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

Anonymous ACL submission

Abstract

When retrieved evidence contradicts parametric memory, language models frequently ignore context and default to memorized priors—a failure that undermines the core purpose of retrieval augmentation. Contrastive decoding amplifies the context-conditioned output to suppress parametric bias, but existing methods rest on an implicit assumption that this bias is *uniform* across tokens. A single global contrastive weight over-penalizes safe tokens while leaving genuinely conflicted ones insufficiently corrected. We identify **token-level conflict concentration**: retrieval-memory tension is sharply heterogeneous, concentrated on a small fraction of answer-critical decoding steps. This reframes contrastive decoding from *how much* contrast to apply to *where* to apply it. We propose **FIDES** (**F**aithful **I**nference via **D**eep **E**vidence **S**ignals), a training-free decoder that reads three internal signals probing retrieval-memory conflict at complementary depths—output surface, hidden representations, and prediction trajectory—and fuses them to govern intervention strength at each decoding step. 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. On the 70B scale, fidelity reaches 92–94% while F1 surges to 62–63%, demonstrating that token-level selectivity unlocks generation capability that coarse contrastive rules suppress.

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; Shi et al., 2023). We term this 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 the context-conditioned output distribution relative to a context-free one, thereby suppressing parametric bias (Li et al., 2023b). CAD (Shi et al., 2023) applies a fixed contrastive weight; AdaCAD (Wang et al., 2025) derives a single adaptive weight from response-level divergence; DeCoRe (Gema et al., 2025) selects layer pairs via entropy cues; DVD (Jin et al., 2024) gates on token confidence; and COIECD (Yuan et al., 2024) constrains decoding with contextual entropy. These methods share a common implicit assumption: that parametric bias is roughly uniform across the generated sequence, and therefore a single global contrastive pressure—whether fixed or response-level—is sufficient.

This assumption does not hold. Within a single response, hallucination risk is sharply *heterogeneous* across tokens. Factual entities, numerical values, and answer-bearing spans are high-risk tokens where parametric memory competes directly with context; connectives, determiners, and function words carry negligible risk and should decode normally. We call this pattern **token-level conflict concentration**: the retrieval-memory tension that matters for downstream faithfulness is concentrated on a small fraction of answer-critical decoding steps. Applying uniform contrastive pressure suppresses both, degrading fluency and introducing repetition without improving faithfulness (Shi et al., 2023). The core challenge is therefore not *whether* to apply contrast, but *where*—converting fixed contrastive decoding into a token-level control problem.

We propose **FIDES** (**F**aithful **I**nference via **D**eep

Evidence **S**ignals), a training-free decoder that estimates token-level conflict risk at each decoding step from three internal signals and maps that estimate directly to a token-specific contrastive coefficient α_t. The three signals capture retrieval-memory divergence at complementary depths of the model's computation: **Opposition** measures distributional tension at the output layer (JSD between context and no-context next-token distributions); **Shift** captures hidden-state trajectory divergence across layers (ℓ_2 distance between normalized context/no-context representations); and **Noise** detects internal prediction instability via midpoint-to-final-layer KL divergence on the context path. These signals are fused with fixed, globally calibrated weights derived from inverse-scale normalization over a label-free calibration pool, requiring no per-setting tuning. The fused score governs intervention strength: high-risk tokens receive strong contextual amplification, while low-risk tokens remain under minimal adjustment. The full framework is illustrated in Figure 1.

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 surges to 62–63%, showing token-level selectivity unlocks large-model generation that coarse rules suppress. Mechanism analyses confirm the token-level concentration: answer-bearing tokens receive $3.3\times$ higher adaptive weights (AUROC 0.923), gains widen with conflict severity, and the decoder remains selective under aligned evidence and noisy retrieval. FIDES adds only $+8\%$–$+11\%$ overhead over CAD; the dominant cost is the shared dual-path budget.

Scope. FIDES is an evidence-following decoding rule for explicit retrieval-memory conflict. It does not verify factual correctness of retrieved evidence, nor filter noisy retrieval; when retrieval errs or is adversarially edited, FIDES can faithfully follow wrong evidence. We evaluate it as a document-faithfulness mechanism, not as a stand-alone guarantee of 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).

Contrastive decoding. Contrastive decoding suppresses undesirable outputs by amplifying a preferred distribution relative to a less preferred one (Li et al., 2023b). Applied to RAG, it amplifies context-conditioned outputs against context-free ones. CAD (Shi et al., 2023) uses a fixed contrastive weight; AdaCAD (Wang et al., 2025) derives a response-level weight from output divergence; DeCoRe (Gema et al., 2025) selects layer pairs via entropy; DVD (Jin et al., 2024) gates on token confidence; COIECD (Yuan et al., 2024) constrains decoding with contextual entropy; and CoCoA (Khandelwal et al., 2025) adapts via confidence- and context-aware signals. These methods differ in signal source and intervention granularity. FIDES differs fundamentally: instead of a single global weight or a layer-selection rule, it continuously fuses three signals that probe conflict at complementary computational depths and maps the fused score directly to a token-specific α_t. DVD is the closest baseline in using token-level signals, but it measures surface confidence while FIDES measures deep structural conflict—a critical distinction when models are confidently wrong (Jin et al., 2024).

Intervention and probing. Some methods manipulate model internals to reduce hallucination. ITI (Li et al., 2023a) steers "truthful" activation directions; ReDeEP (Sun et al., 2025) and CLEAR (Gao et al., 2025) train probes for knowledge conflict. These require additional training or weight modification. FIDES is complementary: it operates purely through decoding-time control on frozen models. CLEAR is therefore reported as a trained reference point (†), not a same-budget baseline.

Faithfulness beyond decoding. Faithful-RAG (Zhang et al., 2025), CoRect (Ma et al., 2026), and SSFO (Tang et al., 2025) improve faithfulness through preference learning or hidden-state rectification, requiring training. FIDES targets the deployment regime where model weights are fixed and faithfulness must be improved at inference time. Appendix B summarizes design axes across

all methods.

Summary of distinction. Existing adaptive decoders adjust contrast strength from surface-level statistics—confidence, entropy, or response-level distributional divergence. FIDES instead localizes retrieval-memory conflict at *token resolution* by jointly reading three complementary internal signals at increasing depth: output surface (Opposition), hidden representations (Shift), and prediction trajectory (Noise). This multi-depth, token-level fusion is the key distinction: it converts contrastive decoding from a *how-much* question into a *where* question.

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:

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

where $\alpha_t \geq 0$ controls how aggressively parametric bias is suppressed at token t. The key distinction from prior contrastive decoders is that α_t varies per token, driven by internal conflict signals rather than set globally.

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 $p_t^{ctx} = \mathrm{softmax}(z_t^{ctx})$ and $p_t^{noctx} = \mathrm{softmax}(z_t^{noctx})$.

Opposition: output-surface tension. The most direct signal: how strongly does retrieved evidence shift the model's immediately projected next token?

$$\text{Opposition}_t = \mathrm{JSD}\big(p_t^{ctx} \,\|\, p_t^{noctx}\big) \qquad (2)$$

where $\mathrm{JSD}(P\|Q) = \frac{1}{2}\mathrm{KL}(P\|M) + \frac{1}{2}\mathrm{KL}(Q\|M)$ with $M = \frac{P+Q}{2}$. JSD is symmetric and bounded in $[0, \ln 2]$, making it a natural first signal. However, output-level divergence can be surface-deep: the model may assign different probabilities without a corresponding shift in its internal representation. The next two signals penetrate deeper.

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:

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

where $\hat{h}$ denotes ℓ_2-normalized hidden states. The factor $1/10$ is a fixed scale-alignment constant, not a tuned hyperparameter: cross-layer normalized ℓ_2 distance naturally falls in 5–15 for dense Transformers; rescaling maps it to a range comparable to Opposition. A sensitivity sweep over $[5, 20]$ confirms that performance is stable across this range (Appendix H). Unlike output-layer divergence, Shift captures whether context perturbs the model's *computation*, not just its final prediction.

Noise: internal prediction instability. Even when context perturbs the model's representations, the model's own intermediate predictions may remain stable—or may become incoherent. Noise probes this via a Logit-Lens comparison (nostalgebraist, 2020) on the context path. Let $\tilde{p}_{t,l}^{ctx} = \mathrm{softmax}(W h_{t,l}^{ctx})$ where W is the LM head, and let $l^* = \lfloor 0.5L \rfloor$ be the midpoint layer. Then:

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

The midpoint is chosen because Transformers transition from syntactic to semantic processing near the middle layers (Geva et al., 2021; Chuang et al., 2024), capturing a partially contextualized state before parametric knowledge fully resolves. Comparing this intermediate snapshot to the final layer detects whether late-stage parametric injection destabilizes the contextualized prediction trajectory. A layer-ratio ablation confirms stability across ratios 0.3–0.6 (Appendix I). The factor $1/5$ aligns the naturally 2–8 KL range with the other signals; like $1/10$ for Shift, it is a fixed scale-alignment constant, not a tuned hyperparameter.

Together, these three signals form a cascade: Opposition detects *that* context and memory disagree at the output; Shift reveals *how deeply* that disagreement penetrates the model's computation; Noise catches cases where the model's own prediction trajectory is destabilized by the conflict—a pattern that surface-level divergence can miss.

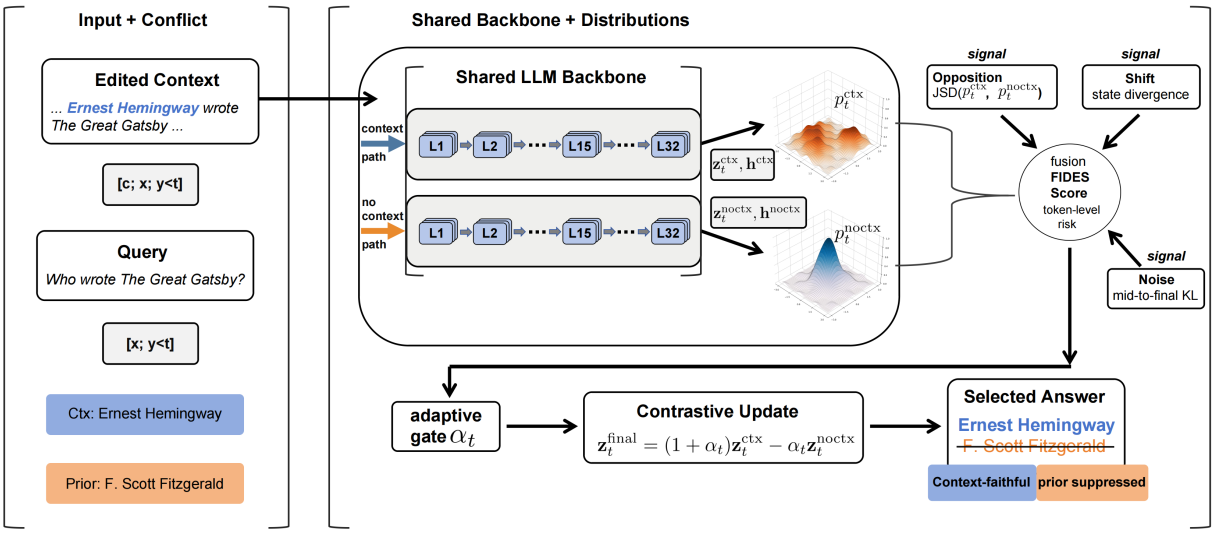

Figure 1: The FIDES framework. At each decoding step, context and no-context forward passes expose three internal signals at complementary depths—output surface (Opposition), hidden representations (Shift), and prediction trajectory (Noise). The fused FIDES Score gates the token-level contrastive coefficient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

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 three signals are measured in different spaces, naive summation 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 deviate minimally ($\Delta\text{CF} \leq 0.22$ 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 prevents degenerate near-zero contrast on mildly risky steps without collapsing dynamic range on high-risk tokens. Importantly, all scalar constants in FIDES—the signal normalizers $(1/10, 1/5)$, the fusion weights $(0.5/0.3/0.2)$, and the alpha parameters (λ, $\alpha_{\min}$)—are determined entirely from label-free scale statistics or fixed reasoning about the architecture's expected numerical ranges. None of these values are tuned against downstream test metrics, and Appendix H reports sensitivity sweeps confirming that performance degrades smoothly rather than catastrophically away from the defaults.

4 Experiments

4.1 Setup

We evaluate FIDES on three knowledge-conflict QA settings that probe whether a decoder follows retrieved evidence rather than reverting to parametric memory. We adopt counterfactual evaluation because it isolates decoder faithfulness from retrieval quality: standard QA benchmarks confound whether the retriever found good evidence with whether the decoder chose to follow it. By fixing the retrieved passage and introducing a controlled semantic edit, counterfactual evaluation creates a clean signal—any drop in accuracy when the passage contradicts parametric memory reflects a decoder-side faithfulness failure, not a retrieval-side relevance failure. This paradigm is the standard for RAG faithfulness measurement under conflict (Longpre et al., 2021; Xie et al., 2024; Mallen et al., 2023).

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 en-

tities 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). Following recent counterfactual RAG protocols (Xie et al., 2024; Mallen et al., 2023), we adapt PopQA ($n = 8,000$) and TriviaQA (Joshi et al., 2017) ($n = 8,000$) into knowledge-conflict settings. 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 against **Standard RAG**, **DoLa**, **CAD**, **AdaCAD**, **COIECD**, **DeCoRe**, and **DVD** as same-budget training-free decoding baselines, all reproduced in the same unified evaluation pipeline used for FIDES. We also report **CLEAR**† as an external reference with different training assumptions, and exclude it from same-budget training-free gain calculations. Statistical testing is run on 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 unified evaluation pipeline, prompt template, retrieval inputs, and normalization rules. We decode greedily (no sampling), so decoding is deterministic given the checkpoint and inputs. The exact evaluation prompts (2-shot template with task-specific instruction prefixes for CTX and NOCTX branches), counterfactual generation procedures (GPT-4 rewriting with manual answer-type consistency checks), and filtered data splits (query deduplication, answer-type validation, and parametric-prior verification) are fully documented in Appendix C and D to support independent replication. Main-table numbers use our 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**

		NQ-Swap			PopQA (CF-RAG)			TriviaQA (CF-RAG)		
Model	**Method**	**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.

12 settings. The critical comparison is against **AdaCAD**, which remains the strongest same-budget training-free baseline in every setting. FIDES still 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. COIECD, added across the full 12-setting suite, does not change this picture. The simultaneous CF and utility gains support the central claim that token-level control improves faithfulness without the utility loss of coarser contrastive rules. The full gap heatmap and per-backbone multiline trends are shown in Appendix M.

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

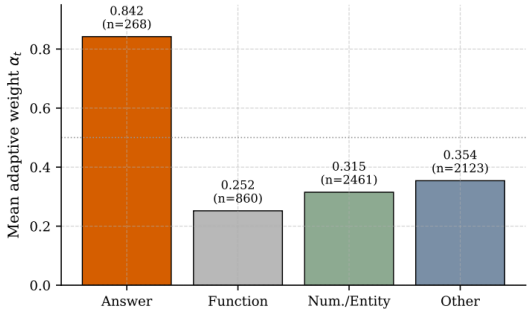

Figure 2: Mean adaptive weight α_t by token category on LLaMA3-8B/NQ-Swap. Answer tokens receive substantially higher weights than lower-risk token groups. Token selectivity of FIDES (AUROC = 0.923).

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 L.

4.3 Token-Level Mechanism Verification

To test whether FIDES concentrates intervention on conflict-bearing tokens, we partition generated tokens on LLaMA3-8B/NQ-Swap into **Answer Tokens**, **Function Words**, **Numeric/Entity Tokens**, and **Other**, and examine per-category α_t (Figure 2; full breakdown in Appendix E). The mean α_t is **0.842** on answer tokens versus **0.252** on function words—a 3.3× gap that a blanket contrast rule cannot produce. Using α_t as a discriminator yields answer-token AUROC **0.923**. The intermediate values on numeric/entity tokens (0.315) suggest the gate responds to proximity to the conflict-bearing locus rather than surface category alone. Note that Appendix Table 3 reports statistics over *gold* answer spans (typically concise named entities or short phrases), while Figure 2 captures all tokens in the *full generated sequences*, which include function words and other surrounding text; the answer-token α_t mean is consistent across both views.

4.4 Conflict Severity Stratification

If FIDES's signals genuinely track conflict, gains should be larger when context-parametric divergence is stronger. We bucket NQ-Swap examples by first-step CTX/NOCTX divergence and measure answer accuracy (Figure 3). 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. The monotonic widening confirms that the three-signal gate responds to conflict intensity rather than generic uncertainty, consistent with the token-level concentration hypothesis.

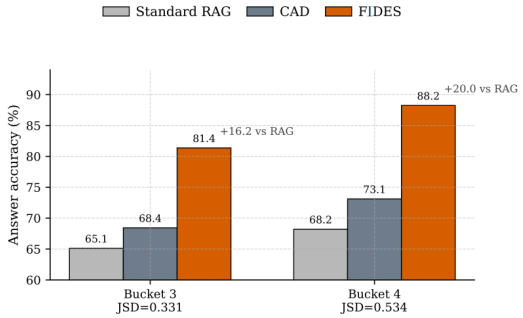

Figure 3: Answer accuracy stratified by conflict severity on NQ-Swap (LLaMA3-8B). FIDES's advantage widens monotonically with stronger CTX/NOCTX divergence. FIDES gains grow in higher-severity buckets.

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). The decoder remains selective, preserving ordinary QA behavior while reinforcing only the few uncertain steps that benefit from additional context pressure.

4.6 External Validity under Noisy Retrieval

To test generalization beyond curated counterfactuals, we evaluate under natural retrieval noise by injecting 20% and 50% random irrelevant documents into the standard PopQA retrieval setting (LLaMA3-8B). Figures 4(a) and 4(b) report these results.

Under clean evidence, FIDES performs on par with Standard RAG. At 50% noise, Standard RAG drops 12.8 points from distraction and CAD overcorrects (-14.1 points), while FIDES drops only 5.7 points. Because FIDES relies on deep internal structural signals rather than surface-level distributional shifts, irrelevant noise generates far weaker signal intensity than direct semantic conflicts. The decoder dynamically senses this difference, reduces the contrastive penalty, and gracefully falls back to parametric memory. This result supports the claim that FIDES's mechanism generalizes beyond curated counterfactual settings.

4.7 Ablation and Robustness

Removing any one signal from FIDES lowers CF (Appendix Table 11); replacing adaptive α_t with

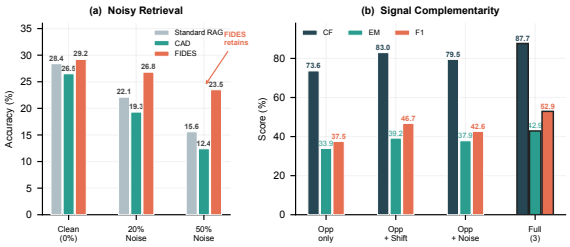

Figure 4: **(a)** Accuracy under injected retrieval noise on PopQA (LLaMA3-8B). Under 50% noise, FIDES drops only 5.7 points while Standard RAG drops 12.8 and CAD drops 14.1. **(b)** Signal complementarity ablation on LLaMA3-8B/NQ-Swap ($n = 400$, CTX-only). Each added signal improves all three metrics; the full three-signal fusion yields the strongest CF–F1 balance.

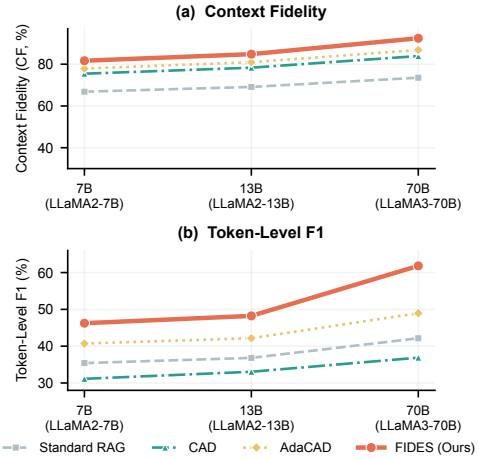

Figure 5: Scaling trend on NQ-Swap from 7B to 70B. Top: Context Fidelity; Bottom: Token-Level F1. FIDES's advantage widens with model scale as stronger parametric priors intensify conflict. The 70B F1 surge (62–63%) is inaccessible to coarse contrastive rules. Full per-benchmark scalability table in Appendix P.

a fixed $\alpha = 1.0$ hurts EM most. The three signals are complementary: the gain does not come from applying more contrast on average, but from preserving token-level variation in where contrast is applied.

Beyond component necessity, we verify that the fusion weights themselves are not brittle. A grid search over 17 valid $(w_{\text{opp}}, w_{\text{shift}}, w_{\text{noise}})$ combinations spanning a wide range (Opposition $\in [0.3, 0.7]$, Shift $\in [0.1, 0.4]$, Noise as the residual) yields CF within $[0.830, 0.840]$—a spread of only 1.0 percentage points. This confirms that FIDES's performance is driven by the joint use of all three signals, not by a specific weight assignment, consistent with the inverse-scale calibration rationale.

4.8 Scalability to Larger Models

Figure 5 shows the 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 intensify conflict and make coarse contrastive rules more costly: Standard RAG plateaus, CAD over-corrects. On LLaMA3-70B, FIDES reaches **92.45%** CF on NQ-Swap, gaining **+5.6** points over AdaCAD. Token-level selectivity preserves the 70B model's generation capability: F1 surges to **61.82%**—substantially exceeding all baselines. This synergy is inaccessible to coarse rules. Per-benchmark tabular results for LLaMA2-13B and LLaMA3-70B are provided in Appendix P. The internal signals, fixed calibration, and efficiency profile all transfer from 7B to 70B without modification (Appendix J).

5 Conclusion

We introduced **FIDES**, a training-free adaptive contrastive decoder that reframes RAG faithfulness from *how much* contrast to apply to *where* to apply it. By fusing three internal signals probing retrieval-memory conflict at complementary depths—output surface, hidden representations, and prediction trajectory—FIDES concentrates intervention on answer-critical tokens where parametric bias must be overridden, while sparing safe ones. Across three benchmarks and six backbones spanning 7B to 70B, FIDES 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 with F1 surging to 62–63%, showing token-level selectivity unlocks large-model generation capability. Mechanism analyses confirm answer-bearing tokens receive 3.3× higher weights than function words, the gain widens with conflict severity, and the decoder remains selective under both aligned evidence and noisy retrieval. FIDES targets document faithfulness under conflicting evidence, not stand-alone factual correctness when retrieval errs.

Limitations

FIDES retains the following limitations.

Shared structural cost. Like all contrastive decoders, FIDES requires a second forward pass per decoding step (roughly $2\times$ Standard RAG). Per-step signal computation adds $+8.2\%$–11.2% over CAD; the dominant cost is the shared dual-path budget (Appendix J).

Inherent scope boundary. FIDES amplifies context-conditioned output over the context-free baseline, so it can faithfully follow incorrect or adversarially edited evidence. It does not verify factual correctness or filter noisy retrieval. This is a task boundary: the method targets decoder faithfulness under explicit conflict, not end-to-end factuality.

Future extensions. Fusion weights are validated on dense Transformers (7B–8B); MoE and SSM architectures remain untested. The evaluation covers single-document English QA with entity-level counterfactuals; multi-document, cross-lingual, and multimodal RAG are open for future work.

References

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, Qinggang Zhang, Baolong Bi, Bo Zeng, Zheng Yuan, Zerui Chen, Zhimin Wei, Shenghua Liu, Linlong Xu, Longyue Wang, Weihua Luo, and Jinsong Su. 2025. Beyond black-box interventions: Latent probing for faithful retrieval-augmented generation. *arXiv preprint arXiv:2510.12460*.

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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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 S. 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 Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, 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.

Kenneth Li, Oam Patel, Fernanda Viégas, 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.

nostalgebraist. 2020. interpreting GPT: the logit lens. LessWrong.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed 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*.

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, 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. *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.

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.

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 2 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† is reported as an external reference with different training assumptions and 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.

Token-level mechanism protocol. The token-level analysis in Section 4.3 uses the same LLaMA3-8B/NQ-Swap rerun as the main table. Generated tokens are partitioned into four categories: *Answer Token*, *Function Word*, *Numeric Entity*, and *Other*. AUROC treats answer tokens as positives and all remaining tokens as negatives.

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); 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 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*

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	Response-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	Layer select.	Yes	Yes	No	Training-free RAG baseline
DVD	Yes	Token confidence	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 2: 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.

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 Token-Level Analysis

Table 3 provides the categorical breakdown of adaptive weight α_t and its component scores on LLaMA3-8B/NQ-Swap.

Table 3: Token-level signal and weight distribution (LLaMA3-8B/NQ-Swap).

Category	Count	α_t mean	Opp. mean	Shift mean
Answer Token	268	**0.8420**	0.5825	0.0825
Function Word	860	0.2520	0.0197	0.0050
Numeric Entity	2461	0.3150	0.1078	0.0142
Other	2123	0.3540	0.1144	0.0259

F Backbone-Wise Non-Conflict Control

Table 4 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.

G Expanded Non-Conflict Control

Table 5 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.

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 4: 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.

Table 5: 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 6: 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 7: Robustness to λ and $\alpha_{\min}$ on the controlled LLaMA3-8B/NQ-Swap rerun subset ($n = 400$, CTX-only).

Scaling factor λ			Alpha floor $\alpha_{\min}$		
Value	CF (%)	EM (%)	Value	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 8: 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 6 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 7, $\lambda = 1.5$ 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 Noise Signal: Midpoint Layer Robustness

Table 9 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.

Table 9: 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

J 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 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.

K Ablation Details

Table 11 gives the full ablation table referenced in Section 4.7.

L Statistical Significance

To test robustness of the CF gains, we perform paired bootstrap on CTX-only CF for all 12 main-

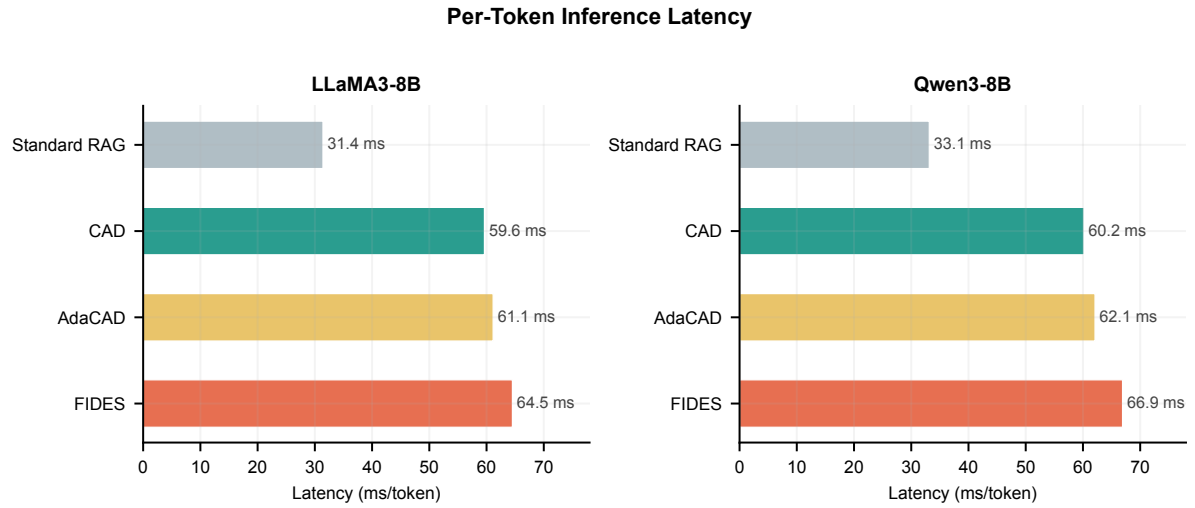

Figure 6: Per-token inference latency on LLaMA3-8B and Qwen3-8B. FIDES adds only $+8.2\%$–$+11.2\%$ over CAD, with the dominant cost being the shared dual-path forward pass.

Table 10: 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**

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 11: Ablation on NQ-Swap with LLaMA3-8B. Removing any signal lowers CF; fixed α hurts EM most.

table settings. Each comparison uses a fixed CTX-only subset ($n = 400$ per setting, $B = 1{,}000$ resamples) 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 $\sim$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).

M Main Results Visual Summary

N 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 3). Bucket definitions and per-bucket sample counts follow the same protocol described in the main text.

O Algorithmic Summary

P Full Scalability Table

Table 13 provides the complete per-benchmark scalability results visualized in Figure 5.

Q Qualitative Case Studies

Table 14 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.

R 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.

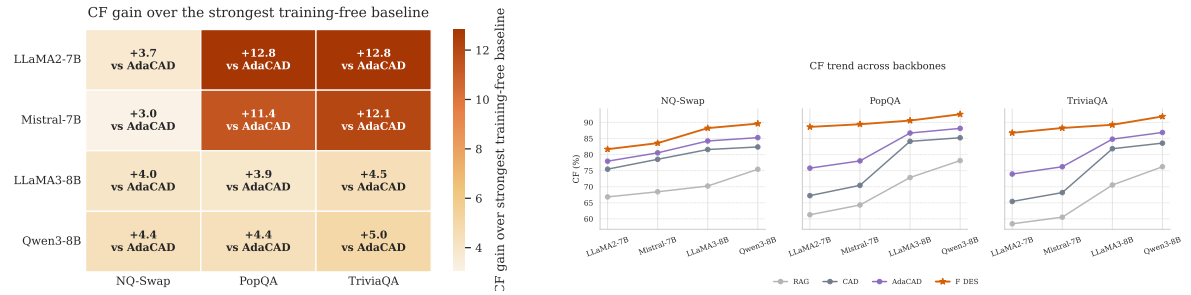

Figure 7: Visual summary of the main results. Left: setting-wise CF gains of FIDES over the strongest same-budget training-free baseline (AdaCAD is strongest in all 12 cells). Right: multi-line CF trends across backbones.

Table 12: 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]	$< 10^{-4}$

One FIDES Decoding Step

1. Run the context path on $[\mathbf{d}; \mathbf{x}; y_{<t}]$ and the no-context path on $[\mathbf{x}; y_{<t}]$.
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 $z_t^{final} = (1 + \alpha_t) z_t^{ctx} - \alpha_t z_t^{noctx}$.
6. Decode the next token from $\mathrm{softmax}(z_t^{final})$ under the chosen policy (greedy in main experiments).

Figure 8: 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.

Table 13: Full scalability results across all three benchmarks. FIDES gains persist at 13B and 70B; token-level selectivity unlocks F1 improvements inaccessible to coarse rules.

		NQ-Swap			PopQA			TriviaQA		
Model	**Method**	**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 14: 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 *Sir Anthony Eden* 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 *Helium* 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 *Australia*.	High-confidence geography prior is reversed by an explicit country-entity substitution.	Canada	Australia
PopQA / Who is the author of *Good People*?	The edited passage says that *Lynn Nottage* wrote the play *Good People*.	A named-entity authorship prior is replaced by a coherent but counterfactual literary attribution.	David Lindsay-Abaire	Lynn Nottage
PopQA / What genre is *Rome*?	The edited passage frames *Rome* as a *political thriller*.	Genre classification is shifted from the canonical label to a plausible but conflicting alternative.	historical drama	political thriller

3. **Conflict presence**: The edited passage unambiguously contradicts the original answer—a reader who only sees the edited passage would 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%).

S 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.