ("Explain Like I'm Five" style with human answers), and QuoteSum (Schuster et al., 2024) (semi-extractive answers from multiple sources).

Evaluation metrics are dataset-specific: CLAPNQ, ExpertQA, and HAGRID use ROUGE-L; ELI5-WebGPT employs claim recall between generated responses and gold sub-claims (Chen et al., 2023a); QuoteSum is evaluated with SEMQA (Schuster et al., 2024). Faithfulness (how well responses are grounded in the provided context) is measured by MiniCheck (Tang et al., 2024a), which scores consistency of statements against source documents. We report the average consistency score (FaithScore) per dataset. Additional details and examples are provided in Appendix A.

Source of Context. For all experiments, we use gold contexts provided by each dataset to ensure consistent and standardized evaluation. For NQ, NQ-SWAP, TriviaQA, HotpotQA, and PopQA, we adopt the gold contexts curated by the AdaCAD authors (Wang et al., 2025). In TabMWP, the accompanying semi-structured tables serve as contextual input.

For summarization tasks (CNN-DM, XSum, and TofuEval) the source documents are used as context, with the task-specific prompts serving as input queries. We expect context distributions to be much more confident compared to prior distributions here, and hence the gating parameter λ_t should increase with decoding steps. In the long-form QA setting, we use the gold contexts included with each dataset: curated web documents in ELI5-WebGPT (Nakano et al., 2021), expert-verified domain documents in ExpertQA (Malaviya et al., 2024), retrieved passages for attributed explanations in HAGRID (Kamalloo et al., 2023), selected passages tailored for long-form answers in CLAPNQ (Rosenthal et al., 2025a), and Wikipedia passages supporting semi-extractive answers in QuoteSum (Schuster et al., 2024). A full summary of input queries x and associated contexts c is presented in Table 9, with prompt details provided in Appendix B.

Models. We evaluate CoCoA using multiple pre-trained models, including LLaMA 2 (13B) (Touvron et al., 2023), LLaMA 3 (8B, 70B) (AI@Meta, 2024), and Mistral (7B) (Jiang et al., 2023), assessing performance on both base and instruction-tuned variants to ensure robustness across model types.

Baselines. We compare **CoCoA** against a range of established and recent test-time decoding methods all of which also take context as part of input. These include: (1) **Greedy decoding**, which selects tokens directly from the model's output without contextual adjustments; (2) **Context-aware decoding (CAD)** (Shi et al., 2024), which applies a fixed contrastive scaling factor α across all examples; (3) **COIECD** (Yuan et al., 2024), which classifies each step as conflicting or not using a threshold λ, switching between fixed decoding modes accordingly; and (4) **ConfCD** (Zhao et al., 2024), which dynamically sets α at each timestep based on confidence differences between context-aware and context-free predictions.

We also include **AdaCAD** (Wang et al., 2025), the strongest baseline method, which adaptively sets α_t using JSD between prior and context-aware distributions to model conflict dynamically. For CAD, we follow prior work with $\alpha = 1.0$ for QA and $\alpha = 0.5$ for summarization. For COIECD, we use $\lambda = 0.25$ and the same α values as CAD. ConfCD sets $\alpha = \max_{y'} p_\theta(y'|c, x, y_{<t})$ if this context-conditioned probability exceeds its context-free counterpart; otherwise, $\alpha = 1 - \max_{y'} p_\theta(y'|x, y_{<t})$.

CoCoA differs by computing adjustment factors via divergence-stabilized entropy ratios and contextual confidence estimation. Following extensive ablation, we fix hyperparameters to $\alpha = 0.5$ (Rényi divergence), $z = 5.0$ (contextual peakedness mixing weight), $\gamma = 1.0$ (entropy gap weight), and $\delta = 1e^{-8}$. All methods are evaluated using zero-shot greedy decoding to ensure fair comparison across QA and summarization tasks. We used the eval scripts obtained from Wang et al. (2025).

4.2 Main Results

	Decoding	NQ	NQ-SWAP	TriviaQA	PopQA	HotpotQA	TabMWP	Avg
Llama2-13B	Greedy	44.26	54.89	85.50	76.65	38.27	38.30	56.31
	CAD	37.91	80.35	71.40	76.83	31.92	19.30	52.95
	COIECD	44.60	59.84	87.00	81.05	42.81	38.80	59.02
	AdaCAD	46.73	67.84	85.40	78.79	37.83	37.50	59.02
	CoCoA	**49.49**	**80.36**	**89.00**	**87.29**	**43.04**	**44.00**	**65.53**
Llama3-8B	Greedy	44.63	47.81	85.70	80.51	51.42	52.20	60.38
	CAD	35.96	77.94	40.20	74.27	39.53	26.60	49.08
	COIECD	43.36	51.16	83.10	78.49	45.63	49.70	58.57
	AdaCAD	45.47	62.34	82.50	81.34	50.53	53.00	62.53
	CoCoA	**49.30**	**79.15**	**90.40**	**93.76**	**51.01**	**62.80**	**71.74**
Llama3-70B	Greedy	44.13	55.74	90.20	86.10	56.11	66.70	66.50
	CAD	34.05	81.32	54.60	75.16	40.86	48.60	55.77
	COIECD	45.09	57.26	88.60	83.60	52.03	64.40	65.16
	AdaCAD	45.43	70.07	88.80	85.68	55.00	67.10	68.68
	CoCoA	**51.80**	**88.32**	**93.00**	**95.90**	**59.29**	**78.90**	**77.87**
Mistral-7B	Greedy	42.56	56.86	80.40	67.56	40.89	38.90	57.65
	CAD	20.98	66.89	24.20	48.54	18.49	20.10	35.82
	COIECD	29.00	58.09	71.60	64.59	35.83	31.60	48.45
	AdaCAD	45.09	67.27	80.20	67.26	41.35	39.70	60.23
	CoCoA	**48.00**	**80.90**	**87.70**	**76.83**	**45.93**	**47.10**	**64.91**

Table 1: Performance metrics for different models and decoding strategies. CoCoA shows improvements over previous methods across all datasets. More details in Appendix C.

QA Tasks. From Table 1, we observe that **CoCoA** consistently outperforms all baselines (greedy decoding, CAD, COIECD, and the strong AdaCAD) across all QA datasets and model scales. On **Llama3-70B**, CoCoA yields an average absolute gain of **11.37 pts** over greedy decoding and **9.19 pts** over AdaCAD, reflecting its robustness across both high- and low-conflict scenarios.

While CAD often suffers from degraded performance in low-conflict contexts, e.g., an average drop of **11.3 pts** across tasks on Llama3-8B, CoCoA maintains high accuracy even on mixed or minimal-conflict datasets such as NQ, TriviaQA, and PopQA. Notably, CoCoA surpasses AdaCAD by **7.9 pts** on **TriviaQA** and by a substantial **12.42 pts** on **PopQA** with Llama3-8B.

On **NQ-SWAP**, a high-conflict dataset, CoCoA demonstrates its ability to dynamically leverage

Decoding	CNN-DM			XSum			TofuEval (AlignScore)		
	ROUGE-L	BERT-P	AlignScore	ROUGE-L	BERT-P	AlignScore	Overall	Main	Marginal
Greedy	24.93	95.41	91.44	14.36	94.05	85.28	76.66	81.64	61.19
CAD	24.76	94.45	91.01	14.59	93.65	84.34	83.23	87.26	73.58
COIECD	23.47	92.06	85.49	13.65	91.04	73.81	60.86	68.06	58.31
ADACAD	25.42	94.91	94.97	14.91	94.29	85.81	85.07	88.06	**75.79**
COCOA	**25.68**	**95.42**	**95.70**	**15.06**	**94.60**	**87.94**	**86.32**	**89.14**	75.51

Table 2: Summarization performance on CNN-DM, XSum and TofuEval with our best performing model (Llama3-70B). COCOA consistently delivers the highest alignment (AlignScore) and strong ROUGE-L and BERT-P results, outperforming both contrastive and adaptive baselines. Full results across all models are in Table 11 in Appendix D.

Model	Decoding	CLAPNQ		ExpertQA		HAGRID		ELI5-WebGPT		QuoteSum		Avg. Faith.
		ROUGE-L	Faith	ROUGE-L	Faith	ROUGE-L	Faith	ROUGE-L	Faith	SEMQA	Faith	
GPT-4o	Greedy	40.53	91.81	46.34	69.48	57.76	90.86	59.04	81.00	42.56	78.51	82.33
GPT-4o-mini	Greedy	37.72	90.35	45.30	66.53	54.87	87.94	56.09	81.89	40.74	78.16	80.97
Llama-3.1-70B-Instruct	Greedy	39.44	88.64	43.02	69.35	49.21	79.08	51.66	74.87	41.24	67.42	75.87
	CAD	38.56	89.75	42.55	70.19	48.15	80.32	50.24	75.25	40.85	68.56	76.41
	ADACAD	41.00	91.32	45.03	71.12	50.04	81.57	52.12	77.11	43.06	70.23	78.27
	COCOA	**42.15**	**92.45**	**46.10**	**72.40**	**52.07**	**82.20**	**54.22**	**78.95**	**44.50**	**71.44**	**79.49**
Mistral-NeMo-12B-Instruct	Greedy	35.28	78.71	42.76	54.19	53.05	80.16	53.84	65.06	39.50	69.85	69.59
	CAD	34.23	79.20	41.81	55.42	51.78	79.55	52.90	66.10	38.90	69.30	69.91
	ADACAD	37.50	81.89	44.10	58.71	55.44	82.22	56.76	69.01	40.90	72.19	72.40
	COCOA	**39.75**	**84.12**	**46.27**	**61.20**	**58.56**	**83.30**	**59.89**	**71.42**	**42.57**	**74.55**	**74.92**
Llama-3.1-8B-Instruct	Greedy	17.14	58.47	31.67	51.22	16.47	55.80	47.11	55.74	25.96	41.70	52.59
	CAD	16.23	60.24	30.58	53.47	15.89	58.20	46.15	57.35	24.30	42.30	54.71
	ADACAD	18.43	62.37	32.87	54.76	17.12	59.90	48.22	58.71	26.80	43.57	55.86
	COCOA	**19.50**	**64.18**	**34.98**	**57.24**	**18.40**	**61.50**	**50.33**	**60.24**	**28.22**	**45.53**	**57.74**

Table 3: Performance on LFQA datasets showing ROUGE-L (RL) and faithfulness (Faith) scores for Greedy, CAD, ADACAD, and COCOA methods. For QuoteSum, we report SEMQA rather than (RL). **COCOA** consistently outperforming others on average factuality. More results in Appendix E.

context-sensitive contrast, achieving **88.32** with Llama3-70B, well above ADACAD (**70.07**) and greedy decoding (**55.74**). Similarly, on the semi-structured **TabMWP** dataset, COCOA reaches **78.90** with Llama3-70B, outperforming ADACAD by **11.8** pts, indicating strong generalization beyond standard QA formats. Across all models, COCOA consistently achieves the highest average accuracy. These results confirm COCOA's effectiveness in balancing contextual information and model confidence across diverse QA conditions.

Summarization Tasks. Table 2 evaluates COCOA on long-form summarization benchmarks (CNN-DM, XSum, and TofuEval) demonstrating consistent improvements over all baselines, including ADACAD. COCOA achieves the highest scores across all metrics, highlighting its ability to enhance both surface-level fluency and deeper factual consistency. On **TofuEval**, a benchmark targeting factual alignment under topic shifts, COCOA attains the highest AlignScore of **86.32**, surpassing greedy decoding by **9.66 pts**, CAD by **3.09 pts**, COIECD by **25.46 pts**, and ADACAD by **1.25 pts**. Notably, on the more challenging *marginal-topic* subset, COCOA scores **75.51**, outperforming all other methods (except ADACAD), demonstrating its strength in preserving factuality even in less salient content regions where prior methods often drift or hallucinate.

On **CNN-DM**, COCOA leads with a ROUGE-L of **25.68**, BERT-P of **95.42**, and AlignScore of **95.70**, showing notable improvements over both COIECD and ADACAD. On the more abstractive **XSum**, it yields modest gains in ROUGE-L (**15.06**) and BERT-P (**94.60**), but a more pronounced improvement in AlignScore (**87.94**, +2.13 pts over ADACAD), indicating stronger factual grounding under abstraction. Together, these results affirm COCOA's advantage in generating context-faithful, coherent summaries, especially under factuality-sensitive or out-of-distribution prompts where traditional contrastive methods struggle.

LFQA Tasks. Table 3 presents COCOA's performance across five long-form QA datasets, comparing it with two closed-source baselines (GPT-4o and GPT-4o-mini (both under greedy decoding)) and three open models (Llama-3.1-70B-Inst., Mistral-NeMo-12B-Inst., and Llama-3.1-8B-Inst.) using CAD, AdaCAD, and COCOA. Despite operating on open models, COCOA achieves results competitive with, and in some cases surpassing, state-of-the-art closed systems. For instance, on Llama-3.1-70B with CLAPNQ, COCOA attains a ROUGE-L of **42.15**, outperforming GPT-

Model Variant	NQ	NQ-SWAP
Full CoCoA	**49.30**	**79.15**
w/o Rényi w/ KL Divergence	48.20	76.69
w/o Rényi Divergence	47.10	70.80
w/o Entropy Gap	46.85	68.90
w/o Contextual peakedness	45.60	65.20
w/o Adaptive Gating (λ=0.5)	44.75	62.50
Greedy Decoding (Baseline)	44.63	47.81

Table 4: Exact Match (EM) accuracy of Llama3-8B on Natural Questions (NQ) and NQ-SWAP under different ablations of the CoCoA decoding framework. Each variant removes or alters a core component to assess its individual contribution to performance.

4o (40.53), and reaches a FaithScore of **92.45**, slightly above GPT-4o's 91.81. Against GPT-4o-mini on CLAPNQ, CoCoA shows even clearer gains (+4.43 ROUGE-L, +2.10 FaithScore), highlighting its ability to elevate smaller open models to the performance tier of much larger proprietary systems. Appendix G shows a qualitative example of improved response generation by CoCoA.

Compared to previous open decoding approaches, CoCoA consistently improves both relevance and factuality. On Llama-3.1-70B, it raises the average ROUGE-L across datasets from 47.05 (AdaCAD) to **48.64** and the FaithScore from 78.27 to **79.49**. Mistral-NeMo-12B exhibits similar trends: CoCoA improves ROUGE-L by +2.67 pts and FaithScore by +2.11 pts over AdaCAD on average all datasets. Even with the smaller Llama-3.1-8B model, CoCoA delivers gains of +1.64 pts ROUGE-L and +1.88 pts FaithScore. These improvements are consistent across all datasets (CLAPNQ, ExpertQA, HAGRID, ELI5-WebGPT, and QuoteSum), reinforcing CoCoA's effectiveness in balancing parametric knowledge and retrieved context for grounded, coherent generation.

Although, CoCoA based decoding from Llama-3.1-70B-Instruct beats all other open source models, it still performs slightly worse on average compared to GPT-40-mini and GPT-4o.

4.3 Ablation Study

As shown in Tab. 4, (1) Rényi Divergence is crucial for detecting nuanced conflicts between the model's knowledge and the context. (2) Entropy Gap provides valuable insights into the confidence level introduced by the context. (3) Contextual peakedness effectively strengthens the model's reliance on highly confident contextual information. (4) Adaptive Gating ensures a balanced integration of the model's prior knowledge and the context, adapting to the specific needs of each decoding step. More details in Appendix F.

Decoding	NQ-SWAP	NQ-SYNTH	Overall
Greedy	51.60	88.20	69.90
CAD	79.60	64.00	71.80
COIECD	50.80	83.60	67.20
AdaCAD	62.80	86.40	74.60
CoCoA	**80.80**	**86.60**	**83.70**

Table 5: Accuracy on conflicting data (NQ-SWAP) and non-conflicting data (NQ-SYNTH) with Llama3-70B.

5 Further Analysis

5.1 High vs. Low Conflict Instances

Setup. To evaluate decoding strategies across varying levels of conflict between context and a model's internal knowledge, we employed **NQ-SYNTH** (non-conflicting) and **NQ-SWAP** (highly conflicting) datasets. These datasets (totaling to 500 samples), established in prior work (Wang et al., 2025), isolate conflict as the primary challenge, allowing us to assess whether decoding methods like CoCoA can generalize effectively across these distinct regimes.

Result. As shown in Table 5, CoCoA demonstrates a significant advantage over prior decoding methods on both high-conflict (NQ-SWAP) and low-conflict (NQ-SYNTH) examples. Unlike CAD, which showed a notable performance drop of 24.2 pts on NQ-SYNTH compared to greedy decoding, CoCoA maintained near-optimal accuracy on non-conflicting data (86.6%) relative to greedy (88.2%). Crucially, CoCoA achieved the highest accuracy on NQ-SWAP (80.8%), surpassing AdaCAD by 18.0 pts, COIECD by 30.0 pts, and greedy decoding by 29.2 pts. This indicates that CoCoA effectively preserves fidelity in low-conflict scenarios without unduly penalizing agreement, a common issue with methods like CAD and COIECD. While AdaCAD aimed for adaptability across conflict levels, CoCoA leverages divergence-based normalization and entropy-aware amplification for a more granular, token-level calibration. This fine-tuning is particularly beneficial in long-context QA with mixed evidence. Consequently, CoCoA offers a robust decoding policy that generalizes effectively across both adversarial and naturally-aligned inputs, yielding an absolute overall accuracy gain of 9.1 pts over AdaCAD.

5.2 JSD does not adequately address conflict

While AdaCAD improves over CAD via a JSD-based gating mechanism, it remains limited in detecting fine-grained or evolving conflicts between

Context: Tom Hanks played multiple roles in The Polar Express, providing motion-capture performances for the Hero Boy, the Hero Boy's father, the Conductor, the Hobo, Santa Claus, and the Narrator. Daryl Sabara as the Hero Boy. (. . .). The voice of the Hero Boy (. . .) Dylan Cash as Boy on Train (voice) (. . .) Dante Pastula played the Little Boy. A group of (. . .) motion-capture. **Question**: Who played the little boy in polar express ? **Gold Answer**: Dante Pastula | **Parametric Knowledge (Llama3-70B)**: Daryl Sabara. **Greedy**: Dylan Cash | **AdaCAD**: Tom Hanks | **CoCoA**: Dante Pastula
Context (Meeting Transcripts): Speaker 0: Agenda Item One: Council Bill 119945, concerning the Civil Service Commissions. This bill amends election procedures for the employee-elected members of both the Civil Service Commission and the Public Safety Civil Service Commission, and includes temporary provisions for the 2020 elections. It amends Sections 4.04.250 and 4.08.040 of the Seattle Municipal Code, and ratifies and confirms certain prior actions. Thank you, Madam Chair. I move to pass Council Bill 119945. Is there a second? Speaker 1: Second. (. . .) Both commissions are holding elections this year, but due to COVID-19-related delaysIn addition, the bill updates outdated language in the Seattle Municipal Code and empowers the City Clerk to modernize election procedures through future Speaker 0: Nine in favor, none opposed. The bill passes. (. . .) **Greedy Decoding**: The bill amends election procedures for the Civil Service Commission and Public Safety Civil Service Commission, allowing the city clerk to hold elections in December instead of November, and permitting electronic ballots. The changes are temporary and only apply to the 2020 elections. **AdaCAD**: This bill amends Seattle Municipal Code (SMC) 4.04.250 and 4.08.040 to allow the City Clerk to hold elections for the Civil Service Commission and Public Safety Civil Service Commission in December 2020 rather than November 2020. The bill also allows for electronic ballots rather than only physical ballots, though the option of paper ballots will still remain, and the schedule changes only for 2020. And these elements are only included in the ordinance, not the **CoCoA**: The bill updates election procedures for employee-elected members of the Civil Service and Public Safety Civil Service Commissions, allowing electronic ballots and rescheduling 2020 elections to December due to COVID-19, while also removing outdated code language and enabling future modernization by the City Clerk.

Table 6: *Qualitative examples showcasing* **CoCoA***'s improved factual grounding and conflict resolution.* In the NQ example (top), CoCoA predicts the correct answer (green) by aligning contextual and parametric signals (blue), unlike greedy and AdaCAD (red). In the TofuEval case (bottom), CoCoA avoids hallucinations (red) and remains faithful to the source (green).

model priors and external context. To assess conflict sensitivity, we compute the Spearman correlation ρ between each method's output distribution and that of greedy decoding across NQ-SWAP (high conflict) and NQ-SYNTH (low conflict). An ideal context-aware decoder should show low ρ in high-conflict and high ρ in low-conflict settings. As shown in Tab. 7, AdaCAD exhibits a narrow correlation gap ($|\Delta\rho| = 0.08$), indicating limited responsiveness to conflict.

CoCoA achieves a substantially larger gap ($|\Delta\rho| = 0.21$), driven by three enhancements: (1) Rényi divergence, which better captures low-probability, contextually relevant alternatives; (2) an entropy gap to model uncertainty shifts; and (3) contextual peakedness to emphasize strong contextual cues. Fig. 2 illustrates CoCoA's more dynamic conflict signal λ_t, which evolves with context and shows greater sensitivity than the plateaued JSD trace in AdaCAD. This enables CoCoA to adaptively modulate generation in response to context-model divergence.

5.3 Qualitative Analysis and Case Study

We provide representative examples from NQ and TofuEval in Tab. 6 to illustrate CoCoA's ability to handle both QA and summarization tasks under knowledge conflict. In the NQ example, greedy decoding and AdaCAD both fail due to incorrect parametric priors, producing "Dylan Cash" and "Tom Hanks" respectively, while CoCoA successfully recovers the correct answer "Dante Pastula" by down-weighting conflicting parametric knowledge and emphasizing grounded context. In the summarization case, CoCoA better captures the COVID-related election delays and modernization efforts, faithfully aligning with the highlighted source sentences. In contrast, AdaCAD introduces unrelated constraints and Greedy omits key updates. These examples underscore CoCoA's advantage in adaptively reconciling model knowledge and context evidence, especially under fine-grained or subtle conflicts.

Decoding	ρ (NQ-SWAP)	ρ (NQ-SYNTH)	$\|\Delta\rho\|$
CAD	0.56	0.57	0.01
AdaCAD	0.86	0.94	0.08
CoCoA	0.74	0.95	**0.21**

Table 7: Spearman rank-order correlation coefficient between original and adjusted output distributions for CAD, AdaCAD and CoCoA on NQ-SWAP and NQ-SYNTH. The difference $|\Delta\rho|$ measures the sensitivity of a decoding method to the degree of conflict (higher is better).

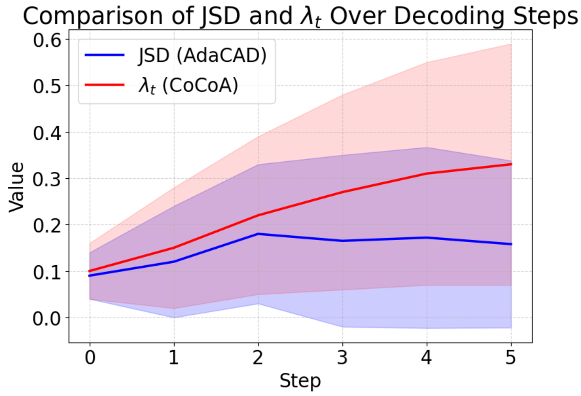

Figure 2: Comparison of JSD values (used in AdaCAD) and λ_t values (used in CoCoA) over the first 5 decoding steps (t) using LLaMA3-70B on TofuEval.

6 Discussion and Conclusion

CoCoA advances decoding by finely detecting and adapting to varying degrees of knowledge conflict between model memory and context. Its entropy-based, token-level conflict measures allow dynamic blending of parametric and contextual signals, avoiding the pitfalls of static or coarse conflict handling. This results in more faithful and accurate generation across QA, summarization, and long-form tasks, without degrading performance when conflict is low. CoCoA's principled approach offers a robust, flexible framework for improving context-aware language generation and sets a strong foundation for future dynamic decoding research.

Limitations

CoCoA relies on fine-grained access to a model's token-level probability distributions, with and without retrieved context, to compute Rényi divergences, entropy gaps, and contextual peakedness. This requirement poses a barrier when working with fully black-box APIs (e.g., GPT-4), which typically expose only sampled text rather than the underlying logits or softmax scores. Developing techniques to approximate or infer these distributions without direct logit access would be a valuable extension, enabling broader applicability to proprietary or mobile-only LLM services.

Our current study is also confined to English-language benchmarks and a handful of widely-used open-source models. As large-scale models emerge in other languages and specialized domains (e.g., legal, medical, or code generation), it will be important to validate CoCoA's conflict-detection and adaptation mechanisms under different linguistic characteristics and domain-specific knowledge structures. Additionally, investigating how model size, instruction-tuning, and alignment procedures interact with CoCoA's entropy- and margin-based signals could reveal further refinements or simplifications.

Finally, while we do not identify any direct ethical or safety concerns with contrastive, context-aware decoding itself, future work should examine potential biases in context selection (retrieval quality) and ensure that CoCoA's stronger reliance on external knowledge does not inadvertently amplify misleading or harmful content.

References

AI@Meta. 2024. Llama 3 model card.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2023a. Understanding retrieval augmentation for long-form question answering. *Preprint*, arXiv:2310.12150.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yujia Chen and 1 others. 2023b. Hagrid: A human-llm collaborative dataset for generative retrieval-augmented question answering. *arXiv preprint arXiv:2307.16883*.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14053, Singapore. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA. Association for Computational Linguistics.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. *Preprint*, arXiv:2304.00740.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting

Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017a. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017b. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv:2307.16883*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019a. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019b. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: assessing temporal generalization in neural language models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. 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, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Shayne Longpre, Pan Lu, and Hal Daumé III. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9578–9589.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023a. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *ICLR*.

Pan Lu, Yichong Zhang, Xinyu Liu, and William Yang Wang. 2023b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *Proceedings of the 2023 International Conference on Learning Representations*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Chaitanya Malaviya and 1 others. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Alex Mallen, Zexuan Zhang, Xinyu Liu, Yichong Wu, Yiming Zhang, and William Yang Wang. 2023b. When not to trust language models: Investigating effectiveness of parametric and retrieval-based knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1234–1245.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Long Ouyang Wu, Christopher Kim, Christopher Hesse, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025a. CLAPnq: Cohesive long-form answers from passages in natural questions for RAG systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025b. Clapnq: Cohesive long-form answers from passages in natural questions for rag systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessí, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tal Schuster, Adam Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William Cohen, and Donald Metzler. 2024. SEMQA: Semi-extractive multi-source question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1363–1381, Mexico City, Mexico. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017a. 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.

Abigail See, Peter J Liu, and Christopher D Manning. 2017b. 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.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. 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.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of* the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. Minicheck: Efficient fact-checking of llms on grounding documents. *Preprint*, arXiv:2404.10774.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Liyan Tang and 1 others. 2024c. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 251–263.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. 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.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024a. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018b. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore. Association for Computational Linguistics.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

Overview of Appendices

- Appendix A: Dataset Details.
- Appendix B: Prompts
- Appendix C: Performance Comparison of Instruction-Tuned Models on QA Benchmarks.
- Appendix D: Performance Comparison on Summarization Tasks
- Appendix E: Performance Comparison on Long-Form QA (LFQA) Datasets
- Appendix F: Ablation Study
- Appendix G: Case Study
- Appendix H: Sensitivity Evaluation
- Appendix I: Latency Considerations

A Dataset Details

We evaluated our approach on a diverse set of question answering (QA) and summarization datasets, adhering to the experimental setup established by AdaCAD (Wang et al., 2025) to enable apples-to-apples comparison. Additionally, we benchmarked on long-form question answering (LFQA) datasets to evaluate faithfulness, defined as the extent to which the model's response is factually grounded in the provided context document. We also present one example from each dataset, as detailed in Table 9. For the synthetically generated QA datasets NQ-SWAP and NQ-SYNTH, we provide examples in Table 8.

Question: Who wrote we're going on a bear hunt ?
NATURAL QUESTION
Original Context: We 're Going on a Bear Hunt is a 1989 children 's picture book written by Michael Rosen and illustrated by Helen Oxenbury .. . . **Original Answer:** Michael Rosen
NQ-SWAP
Substitute Context: We 're Going on a Bear Hunt is a 1989 children 's picture book written by Robert Hooke and illustrated by Helen Oxenbury . . . **Substitute Answer:** Robert Hooke
NQ-SYNTH
Substitute Context: **Original Context:** We 're Going on a Bear Hunt is a 1989 children 's picture book written by Brian Urlacher and illustrated by Helen Oxenbury .. . . **Substitute Answer (generated from LLM):** Brian Urlacher

Table 8: Example from NQ-SWAP and NQ-SYNTH. A *substitute example* for NQ-SWAP is made from the *original example* by replacing the original answer, Michael Rosen, with a similar but conflicting answer, i.e., Robert Hooke. A *substitute example* for NQ-SYNTH is made from the *original example* by replacing the original answer, Michael Rosen, with one generated by Llama3-70B without context, i.e., Brian Urlacher.

A.1 Question Answering Datasets

Natural Questions (NQ) (Kwiatkowski et al., 2019b) A large-scale QA dataset comprising real anonymized queries from Google Search, each paired with a Wikipedia page. Annotators provide long and short answers based on the content of the page. Following AdaCAD (Wang et al., 2025), We utilize a subset of 3,231 validation instances featuring short answers .

NQ-SWAP (Longpre et al., 2021) A synthetic variant of NQ designed to introduce knowledge conflicts by replacing named entities in the context with alternate entities, challenging models to handle conflicting information. Following AdaCAD (Wang et al., 2025), this dataset contains 4,000 instances derived from NQ.

TriviaQA (Joshi et al., 2017b) A challenging reading comprehension dataset with over 650K question-answer-evidence triples. Questions are authored by trivia enthusiasts and paired with evidence documents gathered independently. Following AdaCAD (Wang et al., 2025) we perform evaluation on their sampled 1,000 instances from the Wikipedia domain for evaluation.

PopQA (Mallen et al., 2023b) An open-domain QA dataset consisting of 14,000 question-answer pairs focused on long-tail entities. Each instance includes fine-grained Wikidata entity IDs and relationship type information. We perform benchmarking on the selected 1,600 instances as used in AdaCAD (Wang et al., 2025) where the context contains the gold answer .

HotpotQA (Yang et al., 2018b) A multi-hop QA dataset requiring reasoning over multiple supporting documents. It includes 113,000 question-answer pairs with sentence-level supporting facts to facilitate explainable QA systems. We use the full development set comprising 7,405 instances.

TabMWP (Lu et al., 2023b) A dataset containing 38,431 tabular math word problems that require mathematical reasoning on both textual and tabular data. Each question is aligned with a tabular context presented as an image, semi-structured text, or a structured table. We utilize the "test1k" subset, which includes 1,000 instances .

A.2 Summarization Datasets

CNN/DailyMail (CNN-DM) (See et al., 2017b) An English-language dataset containing over 300,000 news articles from CNN and the Daily Mail, paired with multi-sentence summaries. It supports both extractive and abstractive summarization tasks. We use the same sampled (as obtained from AdaCAD (Wang et al., 2025) authors) set of 500 examples from the test set.

XSum The Extreme Summarization dataset (Narayan et al., 2018b) comprises 226,711 BBC news articles, each accompanied by a one-sentence summary. The dataset covers a wide range of topics and is designed for evaluating abstractive summarization systems. We perform benchmarking on 500 instances as obtained from AdaCAD (Wang et al., 2025) authors.

TofuEval (Tang et al., 2024c) A benchmark for evaluating the factual consistency and topic relevance of summaries, particularly in dialogue or meeting transcription scenarios. It includes 50 documents each from MediaSum and MeetingBank datasets, with three topics per document, resulting in 300 topic-focused summaries.

A.3 Long-Form Question Answering (LFQA) Datasets

ELI5-WebGPT (Nakano et al., 2021) ELI5-WebGPT is a dataset designed for evaluating long-form question answering (LFQA) systems. It comprises 271 questions sourced from the "Explain Like I'm Five" (ELI5) subreddit, as released by WebGPT (Nakano et al., 2021). Each question is paired with human-labeled answers and "gold" documents—relevant and informative passages collected by human annotators using commercial search engines like Bing. These documents serve as high-quality evidence to assess the performance of retrieval-augmented models in generating accurate and informative responses.

ExpertQA (Malaviya et al., 2023) ExpertQA consists of 528 open-ended, information-seeking questions across 32 topics, each paired with relevant documents and expert-verified answers.

HAGRID (Chen et al., 2023b) HAGRID (Human-in-the-loop Attributable Generative Retrieval for Information-seeking Dataset) is a dataset designed for generative information-seeking tasks. It comprises questions paired with manually labeled relevant documents and answers generated by GPT-3.5. These answers are formatted with in-context citations referencing the supporting documents. Human annotators evaluate each answer based on two criteria: informativeness (whether the answer provides a direct response to the question) and attributability (whether the answer's claims are supported by the cited documents). We selected 496 samples where the answers are considered both informative and well-grounded.

CLAPNQ (Rosenthal et al., 2025b).CLAPNQ (Cohesive Long-form Answers from Passages in Natural Questions) is a benchmark dataset designed to evaluate Retrieval-Augmented Generation (RAG) systems. It comprises 4,946 real web search queries sourced from the Natural Questions dataset, each paired with a single gold passage from Wikipedia and a concise, cohesive long-form answer. These answers are typically 2–3 sentences long and are crafted by integrating non-contiguous parts of the passage to ensure fluency and factual grounding. CLAPNQ supports comprehensive evaluation of RAG systems across retrieval, generation, and full pipeline tasks. For our evaluation, we selected the 300 answerable questions from the CLAPNQ test set, which consists of 600 questions in total (300 answerable and 300 unanswerable). This subset allows for focused assessment of model performance on questions with available grounded answers.

QuoteSum (Schuster et al., 2024) QuoteSum is a semi-extractive long-form question answering (LFQA) dataset designed to assess models' abilities to generate grounded, multi-source answers.

Each question is accompanied by relevant documents and human-written answers that explicitly incorporate extracted spans from multiple sources. These answers blend verbatim quotes with connective text to form cohesive, well-grounded responses. The dataset emphasizes the Semi-Extractive Multi-Source Question Answering (SEMQA) task, challenging models to synthesize information while maintaining precise attributions. The test subset of QuoteSum comprises 1,319 examples, each consisting of a question, a set of relevant documents, and a human-written answer. This subset is used to assess the performance of retrieval-augmented generation (RAG) systems in generating accurate and informative answers multiple sources.

A.4 Licenses

The datasets employed in our study are distributed under the following licenses:

- **Natural Questions (NQ)**: Apache License 2.0
- **NQ-SWAP**: MIT License
- **TriviaQA**: Apache License 2.0
- **PopQA**: MIT License
- **HotpotQA**: Apache License 2.0
- **TabMWP**: MIT License
- **CNN/DailyMail (CNN-DM)**: Apache License 2.0
- **XSum**: MIT License
- **TofuEval**: MIT License
- **ELI5**: Creative Commons Attribution-ShareAlike 4.0 International License
- **ExpertQA**: MIT License
- **HaGRID**: Apache License 2.0
- **CLAPNQ**: Apache License 2.0
- **QuoteSum**: Creative Commons Attribution-ShareAlike 4.0 International License

The models utilized in our experiments are governed by the following licenses:

- **LLaMA 2**: Custom license available at https://ai.meta.com/llama/license/
- **LLaMA 3**: Custom license available at https://www.llama.com/llama3/license/
- **Mistral**: Apache License 2.0

B Prompts

The prompts used on pre-trained base language model with and without context for QA, LFQA and summarization tasks are given in Figure 3.

C Performance Comparison of Instruction-Tuned Models on QA Benchmarks

We evaluate the performance of various instruction-tuned language models on multiple QA datasets using four decoding strategies: Greedy, CAD, AdaCAD, and our proposed CoCoA method. The results, as presented in Table 10, demonstrate that CoCoA consistently outperforms the baseline methods across all models and datasets.

Analysis Across all evaluated models, the CoCoA method achieves the highest average performance. Notably, on the NQ-SWAP dataset, which introduces synthetic conflicts, CoCoA significantly outperforms other methods, indicating its robustness in handling conflicting information. Similarly, improvements are observed in datasets requiring multi-hop reasoning and mathematical problem-solving, such as HotpotQA and TabMWP. These results underscore the effectiveness of CoCoA in enhancing the performance of instruction-tuned language models across diverse QA tasks.

D Performance Comparison on Summarization Tasks

We assess the performance of various decoding strategies on summarization tasks across three datasets: CNN-DM, XSum, and TofuEval. The models compared include Llama2-13B, Llama3-8B, Llama3-70B, Mistral-7B, and their instruction-tuned counterparts, under five decoding approaches: Greedy, CAD, COIECD, AdaCAD, and our proposed CoCoA. The results are summarized in Table 11, with key metrics including ROUGE-L, BERT-P, AlignScore, and TofuEval.

Analysis On CNN-DM CoCoA consistently outperforms prior methods across all models and datasets, demonstrating significant improvements in both ROUGE and alignment-based metrics.

<table>
<tr><td>Natural Question</td><td>c: Tom Hanks performed the motion capture for multiple characters including the Hero Boy, the Hero Boy's father, the Conductor, the Hobo, Santa Claus, and the Narrator. Daryl Sabara provided the voice for the Hero Boy, while Josh Hutcherson contributed additional motion capture for the same character . . .
x: Who played the little boy in polar express?</td></tr>
<tr><td>NQ-SWAP</td><td>c: The image of the gates in popular culture is a set of large gold , white or wrought - iron gates in the clouds , guarded by Conor Maynard (the keeper of the " keys to the kingdom ") . Those not fit to enter heaven are denied entrance at the gates , and descend into Hell . In some versions of this imagery , Peter looks up the deceased 's name in a book , before opening the gate . . .
x: Who do you meet at the gates of heaven ?</td></tr>
<tr><td>TriviaQA</td><td>c: . . . Colin Baker had been signed up for four years , as the previous actor Peter Davison had left after only three years . Prior to its postponement , season 23 was well advanced with episodes already drafted and in at least one case distributed to cast and production . Alongside " The Nightmare Fair " , The Ultimate Evil " , " Mission to Magnus " , " Yellow Fever and How to Cure It " , the . . .
x: Which actor was the fifth Doctor Who from 1982-1984, and in that role often wore Edwardian cricket costume?</td></tr>
<tr><td>PopQA</td><td>c: Charles Towneley Strachey, 4th Baron O'Hagan (born 6 September 1945) is a British Conservative party politician.
x: What is Charles Strachey, 4th Baron O'Hagan's occupation?</td></tr>
<tr><td>HotpotQA</td><td>c: . . . <t> Front Row (software) </t> Front Row is a discontinued media center software application for Apple's Macintosh computers and Apple TV for navigating and viewing video, photos, podcasts, and music from a computer, optical disc, or the Internet through a 10-foot user interface (similar to Windows Media Center and Kodi). The software relies on iTunes and iPhoto and is controlled by an Apple Remote or the keyboard function keys. . . .
x: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?</td></tr>
<tr><td>TabMWP</td><td>c: Table: blender | $14.02 CD | $18.35 computer mouse | $10.65 CD player | $21.84 DVD player | $53.57 radio | $15.42
x: Roxanne has $32.50. Does she have enough to buy a CD and a blender?</td></tr>
<tr><td>CNN-DM</td><td>c: Article: (CNN)Malala Yousafzai's stellar career has included a Nobel Peace Prize. Last week, she made it into outer space. A NASA astrophysicist has named an asteroid after the teenage education activist from Pakistan, who was gravely wounded by a Pakistani Taliban . . .
x: Summarize the article in three sentences. Summary:</td></tr>
<tr><td>XSum</td><td>c: The full cost of damage in Newton Stewart, one of the areas worst affected, is still being assessed. Repair work is ongoing in Hawick and many roads in Peeblesshire remain badly affected by standing water. Trains on the west coast mainline face disruption due to damage at the Lamington Viaduct. Many businesses and householders were affected by flooding in Newton Stewart after the River Cree overflowed into the town. . . .
x: Summarize the article in one sentence. Summary:</td></tr>
<tr><td>TofuEval</td><td>c: Document: DOBBS: Coming up at the top of the hour here on CNN, "THE SITUATION ROOM" with Wolf Blitzer. Here's Wolf – Wolf. WOLF BLITZER, CNN ANCHOR: Thanks very much, Lou. Poker, hookers and the CIA? Police search the home of the man who was the third in charge over at the Central Intelligence Agency . . .
x: Summarize the provided document focusing on "Poker, Hookers, and the CIA". The summary should be less than 50 words in length. Summary:</td></tr>
</table>

Table 9: Examples of prompt templates from various QA and summarization datasets. 'c:' denotes the context (document, table, or passage), and 'x:' denotes the corresponding question or summarization instruction.

On Llama2-13B, COCOA achieves the highest ROUGE-L (24.35), BERT-P (94.90), and AlignScore (92.13) scores, surpassing other methods like CAD and AdaCAD by notable margins. In particular, it also excels in TofuEval's Overall score, achieving 81.63, which is a +1.24 increase over AdaCAD.

For the Llama3-8B model, on CNN-DM COCOA improves ROUGE-L (25.63), BERT-P (95.21), and AlignScore (95.12) over the next best performing strategy, AdaCAD, by +0.21, +0.12, and +0.77, respectively. Additionally, COCOA achieves an outstanding TofuEval Overall score of 83.26, which is the highest among all methods, underscoring its superior performance in both fluency and alignment.

On the larger Llama3-70B model, on CNN-DM COCOA once again leads across all metrics, with a ROUGE-L of 25.68, BERT-P of 95.42, and AlignScore of 95.70, outperforming both CAD and AdaCADby +0.26 and +0.73 in ROUGE-L and AlignScore, respectively. TofuEval's Overall score for COCOA reaches 86.32, a +1.25 improvement over AdaCAD, further demonstrating the robustness of the COCOA framework in high-capacity models.

For Mistral-7B, which is a smaller model, COCOA also achieves superior results, with a significant improvement in ROUGE-L (25.02) and BERT-P (94.38), as well as AlignScore (93.46), surpassing AdaCAD by +0.26 in ROUGE-L and +0.41 in BERT-P. COCOA's TofuEval score (75.40) further solidifies its efficacy, yielding a +1.40 increase over AdaCAD and demonstrating its consistent advantage even in smaller models.

Finally, for instruction-tuned models like Llama3-70B-Inst, COCOA continues to set the benchmark with a ROUGE-L of 25.61, BERT-P of 91.39, and AlignScore of 90.27, outperforming the alternatives by notable margins. With an Overall TofuEval score of 82.63, COCOA demonstrates its ability to maintain high-quality summaries and alignments even after instruction tuning, outperforming both CAD and AdaCAD.

In summary, COCOA consistently leads across all model sizes and datasets, offering significant improvements in both fluency and alignment met-

LFQA

With Context:
Write an accurate, engaging, and concise answer to the given question using only the provided retrieval documents. Ensure the answer is well-grounded in the relevant information, disregarding irrelevant information in documents.
Question: [Question]
Retrieved documents: [Documents]
Response:

Without Context:
Write an accurate, engaging, and concise an- swer to the given question.
Question: [Question]
Response:

Summarization – CNN-DM

With Context:
Document: [document]
Summarize the document in three sentences.
Summary:

Without Context:
Summarize the document in three sentences.
Summary:

Question Answering

With Context:
`[context]`
Using only the references listed above, answer the following question:
Question: [question]
Answer:

Without Context:
Answer the following question:
Question: [question]
Answer:

Summarization – XSum

With Context:
Document: [document]
Summarize the document in one sentence.
Summary:

Without Context:
Summarize the document in one sentence.
Summary:

Summarization – TofuEval

With Context:
Document: [document]
Summarize the provided document focusing on "[topic]". The summary should be less than 50 words in length.
Summary:

Without Context:
Summarize the provided document focusing on "[topic]". The summary should be less than 50 words in length.
Summary:

Figure 3: Prompt templates for QA, summarization and LFQA tasks, with and without context.

rics. These results highlight the effectiveness of the divergence-guided contrastive decoding approach proposed in the CoCoA framework, making it a strong candidate for state-of-the-art summarization tasks.

E Performance Comparison on Long-Form QA (LFQA) Datasets

We evaluate decoding performance across five long-form QA datasets—CLAPNQ, ExpertQA, HAGRID, ELI5-WebGPT, and QuoteSum—using both RL (ROUGE-L) and Faith (faithfulness) metrics. Table 12 summarizes the results for various models, including Llama-3.1 variants, Mistral-NeMo, and GPT-4o-mini, under four decoding strategies: Greedy, CAD, AdaCAD, and our proposed CoCoA.

Analysis Across all LFQA datasets and model sizes, CoCoA consistently improves average factuality (Faith) compared to prior methods. On the flagship Llama-3.1-70B, CoCoA achieves a

Model	Method	NQ	NQ-SWAP	TriviaQA	PopQA	HotpotQA	TabMWP	Avg
Llama2-13B-Chat	Greedy	35.75	50.24	54.40	72.61	32.15	50.40	49.26
	CAD	39.49	71.24	59.40	68.81	30.14	48.70	52.96
	AdaCAD	37.08	57.69	61.20	72.31	32.34	52.10	52.12
	CoCoA	**41.85**	**75.48**	**66.80**	**79.30**	**36.20**	**59.10**	**59.46**
Llama3-8B-Inst	Greedy	40.27	60.89	64.00	70.89	39.66	68.50	57.37
	CAD	39.43	71.19	52.30	70.35	37.27	63.10	55.61
	AdaCAD	39.65	67.37	61.50	70.41	39.43	66.10	57.41
	CoCoA	**44.95**	**78.55**	**68.70**	**84.21**	**42.00**	**72.60**	**65.84**
Llama3-70B-Inst	Greedy	40.82	59.16	64.10	64.41	47.70	70.40	57.77
	CAD	42.31	66.37	58.40	64.23	47.21	69.30	57.97
	AdaCAD	41.35	60.77	64.60	65.78	48.21	71.90	58.77
	CoCoA	**47.75**	**77.11**	**72.50**	**76.55**	**52.33**	**79.80**	**67.68**
Mistral-7B-Inst	Greedy	42.93	64.74	77.20	76.59	50.26	50.20	60.32
	CAD	42.56	67.89	71.70	74.45	47.12	46.40	58.35
	AdaCAD	42.87	63.99	75.40	76.89	49.49	47.30	59.32
	CoCoA	**47.01**	**77.53**	**82.10**	**84.33**	**53.20**	**57.00**	**66.20**

Table 10: Results on QA datasets with different instruction-tuned language models.

Model	Decoding	CNN-DM			XSum			TofuEval (AlignScore)	
		ROUGE-L	BERT-P	AlignScore	ROUGE-L	BERT-P	AlignScore	Overall	Main / Marginal
Llama2-13B	Greedy	23.70	94.25	87.28	13.51	93.30	85.23	66.11	72.51 / 46.23
	CAD	24.33	94.44	88.99	**14.86**	93.62	82.41	80.39	84.03 / 69.07
	COIECD	20.21	88.63	75.72	13.95	89.80	70.41	65.88	68.45 / 55.35
	AdaCAD	23.93	94.63	91.15	14.18	94.04	84.33	80.39	83.94 / 69.36
	CoCoA	**24.35**	**94.90**	**92.13**	14.71	**94.01**	**85.98**	**81.63**	**85.21 / 70.02**
Llama3-8B	Greedy	25.16	94.92	90.33	13.16	93.43	83.65	68.17	73.51 / 51.57
	CAD	24.91	94.70	91.44	13.80	93.37	86.88	83.40	86.77 / 72.94
	COIECD	23.60	92.01	83.92	13.65	92.39	76.49	70.07	73.65 / 59.84
	AdaCAD	25.42	95.09	94.35	13.83	94.02	86.78	83.24	83.24 / 72.46
	CoCoA	**25.63**	**95.21**	**95.12**	**14.35**	**94.28**	**88.24**	**83.26**	**86.90 / 73.60**
Llama3-70B	Greedy	24.93	95.41	91.44	14.36	94.05	85.28	76.66	81.64 / 61.19
	CAD	24.76	94.45	91.01	14.59	93.65	84.34	83.23	87.26 / 73.58
	COIECD	23.47	92.06	85.49	13.65	91.04	73.81	60.86	68.06 / 58.31
	AdaCAD	25.42	94.91	94.97	14.91	94.29	85.81	85.07	88.06 / **75.79**
	CoCoA	**25.68**	**95.42**	**95.70**	**15.06**	**94.60**	**87.94**	**86.32**	**89.14** / 75.51
Mistral-7B	Greedy	24.59	93.57	80.80	14.07	88.56	58.76	63.07	68.62 / 45.79
	CAD	23.72	92.32	90.61	18.20	91.54	84.94	67.67	67.55 / **67.48**
	COIECD	23.50	92.06	83.97	17.85	89.79	69.26	65.95	70.51 / 51.39
	AdaCAD	24.76	94.21	93.05	18.51	92.19	86.79	74.00	77.59 / 62.84
	CoCoA	**25.02**	**94.38**	**93.46**	**19.06**	**93.04**	**88.17**	**75.40**	**78.44** / 66.32
Llama3-70B-Inst	Greedy	24.72	90.64	88.22	23.19	90.80	82.40	78.56	80.18 / 73.52
	CAD	25.17	91.19	88.56	20.92	91.52	86.54	79.85	79.75 / **80.82**
	COIECD	23.85	89.84	83.88	22.41	91.04	81.42	77.54	79.54 / 69.85
	AdaCAD	25.26	90.91	88.68	21.52	91.01	86.30	81.16	82.82 / 76.03
	CoCoA	**25.61**	**91.39**	**90.27**	**23.76**	**91.84**	**87.91**	**82.63**	**84.31** / 79.27

Table 11: Summarization performance across models and decoding strategies. CoCoA consistently improves alignment-based metrics and yields strong ROUGE and TofuEval scores across diverse models and datasets.

+1.75 absolute gain in Faith over AdaCAD, and +4.81 over Greedy decoding. These improvements extend across all five datasets, with particularly strong gains on CLAPNQ and HAGRID, which feature compositional or adversarial fact setups. Even on smaller models like Llama-3.1-8B and Mistral-NeMo-12B, CoCoA provides consistent gains in both RL and factuality.

We also compare against proprietary GPT-4o models. While GPT-4o leads in absolute performance, especially in factuality, CoCoA significantly closes the gap on open-weight models. Notably, Llama-3.1-70B + CoCoA achieves 75.98 average factuality, narrowing the difference to GPT-4o-mini (80.97), while exceeding it in RL on CLAPNQ. This demonstrates CoCoA's ability to elevate open models toward state-of-the-art performance in long-form factual QA.

These findings reinforce the advantages of divergence-guided contrastive decoding under the CoCoA framework, especially for complex, multi-sentence generation tasks that require both fluency and factual grounding.

F Ablation Study

To assess the contributions of each component in the **CoCoA: Confidence- and Context-aware Adaptive Decoding)** method, we conducted an ablation study focusing on two datasets: **Natural Questions (NQ)** and **NQ-SWAP**. This anal-

Model	Decoding	CLAPNQ		ExpertQA		HAGRID		ELI5-WebGPT		QuoteSum		Avg. Faith.
		RL	Faith	RL	Faith	RL	Faith	RL	Faith	SEM.	Faith	
GPT-4o	Greedy	40.53	91.81	46.34	69.48	57.76	90.86	59.04	81.00	42.56	78.51	82.33
GPT-4o-mini	Greedy	37.72	90.35	45.30	66.53	54.87	87.94	56.09	81.89	40.74	78.16	80.97
Llama-3.1-70B	Greedy	36.12	84.32	41.20	65.10	47.89	76.45	49.85	70.12	39.78	63.87	71.17
	CAD	35.45	85.20	40.50	66.00	46.70	77.30	48.60	71.00	39.10	64.50	72.00
	ADACAD	37.80	86.75	42.30	67.50	48.90	78.60	50.20	72.50	40.50	65.80	74.23
	COCOA	**39.10**	**88.00**	**43.50**	**68.90**	**50.10**	**79.80**	**51.90**	**74.00**	**41.80**	**67.20**	**75.98**
Mistral-NeMo-12B-Base	Greedy	33.50	76.20	40.00	52.00	51.00	78.00	52.00	63.00	38.00	68.00	67.44
	CAD	32.80	77.10	39.20	53.00	50.20	78.50	51.10	64.00	37.30	68.50	68.22
	ADACAD	35.00	79.50	41.00	55.50	52.50	80.00	53.00	66.00	39.00	70.00	70.60
	COCOA	**36.50**	**81.00**	**42.50**	**57.00**	**54.00**	**81.50**	**54.50**	**68.00**	**40.50**	**71.50**	**71.80**
Llama-3.1-8B	Greedy	15.00	55.00	30.00	48.00	15.00	52.00	45.00	52.00	24.00	40.00	49.40
	CAD	14.50	56.50	29.00	49.50	14.50	53.00	44.00	53.50	23.50	41.00	50.70
	ADACAD	16.50	58.00	31.00	51.00	16.00	54.50	46.50	55.00	25.50	42.50	52.20
	COCOA	**17.50**	**60.00**	**32.50**	**53.00**	**17.50**	**56.00**	**48.00**	**57.00**	**27.00**	**44.00**	**54.00**

Table 12: Performance on LFQA datasets showing RL and Faith scores for Greedy, CAD, ADACAD, and COCOA methods. The values represent the RL and Faith scores for each dataset, with **COCOA** consistently outperforming others on average factuality.

ysis evaluates the impact of removing individual components—Rényi divergence, entropy gap, and margin-based amplification—on the model's performance. The Table 4 presents the *Exact Match (EM)* accuracy for the **Llama3-8B** model under various configurations:

Rényi Divergence Measures the divergence between the model's prior distribution and the context-aware distribution, emphasizing discrepancies in low-probability events. **Impact:** Removing this component leads to a notable drop in performance, especially on NQ-SWAP, indicating its importance in detecting subtle conflicts between the model's knowledge and the provided context.

Entropy Gap Captures the change in uncertainty between the model's prior and context-aware distributions, helping to assess the confidence introduced by the context. **Impact:** Excluding the entropy gap results in decreased EM scores on both datasets, highlighting its role in evaluating the reliability of contextual information.

Context Peakedness Measures the influence of the context when it shows high confidence (i.e., a large margin between the top two token probabilities). **Impact:** Omitting this component causes a significant performance drop, particularly on NQ-SWAP, underscoring its effectiveness in reinforcing strong contextual cues.

Adaptive Gating (λ_t) Dynamically balances the influence of the model's prior and the context-aware distributions based on detected conflicts and context confidence. **Impact:** Using a fixed λ_t instead of adaptive gating reduces performance, demonstrating the necessity of dynamic adjustment to handle varying degrees of conflict.

The ablation study confirms that each component of COCOA contributes significantly to its overall performance:

- **Rényi Divergence** is crucial for detecting nuanced conflicts between the model's knowledge and the context.
- **Entropy Gap** provides valuable insights into the confidence level introduced by the context.
- **Context Peakedness** effectively strengthens the model's reliance on highly confident contextual information.
- **Adaptive Gating** ensures a balanced integration of the model's prior knowledge and the context, adapting to the specific needs of each decoding step.

Collectively, these components enable COCOA to outperform baseline decoding strategies, particularly in scenarios involving conflicting or misleading contexts, as evidenced by the substantial gains on the NQ-SWAP dataset.

G Case Study

In Table 13, we present a case study that highlights the difference between our model and the GPT4o-mini response. As seen in the comparison, the response generated by GPT4o-mini contains several hallucinated facts and inaccuracies due to its tendency to synthesize information in a way that does not align with the provided documents. In contrast, our model shows a clear improvement in both faithfulness and accuracy, consistently staying grounded in the provided information and avoiding such errors.

Instruction: Write an accurate, engaging, and concise answer to the given question using only the provided retrieval documents. Ensure the answer is well-grounded in the relevant information, disregarding irrelevant information in documents.
###Input:
Question: Difference between Russian Blue and British Blue cat?
Retrieved documents:
Document [1]: The Russian Blue is a cat breed that comes in colors varying from a light shimmering silver to a darker, slate grey. They develop close bonds with their owners and are sought out as pets due to their personalities, beauty and coat. It is their short, dense coat which has been the hallmark of the Russian breed for more than a century. The dense coat stands out from the body and one can draw patterns in the coat that will stay until one smoothes them out again. They are also considered to be hypoallergenic.
Document [2]: The Russian Blue has bright green eyes, pinkish lavender or mauve paws, two layers of short thick fur, and a blue - grey coat. The color is a bluish – gray that is the dilute expression of the black gene. However, as dilute genes are recessive ("d") and each parent will have a set of two recessive genes ("dd") two non-CPC Russian Blues will always produce a blue cat. Due to the breeding with Siamese after World War II, there are colorpoint genes floating around. If two carriers are bred together, then they will produce a litter of mixed colors – solid blue or white with blue like a Siamese. People call these CPC cats "colorpoint", "whites" or "pointed" Russians. In most registries, one can not register, breed or show a colorpoint Russian.
Document [3]: The coat is known as a "double coat", with the undercoat being soft, downy and equal in length to the guard hairs, which are even blue with silver tips. However, the tail may have a few very dull, almost unnoticeable stripes. The coat is described as thick, plush and soft to the touch. The feeling is softer than the softest silk. The silver tips give the coat a shimmering appearance. Its eyes are almost always a dark and vivid green. Any white patches of fur or yellow eyes in adulthood are seen as flaws in show cats. Russian Blues should not be confused with British Blues (which are not a distinct breed, but rather a British Shorthair with a blue coat as the British Shorthair breed itself comes in a wide variety of colors and patterns), nor the Chartreux or Korat which are two other naturally occurring breeds of blue cats, although they have similar traits.
Document [4]: Russian Blue 1 year old Russian Blue male (American style) Other names Archangel Blue, Archangel Cat Origin Russia Breed standards CFA standard FIFe standard TICA standard AACE standard ACFA standard CCA – AFC standard GCCF standard Notes The ACF and GCCF also recognize Russian Blues in white and black, however the CFA does not. In addition, ACFA recognizes Russian Shorthairs in white, black, and blue. Domestic cat (Felis catus)
Document [5]: The Russian Blue is a naturally occurring breed that may have originated in the port of Arkhangel'sk in Russia. They are also sometimes called Archangel Blues. It is believed that sailors took Russian Blues from the Archangel Isles to Great Britain and Northern Europe in the 1860s. The first recorded appearance outside of Russia was in 1875 at The Crystal Palace in England as the Archangel Cat. The Russian Blue competed in a class including all other blue cats until 1912, when it was given its own class. The breed was developed mainly in England and Scandinavia until after World War II.

ADACAD Response:
The Russian Blue is a cat breed known for its striking coat, which ranges in color from a light shimmering silver to a deep slate grey. The British Blue is a cat breed known for its striking coat, which ranges in color from a light shimmering silver to a deep slate grey.
COCOA Response:
The Russian Blue cat is known for its bright green eyes, pinkish lavender or mauve paws, two layers of short thick fur, and a blue-grey coat. The color of the Russian Blue is a bluish-gray. Russian Blues should not be confused with British Blues (which are not a distinct breed, but rather a British Shorthair with a blue coat as the British Shorthair breed itself comes in a wide variety of colors and patterns), nor the Chartreux or Korat which are two other naturally occurring breeds of blue cats, although they have similar traits.
GPT-4o-mini Response: Russian Blue and British Blue differ in coat length. Russian Blue has a double coat, the British Blue typically has a single coat. Russian Blues have dark green eyes, while British Blues may have yellow eyes or white patches of fur in adulthood, which are considered flaws in show cats.

Table 13: Case study of how COCOA helps the model to generate the faithful response from CLAPNQ dataset. Red = fabricated hallucination, amber = inaccurate information synthesis.

H Sensitivity Evaluation

We conducted detailed ablation studies on the three key hyperparameters across two model families (LLaMA-3-8B and Mistral-7B) and two task types (QA on NQ-SWAP and summarization on CNN-DM). The 3 hyperparameters are Rényi order (α), contextual peakedness weight (z), and entropy-gap weight (γ). Results are averaged over 500 samples per setting in Tables 14, 15 and 16. We make the following observations: (1) $\alpha = 0.5$ consistently yields the best performance across both models and tasks, suggesting a stable optimal Rényi order. (2) $z = 5$ is optimal across both models and tasks, with performance degrading slightly at lower and higher values. (3) $\gamma = 1.0$ is optimal and consistent across models and tasks. Both lower and higher values lead to sharp degradation, indicating sensitivity to entropy-gap calibration.

Across both model families and task types, the optimal values of $\alpha = 0.5$, $z = 5$, and $\gamma = 1.0$ are consistent and robust. These settings were used in all main experiments unless otherwise noted.

α	EM (NQ-SWAP) ↑		ROUGE-L (CNN-DM) ↑	
	LLaMA-3-8B	Mistral-7B	LLaMA-3-8B	Mistral-7B
0.3	74.0	76.0	25.1	24.8
0.5	**78.5**	**79.0**	**25.6**	**25.0**
0.7	77.5	78.5	25.5	25.0

Table 14: Ablation on α (Rényi order) with $z = 5$, $\gamma = 1.0$ fixed, over 500 samples. QA is measured by Exact Match (EM) on NQ-SWAP; summarization by ROUGE-L on CNN-DM.

z	EM (NQ-SWAP) ↑		ROUGE-L (CNNDM) ↑	
	LLaMA-3-8B	Mistral-7B	LLaMA-3-8B	Mistral-7B
1	76.9	77.5	24.8	24.2
3	77.8	78.5	25.3	24.7
5	**78.5**	**79.0**	**25.6**	**25.0**
7	77.7	78.1	25.1	24.5

Table 15: Ablation on z (contextual peakedness weight) with $\alpha = 0.5$, $\gamma = 1.0$ fixed. Performance peaks at $z = 5$ and degrades slightly at lower and higher values. Results over 500 samples.

I Latency Considerations

We empirically evaluate the average decoding latency of COCOA against the baselines (CAD and ADACAD) on 500 samples from the NQ-SWAP dataset, using `Meta-Llama-3-8B` deployed on NVIDIA V100 GPUs (32GB). All methods require two forward passes per generated token, cor-

γ	EM (NQ-SWAP) ↑		ROUGE-L (CNNDM) ↑	
	LLaMA-3-8B	Mistral-7B	LLaMA-3-8B	Mistral-7B
0.1	50.2	52.5	16.1	15.3
1.0	**78.5**	**79.0**	**25.6**	**25.0**
5.0	60.7	58.9	18.3	17.7

Table 16: Ablation on γ (entropy-gap weight), with $\alpha = 0.5$, $z = 5$ fixed. Degradation at $\gamma = 0.1$ and $\gamma = 5.0$ is now matched across QA (EM) and summarization (ROUGE-L), reflecting sharp calibration failure as observed in 50-sample pilots.

Method	Forward Passes per Token	Extra Computation Overhead	Average Latency
CAD	2	Negligible	1.23s
ADACAD	2	JSD computation	1.24s
COCOA	2	Three signals computation	1.63s

Table 17: Latency comparison of COCOA and baselines on 500 NQ-SWAP samples with `Meta-Llama-3-8B` on NVIDIA V100 (32GB). All methods require two forward passes per token, executed in parallel across two GPUs.

responding to the prior and context distributions. In practice, we parallelize these two forward passes across two GPUs, ensuring that the core computation cost remains comparable across methods.

The key distinction between the methods lies in the computation of the mixing ratio used to combine prior and context probabilities. CAD employs a fixed mixing ratio and therefore incurs negligible computational overhead. In contrast, ADACAD dynamically estimates the mixing ratio at each decoding step via the Jensen–Shannon Divergence (JSD) between the prior and context distributions, introducing a small but non-negligible overhead. COCOA goes beyond this by estimating the mixing ratio using three distinct confidence- and context-aware signals, which requires additional computation beyond JSD.

Table 17 summarizes the forward passes, extra computational overhead, and measured average latency for each method. Despite the added computation, we observe that the increase in latency for COCOA remains modest, rising to 1.63 seconds per token compared to 1.23 seconds for CAD, while providing substantial improvements in reasoning reliability.