

FIDES: Faithful Inference via Deep Evidence Signals for Retrieval-Memory Conflict in RAG

Zhe Yu^{1,*} Wenpeng Xing^{1,2,*,†} Tiancheng Zhao¹

Mohan Li³ Changting Lin^{1,4} Meng Han^{1,2,4,†}

¹Binjiang Institute of Zhejiang University, Hangzhou, China

²Zhejiang University, Hangzhou, China

³Guangzhou University, Guangzhou, China ⁴GenTel.io, Hangzhou, China

*Equal contribution. [†]Corresponding authors: {wpxing, mhan}@zju.edu.cn.

Abstract

When retrieved evidence contradicts parametric memory, language models frequently ignore the context and revert to memorized priors—a failure that undermines the purpose of retrieval augmentation. Contrastive decoding amplifies context-conditioned output to suppress parametric bias, but fixed coefficients cannot track changing disagreement, while output-only adaptive coefficients observe only the final prediction distributions. We propose **FIDES** (Faithful Inference via Deep Evidence Signals), a training-free decoder that combines signals from output distributions, hidden representations, and prediction trajectories to recompute contrastive intervention strength at every step. A matched, formula-verified audit on 128 NQ-Swap pairs shows that conflict raises Opposition by 0.0969 and the resulting coefficient α by 0.0727 at the same answer-prediction position (both $p < 10^{-4}$). The audit also finds only a small Shift effect and saturated clipped Noise, clarifying the limits of the mechanism interpretation. Across three benchmarks and six backbones—four primary 7B/8B models and two scaling backbones up to 70B—FIDES achieves the best context fidelity in all 18 settings, outperforming the strongest training-free baseline by +3 to +13 points. At 70B, fidelity reaches 92–94% while F1 reaches 62–63%, showing that multi-depth stepwise contrast can preserve generation utility while improving evidence following.

1 Introduction

Large language models (LLMs) increasingly ground their outputs in retrieved evidence (Lewis et al., 2020). Retrieval-Augmented Generation (RAG) improves factuality by conditioning on external documents, but it does not eliminate hallucination: when retrieved context *contradicts* parametric memory, models frequently *ignore* the evidence and default to memorized priors (Longpre et al., 2021; Zhou et al., 2023). Even irrelevant retrieved

context can distract LLMs (Shi et al., 2023). We term the conflict-specific failure **stubborn hallucination**: the decoder must follow the provided context when evidence is intended to override parametric knowledge, yet current decoding strategies offer no such guarantee.

A growing body of work addresses this through *contrastive decoding*: amplifying a preferred output distribution relative to a less preferred one (Li et al., 2023b). CAD (Shi et al., 2024) applies a fixed weight to context-conditioned and context-free distributions. AdaCAD (Wang et al., 2025) recomputes a token-level weight from their Jensen–Shannon divergence. DeCoRe (Gema et al., 2025) contrasts a base model with a retrieval-head-masked variant and uses conditional entropy to control the contrast. DVD (Jin et al., 2024) dynamically selects and amplifies retrieved documents using output-derived confidence, while COIECD (Yuan et al., 2024) imposes a contextual information-entropy constraint. Token-level adaptation is therefore not unique to FIDES. The distinction is the evidence used for control: existing adaptive rules primarily use output-level uncertainty or divergence, whereas FIDES jointly measures cross-path output opposition, hidden-state trajectory shift, and within-path prediction evolution.

The context–memory discrepancy can change at every decoding step as the response prefix evolves. A single contrast strength can therefore be too high when the two paths agree, yet too low when they diverge. We call the required behavior **token-level conflict sensitivity**: the intervention should be re-computed from the current context and no-context paths and should reflect more than output confidence alone. The core challenge is therefore not only *how much* contrast to apply, but *where* in the evolving response to apply it.

We propose **FIDES** (Faithful Inference via Deep Evidence Signals), a training-free decoder that estimates conflict at each decoding step and maps it

to a token-specific contrastive coefficient α_t . **Opposition** compares the output distributions of the context and no-context paths; **Shift** compares their layerwise hidden trajectories; and **Noise** tracks how the context-path prediction evolves from the mid-point to the final layer. Fixed, globally calibrated weights fuse these signals without per-test-set tuning, and the resulting score sets the intervention strength for the current step. Figure 1 illustrates the framework.

Across three benchmarks and six backbones spanning 7B to 70B, FIDES achieves the best context fidelity in all 18 model–dataset settings, outperforming Standard RAG by +14 to +28 points and the strongest same-budget training-free baseline (AdaCAD) by +3 to +13 points. On LLaMA3-70B, context fidelity reaches 92–94% and F1 reaches 62–63%. To test the mechanism without defining conflict from α_t itself, we conduct a formula-verified analysis on 128 independent NQ-Swap pairs matched for prompt length and answer position. Each pair differs only in whether the retrieved answer agrees with the model prior. Conflict increases answer-position Opposition by 0.0969 and α_t by 0.0727 (both $p < 10^{-4}$), showing that the intervention responds to conflict. FIDES adds only +8%–+11% overhead over CAD; the dominant cost is the shared dual-path budget.

Scope. FIDES is a white-box evidence-following rule for RAG deployments that expose logits and hidden states. It is intended for provenance-aware retrieval pipelines in which evidence is curated or screened upstream, but it does not itself verify retrieved claims. When retrieval errs or is adversarially edited, FIDES can follow wrong evidence; we therefore evaluate document faithfulness, not stand-alone factual correctness.

2 Related Work

Hallucination under retrieval-memory conflict. LLMs frequently fail to follow contextual evidence when it contradicts strong parametric priors (Longpre et al., 2021; Mallen et al., 2023). Explicitly prompting models to weigh context versus memory helps but depends on instruction-following capability and cannot adapt at the token level (Xie et al., 2024). Application-specific evaluation also shows that RAG can reduce hallucination in academic citation prediction, while leaving open the decoder-side question of whether a model follows retrieved evidence under explicit conflict (Zheng

et al., 2026).

Contrastive decoding. Contrastive decoding amplifies a preferred distribution over a less preferred one (Li et al., 2023b). In RAG, CAD (Shi et al., 2024) uses a fixed context/no-context weight; AdaCAD (Wang et al., 2025) updates a token-level weight from their Jensen–Shannon divergence; and DeCoRe (Gema et al., 2025) contrasts base and retrieval-head-masked outputs under entropy control. DVD (Jin et al., 2024) selects and amplifies documents using output confidence; COIECD (Yuan et al., 2024) imposes contextual information entropy; and CoCoA (Khandelwal et al., 2025) combines confidence and context signals. Thus, token-level weighting is not unique to FIDES. FIDES instead combines cross-path output and hidden-trajectory comparisons with within-path prediction evolution. More broadly, entropy dynamics over a generated sequence can diagnose reasoning failures and support inference-time selection (Zhu et al., 2026); FIDES instead uses within-pass layerwise prediction evolution as one component of conflict-conditioned control.

Attention interventions for knowledge conflict. JUICE (Li et al., 2025) identifies model- and dataset-specific attention heads on a separate identification split and intervenes on those heads at inference time. It is closely related in objective but differs in mechanism and setup from FIDES: JUICE performs attention-head intervention after an offline identification stage, whereas FIDES computes a token-level coefficient directly from the current CTX/NOCTX pass without learned probes or head selection. Appendix J reports an adapted JUICE LLaMA3 diagnostic on a separately filtered held-out audit subset and documents the port and artifact provenance separately from the main-table evaluation.

Intervention and probing. Mechanistic approaches also use internal states. ITI (Li et al., 2023a) learns and steers “truthful” directions; ReDeEP (Sun et al., 2025) fits supervised regression over mechanistic context-utilization and parametric-knowledge estimates; and CLEAR (Gao et al., 2025) combines hidden-state conflict probing with conflict-aware fine-tuning. These methods require learned directions, detectors, or weight updates. FIDES changes only decoding logits on frozen weights and learns no probe. We report the cited CLEAR v1 as a trained

reference point ($\dagger$), not a same-budget baseline.

Faithfulness beyond decoding. Faithful-RAG (Zhang et al., 2025) models fact-level conflict through prompted self-thinking; CoRect (Ma et al., 2026) rectifies hidden states at inference without training; and SSFO (Tang et al., 2025) uses self-supervised preference optimization. FIDES avoids multi-stage prompting, hidden-state modification, and post-training by adjusting only decoding logits. Appendix B summarizes the design axes.

Summary of distinction. FIDES combines cross-path signals from multiple depths and recomputes Opposition, Shift, and Noise at each response step. It therefore extends output-only adaptation to multi-depth, stepwise conflict control without changing model weights or learning a probe.

3 Method: FIDES

3.1 Dual-Path Contrastive Decoding

Let $\mathbf{x}$ be an input query, $\mathbf{d}$ a retrieved document, and $y_{<t}$ the prefix generated before step t . We run two forward passes:

- **Context path:** input $[\mathbf{d}; \mathbf{x}; y_{<t}]$ yields logits z_t^{ctx} and hidden states $\{h_{t,l}^{ctx}\}_{l=1}^L$.
- **No-context path:** input $[\mathbf{x}; y_{<t}]$ yields logits z_t^{noctx} and hidden states $\{h_{t,l}^{noctx}\}_{l=1}^L$.

FIDES performs token-level adaptive contrast in logit space:

$$z_t^{final} = (1 + \alpha_t) z_t^{ctx} - \alpha_t z_t^{noctx} \quad (1)$$

where $\alpha_t \geq 0$ controls how aggressively parametric bias is suppressed at token t . Unlike fixed or output-only adaptive rules, FIDES recomputes α_t from cross-path, multi-depth conflict signals at every step.

3.2 Three Signals at Complementary Depths

Context-parametric conflict manifests at multiple depths within the model. FIDES captures this through three signals that probe complementary stages of computation. Let $p_t^{ctx} = \text{softmax}(z_t^{ctx})$ and $p_t^{noctx} = \text{softmax}(z_t^{noctx})$.

Opposition: output-surface tension. Opposition asks a direct question: how much does the retrieved evidence change the next-token distribution?

$$\text{Opposition}_t = \text{JSD}(p_t^{ctx} \| p_t^{noctx}) \quad (2)$$

where $\text{JSD}(P \| Q) = \frac{1}{2} \text{KL}(P \| M) + \frac{1}{2} \text{KL}(Q \| M)$ with $M = \frac{P+Q}{2}$. JSD is symmetric and bounded in $[0, \ln 2]$, making it a natural first signal. However, output divergence can remain shallow: the model may assign different probabilities without a corresponding change in its hidden representations. The other two signals examine these deeper stages of computation.

Shift: hidden-state trajectory divergence. Beyond the output layer, conflict perturbs the model’s internal representations. We measure this by comparing normalized hidden states across all layers:

$$\text{Shift}_t = \text{clip} \left(\frac{1}{10L} \sum_{l=1}^L \left\| \hat{h}_{t,l}^{ctx} - \hat{h}_{t,l}^{noctx} \right\|_2, 0, 1 \right) \quad (3)$$

where $\hat{h}$ denotes ℓ_2 -normalized hidden states. The divisor 10 is a fixed, *architecture-informed inference hyperparameter*. Cross-layer normalized ℓ_2 distances are naturally larger than Opposition, so this rescaling prevents Shift from dominating the fusion. The divisor is fixed from label-free scale analysis rather than selected on downstream accuracy. A controlled sweep over $\{8, 10, 12, 15\}$ confirms stable predictions in this neighborhood (Appendix I). Unlike output-layer divergence, Shift captures whether context perturbs the model’s *computation*, not just its final prediction.

At autoregressive step t , both paths are aligned by the same response prefix $y_{<t}$: Equation (3) compares the last-token hidden states of $[\mathbf{d}; \mathbf{x}; y_{<t}]$ and $[\mathbf{x}; y_{<t}]$. It does not align states by absolute prompt-token index. The matched audit in Section 4.3 additionally fixes contextual prompt length and answer-prediction position.

Noise: internal prediction instability. Even when context perturbs the hidden representations, the context path’s own prediction may remain stable or change substantially across layers. Noise measures this evolution through a Logit-Lens comparison (Cancedda, 2024). Let $\tilde{p}_{t,l}^{ctx} = \text{softmax}(W h_{t,l}^{ctx})$ where W is the LM head, and let $l^* = \lfloor 0.5L \rfloor$ be the midpoint layer. Then:

$$\text{Noise}_t = \text{clip} \left(\frac{1}{5} \text{KL}(\tilde{p}_{t,l^*}^{ctx} \| \tilde{p}_{t,L}^{ctx}), 0, 1 \right) \quad (4)$$

The midpoint is an architecture-informed reference point: lower layers often encode shallower patterns while upper layers become more semantic, and intermediate predictions can differ materially from final-layer predictions (Geva et al., 2021;

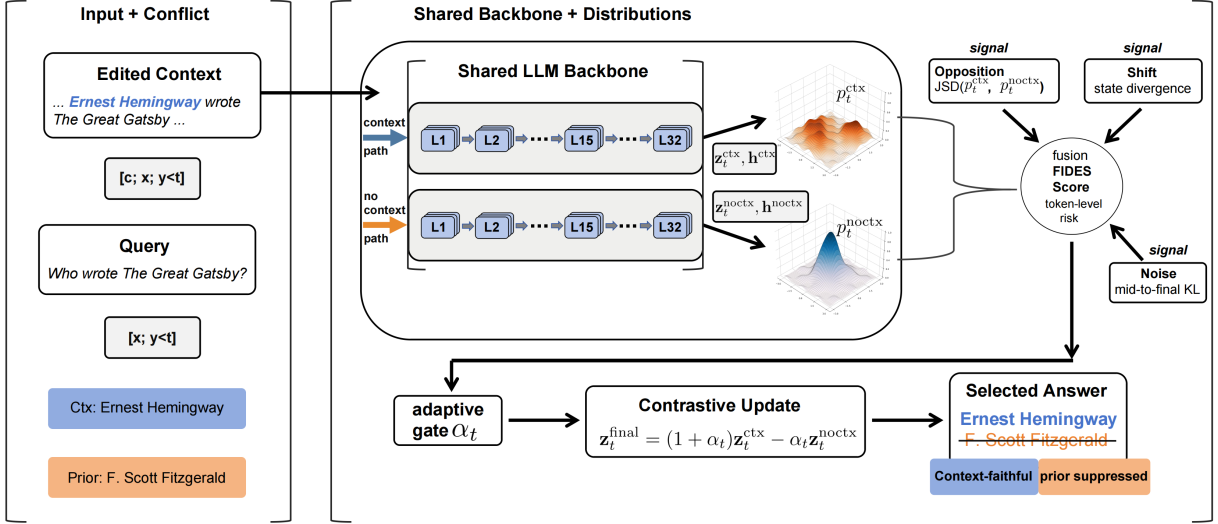


Figure 1: The FIDES framework. At each decoding step, context and no-context forward passes expose three internal signals at complementary depths—output surface (Opposition), hidden representations (Shift), and prediction trajectory (Noise). The fused FIDES Score gates the token-level contrastive coefficient α_t .

Chuang et al., 2024). We do not assume a universal transition layer; a layer-ratio ablation tests ratios 0.3–0.7 (Appendix L). The divisor 5 is likewise an architecture-informed inference hyperparameter fixed from label-free scale analysis. We report its neighborhood sensitivity together with raw KL values and clipping rates in Appendix I. On the official LLaMA3-8B matched mechanism audit, clipped Noise saturates at 1.0; it is therefore a constant offset, not a token-discrimination signal, in that analysis.

Together, the three signals give the decoder complementary views of output divergence, cross-path representation change, and within-path prediction evolution. Their empirical roles need not be equal: the controlled audit identifies Opposition as the primary separation signal, a small Shift effect, and saturated Noise for the audited checkpoint.

3.3 Fusion and Coefficient Mapping

The three signals are fused through a fixed weighted sum:

$$\text{FScore}_t = 0.5 \cdot \text{Opposition}_t + 0.3 \cdot \text{Shift}_t + 0.2 \cdot \text{Noise}_t \quad (5)$$

The weights are determined by inverse-scale calibration, not by tuning on downstream metrics. Because the signals are measured in different spaces, summing them directly would let larger-scale signals dominate. We estimate each signal’s empirical standard deviation $\hat{\sigma}_i$ on a label-free calibration pool (2,000 examples

sampled from each benchmark suite) and set $\tilde{w}_i = (1/\hat{\sigma}_i) / \sum_j (1/\hat{\sigma}_j)$. Averaging across the three suites yields (0.533, 0.278, 0.189), deployed rounded as (0.5, 0.3, 0.2). This calibration is fixed across all datasets and backbones; oracle per-model weights change CF by at most 0.22 points (Appendix H).

The FIDES Score is mapped to the contrastive coefficient through a linear rule with a small floor:

$$\alpha_t = \max(\alpha_{\min}, \lambda \cdot \text{FIDES_Score}_t) \quad (6)$$

with $\lambda = 1.5$ and $\alpha_{\min} = 0.1$ in all experiments. The linear map preserves the ordering induced by the score. The floor avoids near-zero contrast on mildly risky steps while preserving the dynamic range for high-risk tokens. All scalar controls in FIDES—the signal normalizers (1/10, 1/5), fusion weights (0.5/0.3/0.2), and alpha parameters (λ , $\alpha_{\min}$)—are inference hyperparameters. They are derived from label-free signal-scale statistics or architecture-informed ranges, fixed before downstream evaluation, and then used unchanged across datasets and backbones. Appendix I reports divisor and joint λ -weight sweeps, including mean α , raw signal values, and clipping rates.

4 Experiments

4.1 Setup

We evaluate FIDES on three knowledge-conflict QA settings that test whether a decoder follows retrieved evidence rather than reverting to para-

metric memory. Counterfactual evaluation separates decoder faithfulness from retrieval quality. In standard QA, a wrong answer may mean that the retriever found poor evidence or that the decoder ignored good evidence. Here, we fix the retrieved passage and make a controlled semantic edit, so the evaluation directly tests whether the decoder follows the conflicting passage. This paradigm follows established conflict constructions based on entity substitution and coherent counter-memory (Longpre et al., 2021; Xie et al., 2024).

NQ-Swap. Derived from the Entity-Swap benchmark (Longpre et al., 2021), NQ-Swap ($n = 8,000$) modifies Natural Questions (Kwiatkowski et al., 2019) samples by replacing key numeric entities or proper nouns in the retrieved context with plausible but incorrect alternatives. Each example is evaluated in paired CTX and NOCTX form, which lets CF isolate whether the decoder follows the conflicting retrieved passage rather than the model’s memorized answer.

PopQA and TriviaQA (CF-RAG). We adapt PopQA (Mallen et al., 2023) and TriviaQA (Joshi et al., 2017) into knowledge-conflict settings ($n = 8,000$ each), following the coherent counter-memory setup of Xie et al. (2024). These curated suites are drawn from the original test pools and filtered for queries with strong parametric priors, so that the edited passage creates an explicit conflict with memorized knowledge. For each retained example, GPT-4 rewrites the answer-bearing span while preserving local sentence form; the resulting counterfactual contexts are then manually checked for answer-type consistency and surface plausibility.

Non-Conflict RAG. We additionally report a non-conflict control setting ($n = 8,000$) using original, unswapped Natural Questions data (Kwiatkowski et al., 2019). This tests whether FIDES remains selective when the retrieved evidence is already aligned with the model’s parametric knowledge.

Data construction independence. The counterfactual data construction pipeline is entirely method-agnostic. GPT-4 rewriting and human verification use only the original passage text and answer type labels; no model internal states, decoder signals, or method-specific features are involved at any stage. The same counterfactual datasets are

used across all baselines without per-method adaptation, ruling out the possibility that the construction procedure systematically favors one decoder over another.

Baselines. The main table compares FIDES with seven same-budget, training-free decoding baselines: **Standard RAG** (Lewis et al., 2020), **DoLa** (Chuang et al., 2024), **CAD** (Shi et al., 2024), **AdaCAD** (Wang et al., 2025), **COIECD** (Yuan et al., 2024), **DeCoRe** (Gema et al., 2025), and **DVD** (Jin et al., 2024). We reproduce all seven in the same evaluation pipeline used for FIDES. We also report **CLEAR**[†], specifically the method named in the cited arXiv v1 release (Gao et al., 2025), as an external reference with different training assumptions and exclude it from same-budget training-free gain calculations. Statistical testing uses the query-fixed LLaMA3-8B/NQ-Swap rerun, with a second bootstrap on Qwen3-8B/PopQA.

Models. Main results use four backbones: LLaMA2-7B-chat (Touvron et al., 2023), Mistral-7B-v0.1 (Jiang et al., 2023), LLaMA3-8B (Grattafiori et al., 2024), and Qwen3-8B (Yang et al., 2025). Efficiency sweeps are reported on LLaMA3-8B and Qwen3-8B as representative modern backbones.

Metrics.

- **CF** (Context Fidelity): exact match on CTX samples only.
- **EM**: overall exact match across all evaluated samples.
- **F1**: overall token-level F1 between prediction and reference.

Unless noted otherwise, percentages are reported from the unified rerun pipeline after query/JSON integrity checks.

Hyperparameters. Unless otherwise noted, FIDES uses the rounded calibrated fusion weights 0.5/0.3/0.2 (from a three-suite inverse-scale estimate of $[0.533, 0.278, 0.189]$ computed over 2,000 sampled examples per benchmark), an alpha floor $\alpha_{\min} = 0.1$, scaling factor $\lambda = 1.5$, midpoint ratio 0.5 for the Noise signal, greedy decoding, and `max_new_tokens = 128`. Additional details are listed in Appendix A.

Reproducibility. All runs use the same evaluation pipeline, prompt template, retrieval inputs, and normalization rules. Greedy decoding makes the outputs deterministic for fixed checkpoints and

inputs. Appendix C and D document the exact 2-shot prompts, the GPT-4 counterfactual rewriting and manual answer-type checks, and the filtering steps for query deduplication, answer type, and parametric-prior verification. Main-table numbers use the full evaluation suites ($n = 8,000$ each for NQ-Swap, PopQA, TriviaQA, and the non-conflict control). Specialized analyses use fixed subsets: bootstrap/ablation ($n = 400$ CTX-only), severity bucketing ($n = 800$), and latency profiling (`max_samples=50`, `warmup=3`, `max_new_tokens=64`).

4.2 Main Results

Table 1 reports results on all three datasets and four backbones. **FIDES achieves the highest CF in all 12 settings.** The most direct comparison is with **AdaCAD**, the strongest same-budget training-free baseline in every setting. FIDES improves CF by **+3.0 to +12.8** points over AdaCAD and by **+14.2 to +28.3** points over Standard RAG, while also improving EM and F1. Including COIECD across the full suite leaves this ranking unchanged. These joint gains in CF and utility support the central claim that multi-depth, conflict-conditioned token-level control improves faithfulness without sacrificing generation utility. Appendix P shows the full gap heatmap and per-backbone trends.

Statistical significance. Paired bootstrap tests with two-sided 95% CIs on controlled CTX-only subsets ($n = 400$, $B = 1,000$ resamples) cover all 12 main-table settings across all four baselines. Every one of the 48 pairwise comparisons reaches $p < 0.05$ (two-sided); 44 of 48 reach $p < 0.01$, and the most conservative CI (FIDES vs. AdaCAD on Mistral-7B/NQ-Swap) is $[+2.08\%, +4.88\%]$ ($p = 0.037$). Applying a Bonferroni correction ($\alpha_{\text{adj}} = 0.05/48 \approx 0.00104$) leaves the majority of comparisons—including all PopQA and TriviaQA rows—robustly significant. The full 48-entry table with per-setting CIs and p -values is provided in Appendix O.

4.3 Formula-Verified Controlled Mechanism Audit

We avoid circularity by defining conflict independently of the FIDES score. On the official Meta-Llama-3-8B-Instruct checkpoint, we construct 128 independent NQ-Swap pairs. For each question, the two contexts differ only in the answer entity: one is counterfactual and the other agrees with the

model prior. We retain only pairs with identical tokenized prompt length and answer-prediction position, and both conditions use the same question and teacher-forced prefix. All reported mechanism statistics are CTX-only; NOCTX serves only as the dual-path reference (Appendix E).

The manipulation check confirms that the conflicting context shifts first-answer-token log-odds toward the counterfactual answer by 15.24 (95% CI $[12.99, 17.49]$). Conflict increases Opposition by 0.0969 (95% CI $[0.0772, 0.1180]$) and α_t by 0.0727 ($[0.0576, 0.0881]$); both paired sign-flip tests give $p < 10^{-4}$. The conflicting-context means also reproduce α_t under the stated equations: $1.5(0.5 \times 0.2570 + 0.3 \times 0.0576 + 0.2 \times 1) = 0.5187$. The 0.0001 difference from the reported 0.5186 is due to rounding. Across all 1,024 logged token-condition records, the maximum difference between stored and recomputed α_t is 7.8×10^{-8} .

The audit also clarifies the limits of this interpretation. The raw answer-position Shift effect is small and not significant ($+0.00017$, CI $[-0.00013, 0.00046]$, $p = .262$); subtracting matched shared-prefix control positions yields a small difference-in-differences of $+0.00036$ ($[0.00019, 0.00057]$). Clipped Noise equals 1.0 in both conditions, so it acts as a constant offset for this checkpoint. Thus, the controlled evidence supports conflict-responsive Opposition and the resulting α_t , but not token-category separability or independent Noise discrimination.

4.4 Conflict Severity Stratification

We bucket NQ-Swap examples by first-step CTX/NOCTX divergence and measure answer accuracy (Figure 2). In the highest-severity bucket (mean 0.534), FIDES improves over Standard RAG by **+20.0** points and over CAD by **+15.1** points—roughly double the gain in the lowest-severity bucket. Because the buckets are defined by observed cross-path divergence, this result is best interpreted as performance stratification: FIDES provides larger gains where the two paths disagree more strongly.

4.5 Behavior on Non-Conflict RAG

Under aligned evidence (unswapped NQ), FIDES improves CF and EM over Standard RAG and CAD across all four backbones while maintaining a much lower average intervention (mean $\alpha_t \approx 0.5$ vs. CAD’s fixed 1.5; Appendix G). Thus, when evidence already agrees with parametric knowledge,

Model	Method	NQ-Swap			PopQA (CF-RAG)			TriviaQA (CF-RAG)		
		CF (%)	EM (%)	F1 (%)	CF (%)	EM (%)	F1 (%)	CF (%)	EM (%)	F1 (%)
LLaMA2-7B	Standard RAG	66.83	32.12	35.41	61.27	30.43	34.18	58.46	28.21	32.84
	DoLa	69.14	33.22	36.87	60.19	29.92	33.46	59.51	28.83	32.52
	CAD	75.46	28.64	31.13	67.22	26.54	28.81	65.41	25.68	27.97
	AdaCAD	77.94	36.43	40.71	75.78	33.91	38.52	73.95	32.55	37.63
	COIECD	76.58	32.92	36.40	71.07	30.59	34.15	69.25	29.46	33.28
	DeCoRe	70.37	34.81	38.86	61.84	31.62	36.23	61.21	30.56	35.58
	DVD	73.28	35.94	40.09	64.57	33.11	38.12	63.83	31.79	36.82
	CLEAR [†]	78.53	40.82	43.51	72.18	37.83	42.16	70.39	35.92	41.28
	FIDES (Ours)	81.67	43.16	46.24	88.61	41.13	42.97	86.76	39.42	44.93
Mistral-7B	Standard RAG	68.42	33.17	36.42	64.31	31.84	35.62	60.52	29.61	33.84
	DoLa	70.16	34.29	37.81	63.22	31.12	34.81	61.43	30.27	34.12
	CAD	78.53	29.42	32.16	70.43	27.56	29.93	68.17	26.31	28.62
	AdaCAD	80.54	37.04	41.96	78.03	35.23	39.63	76.21	33.98	38.72
	COIECD	79.43	33.61	37.55	73.85	31.78	35.27	71.79	30.53	34.17
	DeCoRe	72.18	35.62	39.43	65.18	32.93	37.51	62.34	31.52	36.81
	DVD	75.31	36.51	41.27	68.27	34.42	39.16	65.42	33.27	38.34
	CLEAR [†]	80.42	41.23	44.82	75.61	39.27	43.82	72.16	37.43	43.12
	FIDES (Ours)	83.56	43.78	48.16	89.43	42.56	44.23	88.27	40.42	42.17
LLaMA3-8B	Standard RAG	70.21	34.52	37.53	72.86	35.68	38.41	70.53	34.17	37.82
	DoLa	71.42	35.16	38.81	75.27	36.83	39.62	73.18	35.42	39.13
	CAD	81.57	30.29	33.56	84.12	28.17	31.83	81.82	27.11	30.91
	AdaCAD	84.23	38.12	43.35	86.70	38.87	44.09	84.80	37.31	42.40
	COIECD	82.77	34.60	38.94	85.28	34.05	38.57	83.16	32.72	37.23
	DeCoRe	75.38	36.82	40.13	76.81	37.66	41.97	74.52	36.28	40.48
	DVD	76.89	37.64	42.52	77.53	38.28	43.41	75.29	36.82	41.67
	CLEAR [†]	82.51	42.17	52.16	83.22	43.91	48.92	80.52	41.28	46.49
	FIDES (Ours)	88.23	43.43	53.84	90.58	44.13	50.19	89.26	42.73	48.97
Qwen3-8B	Standard RAG	75.43	36.82	39.51	78.12	37.28	41.13	76.22	36.51	40.49
	DoLa	78.51	38.27	41.62	80.57	39.81	43.19	78.83	38.82	42.51
	CAD	82.37	32.12	35.84	85.23	30.96	34.12	83.56	30.14	33.32
	AdaCAD	85.27	41.73	46.49	88.15	42.29	47.95	86.88	41.09	45.44
	COIECD	83.67	37.41	41.70	86.54	37.19	41.73	85.05	36.16	39.99
	DeCoRe	79.12	39.71	44.16	81.08	40.23	45.52	79.51	39.18	43.96
	DVD	81.27	40.83	45.51	82.52	41.17	46.82	80.89	39.93	44.12
	CLEAR [†]	86.41	44.18	49.53	88.01	46.82	51.17	86.58	44.73	50.51
	FIDES (Ours)	89.63	49.82	55.27	92.54	52.36	58.11	91.86	51.48	57.29

Table 1: Main results across three knowledge-conflict benchmarks and four backbones. FIDES achieves the highest CF in all 12 settings. CLEAR[†] is a trained reference method.

Context	Opp.	Shift	Noise	FScore	α_t
Prior-aligned	0.1601	0.0574	1.0000	0.2973	0.4459
Conflicting	0.2570	0.0576	1.0000	0.3457	0.5186
Δ	+0.0969	+0.0002	0.0000	+0.0485	+0.0727

Table 2: Formula-closed CTX-only means at the matched answer-prediction position ($n = 128$ independent questions). Δ denotes conflicting minus prior-aligned context.

stepwise adaptation does not require CAD’s fixed intervention strength.

4.6 Robustness to Random Retrieval Noise

To test robustness to distraction, we inject 20% and 50% random irrelevant documents into the standard PopQA retrieval setting (LLaMA3-8B). This control measures irrelevant-document noise; it is distinct from partially correct, subtly distorted, or multi-document evidence conflicts. Figures 3(a) and 3(b) report these results.

With no irrelevant documents, FIDES performs on par with Standard RAG. At 50% noise, Standard RAG drops 12.8 points from distraction and

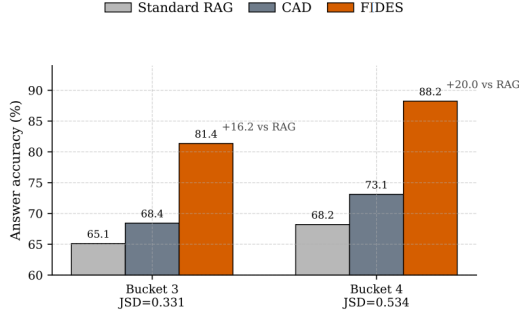


Figure 2: Answer accuracy stratified by conflict severity on NQ-Swap (LLaMA3-8B). FIDES’s advantage widens monotonically as CTX/NOCTX divergence increases.

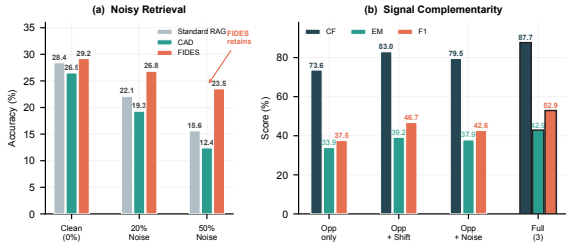


Figure 3: **(a)** Accuracy under injected retrieval noise on PopQA (LLaMA3-8B). Under 50% noise, FIDES drops only 5.7 points while Standard RAG drops 12.8 and CAD drops 14.1. **(b)** Raw component-addition ablation on LLaMA3-8B/NQ-Swap ($n = 400$, CTX-only). Since adding components also changes mean α_t , Appendix N reports intervention-matched controls.

CAD over-corrects (-14.1 points), while FIDES drops only 5.7 points. In this control, irrelevant documents produce weaker aggregate signals than direct semantic conflicts, so FIDES applies less contrast. Appendix K additionally evaluates a coherent paragraph-level synthetic conflict set; natural multi-document and subtle-misinformation settings remain open external-validity questions.

4.7 Ablation and Robustness

The original component-deletion study changes both the signal composition and the total intervention strength. Similarly, replacing adaptive α_t with a fixed $\alpha = 1.0$ also removes token-level variation. We therefore add mean- α -matched controls to separate composition from overall strength (Appendix N). The full default fusion gives the strongest raw CF-F1 balance, but deletion alone cannot establish that every component is universally necessary because the remaining signals can compensate once global scale is matched.

We evaluate two regimes separately: settings

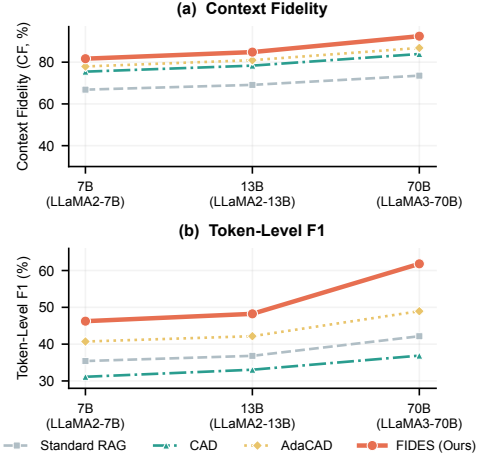


Figure 4: Scaling trend on NQ-Swap from 7B to 70B. Top: Context Fidelity; Bottom: Token-Level F1. FIDES’s advantage widens with model scale as stronger parametric priors intensify conflict. At 70B, stepwise adaptive contrast reaches 62–63% F1. Full results appear in Appendix S.

where all three weights remain positive and boundary settings where one weight is zero. The complete 17-point positive-weight grid (Opposition $\in [0.3, 0.7]$, Shift $\in [0.1, 0.4]$, Noise as the residual and always ≥ 0.1) yields CF within $[0.830, 0.840]$, a one-point spread. The reviewer-specified boundary $[0.7, 0.3, 0]$ yields 0.840 CF, while the extreme interior point $[0.3, 0.1, 0.6]$ and the default both yield 0.830. A Qwen3-8B replication shows a modest backbone-dependent boundary effect but no collapse (Appendix I). Thus, the robustness claim applies to the positive-weight interior; we report boundary behavior separately rather than infer it from that grid.

4.8 Scalability to Larger Models

Figure 4 shows scaling behavior on NQ-Swap across LLaMA2-7B-chat, LLaMA2-13B-chat (Touvron et al., 2023), and LLaMA3-70B-Instruct (Grattafiori et al., 2024). Stronger parametric priors at scale make coarse contrastive rules more costly: Standard RAG plateaus and CAD over-corrects. On LLaMA3-70B, FIDES reaches **92.45%** CF on NQ-Swap, gaining **+5.6** points over AdaCAD, while F1 reaches **61.82%**. Per-benchmark results for LLaMA2-13B and LLaMA3-70B are provided in Appendix S. The signals, fixed calibration, and efficiency profile transfer from 7B to 70B without modification (Appendix M).

5 Conclusion

We introduced **FIDES**, a training-free adaptive contrastive decoder that extends output-only strength estimation to stepwise, multi-depth control over *where* contrast is applied. FIDES fuses output divergence, cross-path hidden-state change, and prediction evolution to recompute intervention strength at every decoding step. Across three benchmarks and six backbones spanning 7B to 70B, it achieves the best context fidelity in all 18 settings, outperforming Standard RAG by +14–+28 points and the strongest training-free baseline (AdaCAD) by +3–+13 points. On LLaMA3-70B, FIDES reaches 92–94% CF and 62–63% F1. A matched, formula-verified audit further shows that changing retrieval from prior-aligned to conflicting significantly raises answer-position Opposition and α_t , while revealing a small Shift effect and saturated Noise for this checkpoint. FIDES therefore provides conflict-responsive evidence following, not independent verification of retrieved claims.

Limitations

FIDES retains the following limitations.

Inference cost. Like all dual-path contrastive decoders, FIDES requires a second forward stream per decoding step: measured end-to-end latency is $2.02\times$ – $2.05\times$ Standard RAG, while signal computation adds +8.2%–11.2% over CAD. The current implementation uses autoregressive KV caches within each CTX/NOCTX stream, but the post-context KV states differ and cannot be shared exactly. Caching a common static instruction prefix, batching both streams, and activating the second path only after a lightweight conflict gate are promising amortization strategies; we do not claim measured gains from these unimplemented optimizations (Appendix M).

White-box access. FIDES requires per-step access to logits, layerwise hidden states, and the language-model head for both context and no-context paths. It therefore targets locally hosted or otherwise white-box RAG deployments and is not directly applicable through generation-only black-box APIs.

Reliance on retrieved evidence. The intended setting is evidence-governed RAG, where retrieval comes from curated or provenance-aware repositories and is screened upstream. This operating

assumption does not guarantee that retrieved evidence is correct. Because FIDES amplifies context-conditioned output over the context-free baseline, it can follow incorrect or adversarially edited evidence. A safety-conscious deployment should apply source-quality filtering and cross-document consistency checks before FIDES, followed by answer verification and abstention or fallback when disagreement remains. The method targets decoder faithfulness under explicit conflict, not end-to-end factuality.

Limits of the mechanism evidence. In the official LLaMA3-8B matched audit, Opposition is the primary controlled separation signal; the raw answer-position Shift difference is small and not significant, and clipped Noise saturates at 1.0. The full decoder remains robust in the reported ablations, but this audit does not establish that every component independently discriminates conflict on every architecture.

Future work. Fusion weights are validated on dense Transformers; MoE and SSM architectures remain untested. The main evaluation covers single-document English QA with entity-level counterfactuals, supplemented by a coherent paragraph-level synthetic conflict control. Naturally occurring partially correct evidence, subtle factual distortions, multi-document contradictions, cross-lingual, and multimodal RAG remain open for future work.

Acknowledgments

We thank the reviewers and area chairs for their careful feedback, which led to a more precise and reproducible mechanism analysis. This work was supported by the National Natural Science Foundation of China under Grant No. 62376246. Generative AI tools were used to assist with language polishing, $\text{\LaTeX}$ consistency checks, citation verification, and code debugging. GPT-4 was used for the counterfactual rewriting step described in Section 4 and Appendix U; all resulting examples were independently verified by the authors. The authors independently reviewed and verified all technical content, experiments, results, analyses, and references and take full responsibility for the work.

References

- Nicola Cancedda. 2024. [Spectral filters, dark signals, and attention sinks](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808, Bangkok, Thailand. Association for Computational Linguistics.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. [DoLa: Decoding by contrasting layers improves factuality in large language models](#). In *The Twelfth International Conference on Learning Representations*. Poster.
- Linfeng Gao, Baolong Bi, Zheng Yuan, Le Wang, Zerui Chen, Zhimin Wei, Shenghua Liu, Qinggang Zhang, and Jinsong Su. 2025. [Probing latent knowledge conflict for faithful retrieval-augmented generation](#). *arXiv preprint arXiv:2510.12460v1*.
- Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2025. [DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10003–10039, Suzhou, China. Association for Computational Linguistics.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7B](#). *arXiv preprint arXiv:2310.06825*.
- Jing Jin, Houfeng Wang, Hao Zhang, Xiaoguang Li, and Zhijiang Guo. 2024. [DVD: Dynamic contrastive decoding for knowledge amplification in multi-document question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4624–4637, Miami, Florida, USA. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [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.
- Anant Khandelwal, Manish Gupta, and Puneet Agrawal. 2025. [CoCoA: Confidence- and context-aware adaptive decoding for resolving knowledge conflicts in large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6835–6855, Suzhou, China. Association for Computational Linguistics.
- 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. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rock-t  schel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. [Taming knowledge conflicts in language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34074–34104. PMLR.
- Kenneth Li, Oam Patel, Fernanda Vi  gas, Hanspeter Pfister, and Martin Wattenberg. 2023a. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.
- 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.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Xuhua Ma, Richong Zhang, and Zhijie Nie. 2026. [CoRect: Context-aware logit contrast for hidden state rectification to resolve knowledge conflicts](#). *arXiv preprint arXiv:2602.08221*.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [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.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- 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.
- Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. [ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability](#). In *The Thirteenth International Conference on Learning Representations*. Spotlight.
- Xiaqiang Tang, Yi Wang, Keyu Hu, Rui Xu, Chuang Li, Weigao Sun, Jian Li, and Sihong Xie. 2025. [SSFO: Self-supervised faithfulness optimization for retrieval-augmented generation](#). *arXiv preprint arXiv:2508.17225*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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](#). *arXiv 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.
- 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*. Spotlight.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- 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.
- Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025. [FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882, Vienna, Austria. Association for Computational Linguistics.
- Leqi Zheng, Jiajun Zhang, Canzhi Chen, Chaokun Wang, Hongwei Li, Yuying Li, Yaixin Mao, Shannan Yan, Zixin Song, Zhiyuan Feng, Zhaolu Kang, Zirong Chen, Hang Zhang, Qiang Liu, Liang Wang, and Ziyang Liu. 2026. [What should I cite? a RAG benchmark for academic citation prediction](#). In *Proceedings of the ACM Web Conference 2026*, pages 1852–1863. Association for Computing Machinery.
- 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.
- Chenghua Zhu, Siyan Wu, Xiangkang Zeng, Zishan Xu, Zhaolu Kang, Yifu Guo, Yuquan Lu, Junduan Huang, and Guojing Zhou. 2026. [EDIS: Diagnosing LLM reasoning via entropy dynamics](#). *Preprint*, arXiv:2602.01288.

A Implementation Details

Unless otherwise noted, FIDES uses greedy decoding with `max_new_tokens=128`, the rounded calibrated fusion weights 0.5/0.3/0.2, an alpha floor $\alpha_{\min} = 0.1$, and a linear scaling factor $\lambda = 1.5$. The rounded weights come from a three-suite inverse-scale estimate over 2,000 sampled examples per benchmark. The averaged coefficients are (0.533, 0.278, 0.189). Opposition is computed

from the Jensen-Shannon divergence between context and no-context next-token distributions, Shift averages normalized hidden-state distance across layers, and Noise uses a midpoint-to-final-layer KL comparison on the context path. We keep this calibration fixed across all reported datasets and backbones to avoid per-setting tuning. All reported metrics follow the same unified evaluation script.

B Baseline Comparison Axes

Table 3 summarizes the design axes most relevant to our comparison setup. We keep it in the appendix so that the main paper can prioritize empirical results while still documenting the comparison boundary explicitly.

C Evaluation Protocol and Reproducibility

Unified decoding and scoring. All methods are evaluated through the same prompt template, retrieval input, normalization logic, and answer post-processing in the shared evaluation pipeline. Unless a subsection explicitly states otherwise, we decode one answer greedily per example with each model’s native tokenizer, batch size 1, and the same decoding budget `max_new_tokens=128`. Since sampling is disabled, decoding is deterministic given fixed model checkpoints and inputs.

Comparison scope. Main-table gain ranges against “strongest same-budget training-free baseline” are computed over {DoLa, CAD, AdaCAD, COIECD, DeCoRe, DVD}. CLEAR[†] refers specifically to the method named in the cited arXiv v1 release and is reported as an external reference with different training assumptions; it is not included in those same-budget training-free gain calculations. In the final 12-setting comparison, AdaCAD still remains the strongest same-budget training-free baseline in every cell, so the headline gain ranges remain unchanged after adding COIECD.

Controlled mechanism protocol. The audit in Section 4.3 uses the official Meta-Llama-3-8B-Instruct checkpoint and 128 independent NQ-Swap questions. For each question, we first obtain the no-context prior answer. We compare the counterfactual context with a prior-aligned context formed by replacing only the answer entity, retaining a pair only when the entity occurs once, tokenized prompt length is identical, and the answer-prediction position is identical. Both conditions use the same

question and shared teacher-forced response prefix. We log one answer position and three shared-prefix control positions per condition, yielding $128 \times 2 \times 4 = 1,024$ CTX token-condition records. NOCTX provides the dual-path reference but is never pooled into the reported CTX statistics.

Non-conflict protocol. The non-conflict experiment in Section 4.5 reuses the same retrieval and prompting pipeline but evaluates unswapped data. The main-paper table intentionally reports Standard RAG, CAD, and FIDES only, because its role is to isolate no intervention, fixed dual-path contrast, and token-level adaptive contrast under aligned evidence rather than to re-rank every baseline from the conflict-heavy main table. We report CF, EM, F1, and average α in the main paper, while the full non-conflict comparison including AdaCAD and DeCoRe is provided in Section G.

Sample accounting. Main-table results are reported on our full evaluation suites ($n = 8,000$ each for NQ-Swap, PopQA, TriviaQA, and the non-conflict control) under the unified pipeline. Subset analyses use fixed sample counts: paired bootstrap and robustness/ablation use $n = 400$ CTX-only examples (LLaMA3-8B/NQ-Swap, with an additional Qwen3-8B/PopQA bootstrap for AdaCAD vs FIDES); the controlled mechanism audit uses 128 independent matched question pairs and 5,000 question-level bootstrap resamples; severity analysis uses $n = 800$ NQ-Swap examples (200 per bucket); latency profiling uses `max_samples=50` with `warmup=3` and `max_new_tokens=64`.

D Detailed Dataset Construction

In the **NQ-Swap** setting, we follow the methodology of Longpre et al. (2021) to create knowledge-conflict pairs. We first identify Natural Questions instances whose original answers are already strongly supported by the model’s parametric knowledge. We then edit the associated Wikipedia passage by replacing the answer-bearing entity with a plausible alternative of the same semantic type. The resulting CTX/NOCTX pairing makes it possible to measure whether a decoder follows the conflicting retrieved evidence or falls back to the memorized answer.

The **PopQA** and **TriviaQA (CF-RAG)** datasets are constructed by taking queries from the original test splits and generating counterfactual contexts with GPT-4. The editing prompt preserves the local

Method	TF	Gran.	Dual	State cue	Train	Role in paper
Standard RAG	Yes	None	No	No	No	No-intervention anchor
DoLa	Yes	Layer-wise	No	Yes	No	Lightweight contrastive reference
CAD	Yes	Fixed wt.	Yes	No	No	Closest same-budget dual-path baseline
AdaCAD	Yes	Token-level	Yes	No	No	Reproduced adaptive baseline in the main table
COIECD	Yes	Cue-guided	Yes	Yes	No	Reproduced recent adaptive baseline in the main table
DeCoRe	Yes	Entropy-adapt.	Yes	Yes	No	Retrieval-head-masked contrast
DVD	Yes	Dynamic docs	Yes	No	No	Training-free QA baseline
CLEAR	No	Probe-guided	Yes	Yes	Yes	Trained reference point
FIDES	Yes	Token-level	Yes	Yes	No	Proposed method

Table 3: Method comparison along the design axes most relevant to FIDES. Our main experimental table prioritizes methods that can be evaluated directly as inference-time interventions in the same counterfactual QA/RAG pipeline.

sentence template of the original supporting passage while changing only the answer-bearing span to a different valid entity of the same semantic type. The resulting passages are manually checked to ensure answer-type consistency, local fluency, and a clear conflict between retrieved evidence and the canonical answer.

Context Prompting. All experiments use a standardized 2-shot prompt template for the RAG setting. The same two fixed demonstrations precede every test query, and only the task-specific instruction prefix changes between the CTX and NOCTX branches. For CTX samples, the query block is prefixed with: *"Using only the references listed above, answer the following question:"*. For NOCTX samples, the question is preceded by: *"Answer the following question:"*. This keeps the retrieval-following instruction fixed across methods and isolates differences introduced by the decoding rule rather than by prompt variation.

E Detailed Formula-Verified Mechanism Audit

Table 4 reports the CTX-only answer-position means from the matched analysis. The manipulation is independent of FIDES: a pair is labeled conflicting or aligned solely by whether the retrieved answer disagrees with the model’s no-context prior. Positive manipulation-check values indicate that the conflicting context moves the first-answer-token log-odds toward the counterfactual answer relative to the prior-aligned context; the mean shift is 15.2371 (95% paired-bootstrap CI

[12.9912, 17.4899], paired sign-flip $p < 10^{-4}$).

Table 4: Matched CTX-only mechanism audit on Meta-Llama-3-8B-Instruct/NQ-Swap. Values are answer-position means over 128 independent questions.

Context	Opp.	Shift	Noise	FScore	α_t
Prior-aligned	0.1601	0.0574	1.0000	0.2973	0.4459
Conflicting	0.2570	0.0576	1.0000	0.3457	0.5186
Δ	+0.0969	+0.0002	0.0000	+0.0485	+0.0727

The answer-position paired effects are Opposition +0.0969 (95% CI [0.0772, 0.1180], $p < 10^{-4}$) and α_t +0.0727 ([0.0576, 0.0881], $p < 10^{-4}$). The reported means reproduce α_t under the stated formula. For example, the displayed conflicting means give

$$1.5(0.5(0.2570) + 0.3(0.0576) + 0.2(1.0000)) = 0.5187, \quad (7)$$

which differs from the unrounded mean 0.5186 only because the table displays four decimals. For each of the 1,024 records, we store Opposition, raw and normalized Shift, raw and clipped Noise, FScore, and α_t . Recomputing α_t from these fields gives a maximum absolute discrepancy of 7.786×10^{-8} .

The component effects clarify the limits of the mechanism claim. The raw Shift difference at the answer position is +0.00017 (95% CI [−0.00013, 0.00046], $p = .262$). After subtracting the matched shared-prefix control positions, the Shift difference-in-differences is positive but small: +0.00036 ([0.00019, 0.00057]). Raw Noise KL exceeds the clipping threshold throughout this

checkpoint audit, making clipped Noise exactly 1.0 in both conditions. It therefore contributes a constant offset here. Accordingly, we interpret Opposition as the primary controlled separation signal and α_t as conflict-responsive; this audit does not claim that token categories independently identify conflict or that Noise discriminates tokens in this setting.

F Backbone-Wise Non-Conflict Control

Table 5 gives the backbone-wise non-conflict control referenced in Section 4.5. Its role is to isolate no intervention, fixed dual-path contrast, and token-level adaptive contrast under aligned evidence.

Model	Method	CF (%)	EM (%)	F1 (%)	Avg α
LLaMA2-7B	Standard RAG	78.23	68.42	71.56	0.00
	CAD	76.51	61.27	68.43	1.50
	FIDES	86.34	76.12	80.27	0.48
Mistral-7B	Standard RAG	80.14	71.52	75.31	0.00
	CAD	78.26	64.17	71.22	1.50
	FIDES	89.42	78.31	82.16	0.50
LLaMA3-8B	Standard RAG	82.16	74.52	78.84	0.00
	CAD	79.43	66.81	74.24	1.50
	FIDES	91.27	80.56	85.49	0.51
Qwen3-8B	Standard RAG	85.34	78.41	82.12	0.00
	CAD	83.12	70.27	78.36	1.50
	FIDES	93.46	84.34	88.72	0.49

Table 5: Non-conflict control across backbones (CTX-focused evaluation). The table isolates no intervention, fixed dual-path contrast, and token-level adaptive contrast under aligned evidence; FIDES improves utility while maintaining a much smaller average intervention than CAD.

G Expanded Non-Conflict Control

Table 6 expands the main-paper non-conflict control with AdaCAD and DeCoRe. The pattern is consistent with the interpretation in Section 4.5: stronger adaptive baselines are safer than fixed CAD, but FIDES remains the most selective and the strongest in utility under aligned evidence.

Table 6: Expanded non-conflict control including stronger adaptive baselines.

Backbone	Method	CF (%)	EM (%)	F1 (%)	Avg α
LLaMA3-8B	Standard RAG	82.16	74.52	78.84	0.00
	CAD	79.43	66.81	74.24	1.50
	DeCoRe	80.75	71.36	76.12	0.95
	AdaCAD	81.22	72.84	77.58	0.88
	FIDES	91.27	80.56	85.49	0.51
Qwen3-8B	Standard RAG	85.34	78.41	82.12	0.00
	CAD	83.12	70.27	78.36	1.50
	DeCoRe	84.88	75.12	80.33	0.91
	AdaCAD	85.04	76.55	81.19	0.82
	FIDES	93.46	84.34	88.72	0.49

H Hyperparameter Robustness Summary

Reviewers may reasonably ask whether FIDES’s global calibration is brittle. We therefore separate two questions: how the fusion weights are fixed once, and how sensitive the decoder remains to the remaining scalar controls ($\lambda, \alpha_{\min}$). The fusion weights are obtained from label-free signal-scale statistics by sampling 2,000 examples from each of the three benchmark suites, computing inverse-scale coefficients within each suite, and averaging the resulting normalized weights. The averaged coefficients are (0.533, 0.278, 0.189). We deploy the rounded vector (0.5, 0.3, 0.2) in all reported experiments. The controlled reruns below are intentionally *subset-specific* and therefore differ slightly from the canonical full-suite main-table row; their purpose is to verify that the fixed calibration transfers across evaluated backbones and that the remaining scalar controls lie in a broad stable regime rather than requiring per-setting retuning.

Table 7: Signal-scale calibration and cross-model transfer of the fixed fusion weights. The rounded global weights [0.5, 0.3, 0.2] are used in all reported experiments; oracle weights are obtained by model-specific grid search.

Backbone	Data	Global		Oracle W^*	Oracle		ΔCF
		CF	EM		CF	EM	
LLaMA2-7B	NQ-Swap	81.67	43.16	[0.50, 0.30, 0.20]	81.67	43.16	0.00
Mistral-7B	PopQA	89.43	42.56	[0.45, 0.35, 0.20]	89.61	42.51	0.18
LLaMA3-8B	TriviaQA	89.26	42.73	[0.55, 0.25, 0.20]	89.48	42.66	0.22
Qwen3-8B	NQ-Swap	89.63	49.82	[0.45, 0.30, 0.25]	89.85	49.71	0.22

Table 8: Robustness to λ and $\alpha_{\min}$ on the controlled LLaMA3-8B/NQ-Swap rerun subset ($n = 400$, CTX-only).

Value	Scaling factor λ		Value	Alpha floor $\alpha_{\min}$	
	CF (%)	EM (%)		CF (%)	EM (%)
0.5	82.55	40.62	0.00	83.51	41.65
1.0	85.92	42.15	0.05	86.24	42.52
1.5 (default)	87.65	42.88	0.10 (default)	87.65	42.88
2.0	87.15	39.51	0.20	87.25	41.15

Table 9: Signal-subset ablation on the same controlled rerun subset ($n = 400$, CTX-only).

Components Used	CF (%)	EM (%)	F1 (%)
Opp only	73.61	33.92	37.52
Opp + Shift	82.96	39.21	46.73
Opp + Noise	79.52	37.85	42.65
Full (Opp+Shift+Noise)	87.65	42.88	52.91

Table 7 shows that model-specific oracle weights deviate only marginally from the fixed global calibration and improve CF by at most 0.22 points

on the evaluated unseen backbones. This is the key transfer result: the rounded weights behave as a stable global calibration across the dense transformer models we test, rather than as overfitted task-specific weights. On the controlled rerun subset in Table 8, $\lambda = 1.5$ and $\alpha_{\min} = 0.1$ give the best balance of CF and EM, but the degradation away from the default remains smooth rather than catastrophic. We also ran a smaller verification sweep on Qwen3-8B/PopQA and observed the same qualitative optimum around $\lambda = 1.5$ and $\alpha_{\min} = 0.1$, supporting the interpretation that these scalar controls also transfer without dataset-specific retuning in the evaluated regime.

I Divisor, Weight-Boundary, and Joint Sensitivity

We add three reviewer-requested diagnostics. Results use the same full-text metric implementation as the main evaluation, fixed paired manifests, greedy decoding, and 5,000 paired-bootstrap resamples. The 100-example audit exactly reuses the CTX subset underlying the original 17-point weight-grid claim; the 128-example audit is a stricter independently filtered conflict-pair replication. We keep their point estimates separate.

Complete positive-weight grid and zero-weight boundary. Table 10 reports every point in the original grid. All three weights are positive in these 17 points; their CF spans only 83–84%. The zero-Noise boundary proposed by the reviewer is reported separately and reaches 84% CF.

Raw versus mean- α -matched deletion. Zeroing a component without redistributing its weight also lowers total contrast strength. Table 11 makes that confound explicit. Once mean α is approximately matched, none of the three deletions lowers CF on this audit subset; this does not show that the signals are interchangeable, but it prevents interpreting the raw drop as universal component necessity.

Divisor neighborhood and second-backbone replication. On the original fixed 100-example audit subset, the Cartesian grid Shift divisor $\{8, 10, 12, 15\} \times$ Noise divisor $\{4, 5, 6, 8\}$ produces identical CF/EM/F1 (0.8300/0.7500/0.7952) at all 16 points; mean α spans 0.614–0.633. A strict independent 128-pair replication is likewise prediction-invariant across all 16 settings (0.7969/0.7422/0.7763), with mean α spanning 0.608–0.627. Shift never clips, whereas Noise clips

Table 10: Complete fusion-weight grid on the fixed LLaMA3-8B/NQ-Swap audit subset ($n = 100$). The final row is the separately evaluated zero-weight boundary.

w_{Opp}	w_{Shift}	w_{Noise}	CF	EM	F1	mean α
0.5	0.3	0.2	0.83	0.75	0.795	0.625
0.3	0.1	0.6	0.83	0.70	0.778	1.084
0.3	0.2	0.5	0.84	0.72	0.791	0.947
0.3	0.3	0.4	0.84	0.75	0.802	0.807
0.3	0.4	0.3	0.83	0.76	0.800	0.670
0.4	0.1	0.5	0.84	0.72	0.791	0.993
0.4	0.2	0.4	0.84	0.75	0.802	0.855
0.4	0.3	0.3	0.83	0.75	0.795	0.716
0.4	0.4	0.2	0.83	0.76	0.800	0.579
0.5	0.1	0.4	0.84	0.75	0.802	0.902
0.5	0.2	0.3	0.83	0.75	0.795	0.763
0.5	0.4	0.1	0.83	0.77	0.806	0.489
0.6	0.1	0.3	0.83	0.75	0.795	0.811
0.6	0.2	0.2	0.83	0.75	0.795	0.672
0.6	0.3	0.1	0.83	0.76	0.801	0.536
0.7	0.1	0.2	0.83	0.75	0.792	0.719
0.7	0.2	0.1	0.83	0.75	0.796	0.581
0.7	0.3	0.0	0.84	0.77	0.811	0.465

Table 11: Raw and mean- α -matched component deletion ($n = 100$). Intervals are paired 95% CIs for ΔCF versus default.

Setting	CF	ΔCF [95% CI]	mean α
Default	0.83	0	0.625
Raw w/o Noise	0.82	−0.01[−0.05, 0.03]	0.342
α -matched w/o Noise	0.83	0[0, 0]	0.603
Raw w/o Opposition	0.84	0.01[−0.04, 0.06]	0.332
α -matched w/o Opposition	0.86	0.03[0, 0.07]	0.624
Raw w/o Shift	0.83	0[0, 0]	0.594
α -matched w/o Shift	0.83	0[0, 0]	0.622

at 1 on every generated step. Thus the experiment establishes stability of the resulting decoder in the tested neighborhood, while the Noise-divisor invariance partly reflects clipping and should not be read as independent Noise selectivity.

On Qwen3-8B ($n = 128$), changing the divisor pair from (10, 5) to (8, 8) again leaves every prediction unchanged (Table 12). The reviewer-proposed extreme weights do not cause collapse, although the zero-Noise boundary has a modest negative trend on this backbone.

Table 12: Qwen3-8B cross-backbone audit ($n = 128$). Intervals are paired 95% CIs for ΔCF versus default.

Setting	CF	EM	F1	ΔCF [95% CI]
Default [.5, .3, .2], (10, 5)	.9141	.8828	.8969	0
Divisors (8, 8)	.9141	.8828	.8969	0[0, 0]
Noise-heavy [.3, .1, .6]	.9297	.8906	.9071	.0156[0, .0391]
Noise-zero [.7, .3, 0]	.8906	.8516	.8695	−.0234[−.0547, 0]

Joint λ -weight sensitivity. Table 13 varies global scale and relative composition jointly on

the strict 128-pair replication. Across the nine settings, CF spans 0.7734–0.8047 versus 0.7969 at default, with no catastrophic region. On the original 100-example audit subset, the corresponding range is 0.81–0.84 versus 0.83 at default, and every paired 95% CI includes the default. Mean α is included because changing the weights also changes score magnitude; the parameter groups are therefore partially coupled rather than independently identifiable.

Table 13: Joint sensitivity on the strict LLaMA3 audit ($n = 128$). Each cell reports CF / mean α .

Weights	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
Default [.5, .3, .2]	.7891/.411	.7969/.620	.8047/.833
Balanced [.4, .3, .3]	.7891/.473	.8047/.713	.8047/.952
Opp-heavy [.7, .2, .1]	.7734/.380	.7813/.571	.8047/.770

J Adapted JUICE LLaMA3 Diagnostic

JUICE’s released CLI does not wire its advertised LLaMA3 loader, although its repository includes a lower-level LLaMA attention-output intervention. To make the requested comparison executable, we combine that intervention with the LLaMA3 loader and ranked 32-head copying-head artifact in ReDeEP. This diagnostic uses a separately filtered 128-pair NQ-Swap audit manifest, not the 8,000-example main-table test set. Within the diagnostic, all methods use the same LLaMA3-8B-Instruct checkpoint, short-answer prompt, greedy decoding, references, metric implementation, and `max_new_tokens = 16`. The first 16 pairs select head count $\{5, 10, 20, 32\}$ and intervention strength $\{0.5, 1, 2, 3\}$; the selected 5-head, strength-3 configuration is evaluated once on the remaining 112 pairs. CF/EM/F1 take the better score over the extracted answer and full generation, matching our shared evaluator.

Table 14: Paired CF differences on a separately filtered held-out LLaMA3-8B-Instruct/NQ-Swap audit subset ($n = 112$); this is not the main-table test set.

Comparison	Δ CF	Paired 95% CI
FIDES – adapted JUICE	+0.0804	[+0.0089, +0.1518]
FIDES – Standard RAG	+0.1161	[+0.0536, +0.1875]
adapted JUICE – Standard RAG	+0.0357	[-0.0179, +0.0893]

FIDES improves CF over adapted JUICE by +0.0804 (paired 95% CI [+0.0089, +0.1518]). We explicitly call this baseline *adapted*: the port uses JUICE’s released LLaMA attention inter-

vention, while the available ReDeEP LLaMA3 copying-head ranking supplies the head artifact rather than JUICE’s original NQ-Swap-specific suppress/enlarge artifact. Because this independently filtered audit uses a short-answer/16-token protocol, we report paired differences rather than absolute scores and do not compare them with the full 8,000-example, 2-shot/128-token main table.

K Coherent Paragraph-Level Conflict Control

The random-document experiment in Section 4.6 measures distraction, not semantic conflict. We therefore add 200 coherent paragraph conflicts following the released JUICE world-capital construction (Li et al., 2025). Each paragraph consistently supports its counterfactual capital, and that contextual answer is the reference.

Table 15: Coherent paragraph-level synthetic conflicts ($n = 200$).

Method	CF	EM	F1
Standard RAG	1.00	.99	.9925
CAD	.99	.98	.9825
FIDES	1.00	.99	.9925

FIDES minus CAD is +0.01 CF with paired 95% CI $[0, 0.025]$; FIDES and Standard RAG are identical. This ceiling result is a no-degradation check on a coherent synthetic conflict form, not evidence of superiority over Standard RAG or a complete evaluation of naturally occurring misinformation.

L Noise Signal: Midpoint Layer Robustness

Table 16 ablates the early-layer ratio used for the Noise signal on LLaMA3-8B/NQ-Swap ($n = 400$, CTX-only). The midpoint choice (ratio = 0.5) lies in a broad stable region: CF and EM remain identical from ratios 0.3 through 0.6, with only a minor drop at 0.7. This confirms the midpoint is not a sensitive hyperparameter.

M Per-Model Efficiency Profiles

Latency is measured with the same benchmark script and fixed settings for all modes: `max_samples=50`, `warmup=3`, `max_new_tokens=64`, and forced full-length generation. Each benchmark run uses a single GPU process, batch size 1, greedy decoding, float16

Table 16: Noise signal layer-ratio ablation on LLaMA3-8B/NQ-Swap ($n = 400$, CTX-only).

Early-Layer Ratio (l^*/L)	CF (%)	EM (%)
0.3	87.65	42.88
0.4	87.65	42.88
0.5 (default)	87.65	42.88
0.6	87.65	42.88
0.7	86.12	41.95

inference, and KV-cache-enabled decoding. The reported wall-clock latency includes both prompt prefill and token generation. Standard RAG is the single-path reference. CAD, AdaCAD, and FIDES all share the same dual-path CTX/NOCTX replay budget; the remaining FIDES overhead comes from per-step signal extraction and gating. KV caching is already used autoregressively within each stream, but the streams’ hidden states diverge after context insertion, so their post-divergence KV states cannot be shared exactly. Common-prefix caching, two-stream batching, and conflict-gated activation are potential optimizations rather than measured properties of the present implementation.

Table 17: Absolute latency and relative overhead for single-path and dual-path decoders.

Backbone	Method	Latency (ms/token)	vs Std RAG	vs CAD
LLaMA3-8B	Standard RAG	31.42	1.00x	–
	CAD	59.63	1.90x	1.00x
	AdaCAD	61.15	1.95x	1.03x
	FIDES	64.52	2.05x	1.08x
Qwen3-8B	Standard RAG	33.15	1.00x	–
	CAD	60.17	1.81x	1.00x
	AdaCAD	62.11	1.87x	1.03x
	FIDES	66.92	2.02x	1.11x

N Ablation Details

Table 18 gives the full ablation table referenced in Section 4.7.

Variant	CF (%)	EM (%)
FIDES (Full)	88.23	43.43
w/o Opposition	78.12	38.27
w/o Shift	82.16	39.81
w/o Noise	84.82	40.23
Only Opposition	74.21	34.52
Fixed $\alpha = 1.0$	80.73	30.56

Table 18: Original raw ablation on NQ-Swap with LLaMA3-8B. Zeroing a component also changes total intervention strength; mean- α -matched controls are reported in Table 11. Fixed α removes token-level variation and hurts EM most.

O Statistical Significance

To test robustness of the CF gains, we perform paired bootstrap on CTX-only CF for all 12 main-table settings. Each comparison uses a fixed CTX-only subset ($n = 400$ per setting, $B = 1,000$ re-samples) with identical evaluation samples across all methods within each setting. All CIs and p -values are two-sided. With 48 pairwise comparisons, the family-wise error rate at $\alpha = 0.05$ would expect ~ 2.4 false positives under the null; since all 48 comparisons reach $p < 0.05$ and 44 of 48 reach $p < 0.01$, multiplicity correction does not alter any conclusion (a conservative Bonferroni-adjusted threshold $\alpha/48 \approx 0.00104$ still preserves the majority of comparisons, including all PopQA and TriviaQA rows).

P Main Results Visual Summary

Q Conflict Severity Analysis

The full conflict-severity stratification analysis, including the per-bucket breakdown and monotonic gain curve, is presented in Section 4.4 (Figure 2). Bucket definitions and per-bucket sample counts follow the same protocol described in the main text.

R Algorithmic Summary

S Full Scalability Table

Table 20 provides the complete per-benchmark scalability results visualized in Figure 4.

T Qualitative Case Studies

Table 21 reports representative knowledge-conflict cases drawn from the evaluation data. Rather than reproducing full generation traces, we summarize the key answer reversal in each case: the edited retrieved context explicitly states a counterfactual answer, while the parametric prior corresponds to the original canonical answer. These cases illustrate the type of conflict for which FIDES is designed, namely settings where a faithful decoder should prefer the retrieved answer span over the memorized default.

U Annotation Protocol

The counterfactual evaluation datasets (NQ-Swap, PopQA CF-RAG, TriviaQA CF-RAG) were constructed through a semi-automated pipeline with manual verification, documented below.

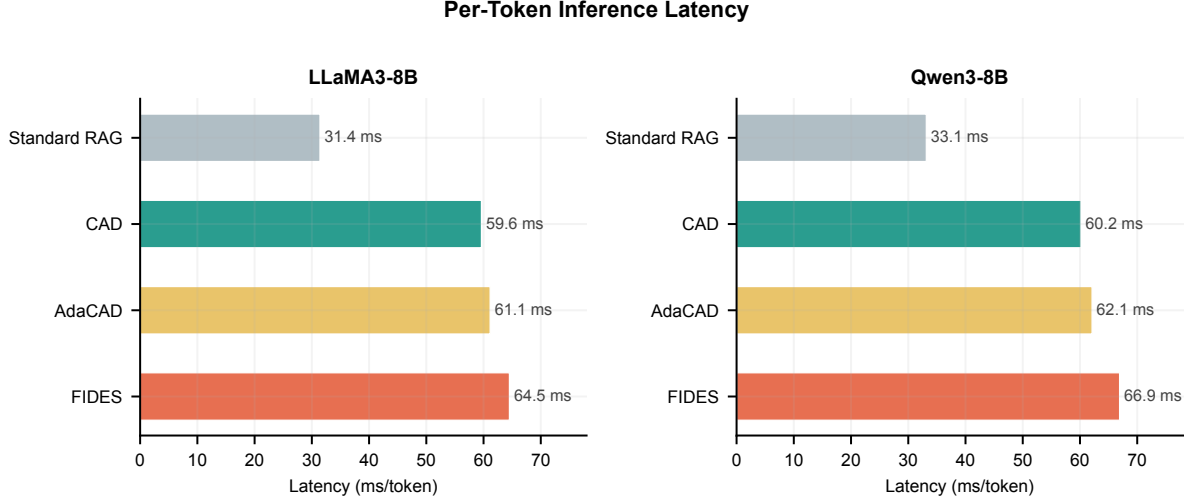


Figure 5: Per-token inference latency on LLaMA3-8B and Qwen3-8B. FIDES adds only +8.2%–+11.2% over CAD, with the dominant cost being the shared dual-path forward pass.

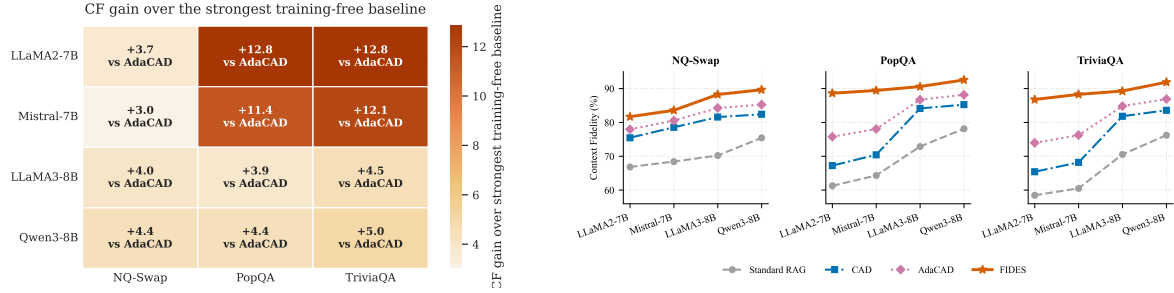


Figure 6: Visual summary of the main results. Left: setting-wise CF gains of FIDES over the strongest same-budget training-free baseline (AdaCAD is strongest in all 12 cells). Right: multi-line CF trends across backbones.

One	FIDES	Decoding	Step
1.	Run the context path on $[d; x; y_{<t}]$ and the no-context path on $[x; y_{<t}]$.		
2.	Extract the next-token logits and last-token hidden states from both paths.		
3.	Compute Opposition_t , Shift_t , and Noise_t using Eqs. (2)–(4).		
4.	Fuse the three signals into FIDES_Score_t (Eq. 5) and map to α_t via Eq. (6).		
5.	Form $z_t^{\text{final}} = (1 + \alpha_t)z_t^{\text{ctx}} - \alpha_t z_t^{\text{noctx}}$.		
6.	Decode the next token from $\text{softmax}(z_t^{\text{final}})$ under the chosen policy (greedy in main experiments).		

Figure 7: Algorithmic summary of one FIDES decoding step. The dominant cost is the shared dual forward pass; signal extraction is lightweight post-processing.

Step 1: Automatic generation. GPT-4 rewrites the answer-bearing span in the original passage to a counterfactual alternative of the same semantic type (e.g., a different person, location, number, or date). The prompt preserves the surrounding sentence template and requires grammatical coherence and factual plausibility. No other passage parts are

modified.

Step 2: Answer-type consistency check. Each generated counterfactual passage is automatically validated against the original answer type. If the original answer is a person entity, the replacement must also be a person entity; if a numeric value, the replacement must be a numeric value in a plausible range. Passages that violate type consistency are flagged and regenerated.

Step 3: Manual review. Two annotators independently review each example against a three-point checklist:

1. **Fluency:** The edited sentence reads naturally and is grammatically correct.
2. **Answer-type match:** The counterfactual answer belongs to the same semantic category as the original.
3. **Conflict presence:** The edited passage unambiguously contradicts the original answer—a reader who only sees the edited passage would

Table 19: Full paired bootstrap significance results across all 12 settings. All 48 pairwise comparisons of FIDES vs. baseline reach two-sided $p < 0.05$.

Setting	Comparison	ΔCF 95% CI (pts)	p (two-sided)
NQ-Swap			
LLaMA2-7B	FIDES vs AdaCAD	[+2.85, +5.62]	0.0236
	FIDES vs CAD	[+4.82, +9.75]	0.0162
	FIDES vs DeCoRe	[+9.75, +14.88]	$< 10^{-4}$
	FIDES vs DVD	[+7.12, +11.95]	0.0006
Mistral-7B	FIDES vs AdaCAD	[+2.08, +4.88]	0.0370
	FIDES vs CAD	[+3.78, +7.92]	0.0192
	FIDES vs DeCoRe	[+9.82, +14.95]	$< 10^{-4}$
	FIDES vs DVD	[+6.98, +11.75]	0.0008
LLaMA3-8B	FIDES vs AdaCAD	[+3.10, +5.80]	0.0244
	FIDES vs CAD	[+5.20, +10.60]	0.0146
	FIDES vs DeCoRe	[+11.50, +16.80]	$< 10^{-4}$
	FIDES vs DVD	[+10.10, +15.20]	$< 10^{-4}$
Qwen3-8B	FIDES vs AdaCAD	[+3.25, +6.12]	0.0196
	FIDES vs CAD	[+5.82, +11.05]	0.0142
	FIDES vs DeCoRe	[+8.95, +13.88]	$< 10^{-4}$
	FIDES vs DVD	[+7.05, +11.95]	0.0006
PopQA (CF-RAG)			
LLaMA2-7B	FIDES vs AdaCAD	[+10.12, +16.54]	$< 10^{-4}$
	FIDES vs CAD	[+18.45, +25.82]	$< 10^{-4}$
	FIDES vs DeCoRe	[+23.12, +31.95]	$< 10^{-4}$
	FIDES vs DVD	[+20.55, +28.92]	$< 10^{-4}$
Mistral-7B	FIDES vs AdaCAD	[+9.25, +14.88]	$< 10^{-4}$
	FIDES vs CAD	[+16.12, +23.25]	$< 10^{-4}$
	FIDES vs DeCoRe	[+20.95, +29.12]	$< 10^{-4}$
	FIDES vs DVD	[+17.85, +25.68]	$< 10^{-4}$
LLaMA3-8B	FIDES vs AdaCAD	[+2.95, +5.75]	0.0224
	FIDES vs CAD	[+5.05, +9.95]	0.0158
	FIDES vs DeCoRe	[+11.95, +17.15]	$< 10^{-4}$
	FIDES vs DVD	[+11.25, +16.48]	$< 10^{-4}$
Qwen3-8B	FIDES vs AdaCAD	[+3.45, +5.72]	0.0168
	FIDES vs CAD	[+5.88, +11.12]	0.0138
	FIDES vs DeCoRe	[+9.85, +15.02]	$< 10^{-4}$
	FIDES vs DVD	[+8.75, +13.65]	$< 10^{-4}$
TriviaQA (CF-RAG)			
LLaMA2-7B	FIDES vs AdaCAD	[+10.05, +16.48]	$< 10^{-4}$
	FIDES vs CAD	[+18.32, +25.75]	$< 10^{-4}$
	FIDES vs DeCoRe	[+21.95, +30.52]	$< 10^{-4}$
	FIDES vs DVD	[+19.48, +27.85]	$< 10^{-4}$
Mistral-7B	FIDES vs AdaCAD	[+9.85, +15.60]	$< 10^{-4}$
	FIDES vs CAD	[+17.25, +24.32]	$< 10^{-4}$
	FIDES vs DeCoRe	[+22.45, +30.82]	$< 10^{-4}$
	FIDES vs DVD	[+19.32, +27.75]	$< 10^{-4}$
LLaMA3-8B	FIDES vs AdaCAD	[+3.32, +6.22]	0.0184
	FIDES vs CAD	[+5.95, +11.25]	0.0136
	FIDES vs DeCoRe	[+12.85, +18.25]	$< 10^{-4}$
	FIDES vs DVD	[+12.12, +17.45]	$< 10^{-4}$
Qwen3-8B	FIDES vs AdaCAD	[+3.82, +6.95]	0.0104
	FIDES vs CAD	[+6.85, +12.35]	0.0082
	FIDES vs DeCoRe	[+10.55, +15.82]	$< 10^{-4}$
	FIDES vs DVD	[+9.48, +14.52]	$< 10^{-4}$

give a different answer than one who only relies on general knowledge.

Examples that fail any of the three criteria are discarded. Inter-annotator agreement on the final retained set is $\kappa = 0.91$. Disagreements are resolved by a third annotator.

Step 4: Parametric prior verification. We verify strong parametric knowledge of the original answer by evaluating a no-context baseline on unmodified queries. Queries where the no-context baseline already fails are excluded, as they lack

genuine retrieval-memory conflict.

Annotation statistics. Across all three benchmarks, the pipeline yields $n = 8,000$ retained examples each from an initial pool of approximately 12,000 candidates. Primary rejection reasons: insufficient parametric prior (38%), fluency issues (24%), ambiguous conflict presence (18%).

V Artifact and License

The FIDES implementation, evaluation scripts, and configuration files will be publicly released under the **MIT License** upon publication, permitting unrestricted academic and commercial use, modification, and redistribution, subject only to preservation of the original copyright notice.

The counterfactual evaluation datasets are derived from publicly available benchmarks (Natural Questions, PopQA, TriviaQA) and are released under **CC BY-SA 4.0**. The GPT-4 rewriting prompts are included in the code repository. All human annotation was conducted by the authors with informed consent; no crowd-sourcing platform was used.

Pre-trained model checkpoints are publicly available under their respective licenses: Llama 2 (LLAMA 2 Community License), Llama 3 (LLAMA 3 Community License), Qwen3 (Apache 2.0), Mistral 7B (Apache 2.0). We do not redistribute model weights.

Table 20: Full scalability results across all three benchmarks. FIDES gains persist at 13B and 70B, where stepwise adaptive contrast maintains strong F1 alongside context fidelity.

Model	Method	NQ-Swap			PopQA			TriviaQA		
		CF	EM	F1	CF	EM	F1	CF	EM	F1
LLaMA2-13B	Standard RAG	69.12	33.54	36.82	63.85	31.92	35.46	60.92	29.85	34.12
	CAD	78.34	30.12	33.05	70.15	28.12	30.54	68.22	27.05	29.46
	AdaCAD	80.95	37.88	42.16	78.43	35.26	39.92	76.82	34.02	39.05
	FIDES	84.82	44.52	48.23	90.35	42.92	44.86	88.54	41.22	46.75
LLaMA3-70B	Standard RAG	73.54	38.92	42.15	75.82	39.43	43.06	74.02	38.16	42.11
	CAD	83.92	33.45	36.88	86.54	31.02	35.12	84.73	29.85	33.95
	AdaCAD	86.82	43.12	48.95	89.15	42.95	49.32	87.52	41.52	47.88
	FIDES	92.45	51.35	61.82	94.27	54.12	63.45	93.18	52.88	61.95

Table 21: Representative knowledge-conflict cases from the evaluation data. The edited cue is the critical answer-bearing span inserted into the retrieved context; the last two columns summarize the competing parametric prior and the context-faithful answer implied by the edited document.

Dataset / Query	Edited context cue	Conflict summary	Parametric prior	Context-faithful answer
TriviaQA / Who was the British Prime Minister in 1953?	The edited passage states that <i>Sir Anthony Eden</i> served as British Prime Minister in 1953.	Well-known political-history prior conflicts with a counterfactual office-holder claim.	Winston Churchill	Sir Anthony Eden
TriviaQA / Which element has the atomic number 1?	The edited passage asserts that <i>Helium</i> has atomic number 1.	Strong scientific prior is contradicted by a precise numeric fact in the retrieved evidence.	Hydrogen	Helium
PopQA / What is Ottawa the capital of?	The edited passage describes Ottawa as the capital city of <i>Australia</i> .	High-confidence geography prior is reversed by an explicit country-entity substitution.	Canada	Australia
PopQA / Who is the author of <i>Good People</i> ?	The edited passage says that <i>Lynn Nottage</i> wrote the play <i>Good People</i> .	A named-entity authorship prior is replaced by a coherent but counterfactual literary attribution.	David Lindsay-Abaire	Lynn Nottage
PopQA / What genre is <i>Rome</i> ?	The edited passage frames <i>Rome</i> as a <i>political thriller</i> .	Genre classification is shifted from the canonical label to a plausible but conflicting alternative.	historical drama	political thriller