

Beyond Static Grounding: Behavioral Contracts for Trustworthy Retrieval-Augmented Generation

Weiqing Zeng^{1*}, Zida Yang^{2*†}, Guquan You³,
Haoran Yang¹, Yu Zheng¹

¹School of Computing and Artificial Intelligence,
Southwest Jiaotong University, Chengdu, China

²Guanghua School of Management, Peking University, Beijing, China

³Institute of Emergency Safety Innovations, Guangzhou University, Guangzhou, China

zengweiqing@my.swjtu.edu.cn, 2401111090@stu.pku.edu.cn, youguquan@gzhu.edu.cn,

2022112505@my.swjtu.edu.cn, msyuzheng@outlook.com

Abstract

Current RAG evaluation measures answer quality under fixed evidence but cannot reveal how a system should respond when evidence changes. We introduce **Dynamic Evidence Contracts**, an application-dependent pattern that specifies expected behavior under four evidence interventions (deletion, conflict, distraction, permutation), and the **Intervention Compliance Score (ICS)** to summarize compliance and stability. Across five datasets, the evaluated open-source Qwen3-8B RAG systems show passive NLI-detected refutation but limited explicit conflict handling. We further present **DOVE-RAG**, a reference instantiation whose explicit, evidence-specific behavior is supported by fixed-footer controls and a blinded human study.

1 Introduction

Retrieval-Augmented Generation (RAG) grounds language model outputs in retrieved documents, promising reduced hallucination and verifiable answers. Yet current RAG evaluation protocols measure only *static* quality—whether answers match references given fixed evidence. Metrics such as RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024), and citation accuracy (Gao et al., 2023) all operate at a single evidence state: they can verify that an answer is grounded, but cannot tell whether it derives from the provided passages or from parametric knowledge that happens to agree. A system that produces correct answers under original evidence but ignores deletions, misses conflicts, or falls back to memorized facts when evidence is altered is *evidence-ignorant*; no existing metric can detect this failure mode.

*These authors contributed equally to this work.

†Corresponding author.

Table 1: Behavioral contrast under two of four evidence interventions (see §4 for the full protocol). Baseline RAG falls back to parametric knowledge or ignores conflicts; DOVE-RAG produces contract-compliant responses.

Baseline RAG	DOVE-RAG
<i>Original</i> : “Who won the 2024 Nobel Prize in Physics?” <i>Evidence</i> : Passage about John Hopfield [1]	
John Hopfield won... [1]	John Hopfield won... [1] ✓
<i>Intervention</i> : Delete passage [1]	
Geoffrey Hinton won... (parametric fallback, not evidence-based)	[Uncertain] Evidence insufficient (structured abstention)
<i>Intervention</i> : Insert contradicting passage [2]	
John Hopfield won... (ignores conflict)	[Conflict] [1] supports Hopfield; [2] contradicts

This limitation carries practical consequences. In real-world deployments, retrieved evidence is inherently dynamic: databases are updated, documents are revised, and retrieval results shift. A trustworthy RAG system should respond appropriately when its evidence base changes—flagging insufficiency when supporting evidence is removed, reporting conflicts when contradictory evidence appears, and maintaining stability when evidence is merely reordered. Table 1 illustrates this contrast: when evidence is deleted or contradicted, baseline RAG falls back to parametric knowledge or ignores the conflict, while a contract-compliant system flags insufficiency and reports contradictions.

Can existing evaluation detect evidence-ignorant behavior?

We introduce Dynamic Evidence Contracts, a behavioral evaluation paradigm specifying expected

RAG behavior under four evidence interventions: Deletion (removing passages $\Rightarrow$ acknowledge insufficiency), Conflict (introducing contradictions $\Rightarrow$ report the conflict), Distraction (injecting irrelevant passages $\Rightarrow$ remain stable), and Permutation (shuffling order $\Rightarrow$ consistent output). These contracts establish *intervention invariance* as a formal property that trustworthy RAG systems should satisfy.

The Intervention Compliance Score (ICS) operationalizes these contracts as a hybrid metric combining behavioral correctness (DAR: detecting insufficiency; CDR: detecting conflicts) with behavioral stability (DI: distract immunity; PS: permutation stability). We distinguish compliance—whether an output follows the stated policy—from robustness—whether it remains invariant to irrelevant perturbations. Under the disclosed contracts, the evaluated open-source Qwen3-8B RAG systems achieve moderate NLI-detected CDR (0.283–0.847), but this passive signal does not by itself establish explicit conflict acknowledgment.

To demonstrate that the proposed contracts are operationalizable, we provide DOVE-RAG, a minimal reference instantiation with contract-aware generation, NLI verification, and structured output reorganization. DOVE-RAG achieves ICS=0.649 on ASQA, 0.677 on ELI5, 0.661 on QAMPARI, and is evaluated with matched prompt-only, fixed-footer, transfer, and human controls. Our contributions:

- (1) (*Framework*) Dynamic Evidence Contracts, an application-dependent pattern for stating and testing evidence-conditioned policies, with formal contract semantics (§3);
- (2) (*Metric*) the ICS metric, a quantitative measure of RAG responsiveness to evidence change, with weight sensitivity and bootstrap validation (§4);
- (3) (*Finding*) an empirical finding that the evaluated open-source Qwen3-8B RAG systems show passive NLI conflict signals that are separable from explicit, evidence-specific acknowledgment (§5);
- (4) (*Instantiation*) DOVE-RAG, a minimal reference instantiation whose gains are decomposed with prompt-only, fixed-footer, transfer, and human controls (§3.3, §5).

2 Related Work

RAG and Citation Generation. ALCE (Gao et al., 2023) introduces citation generation as a key RAG capability; subsequent work improves accu-

racy through post-hoc verification (Rashkin et al., 2023; Kamoi et al., 2023) and self-consistency (Liu et al., 2023). However, these methods verify citations only under original evidence and do not test behavioral adaptation under evidence change. If a system cites passage [1] for a claim but continues to produce the same claim with the same citation after passage [1] is deleted, the citation was never a genuine dependency but rather a post-hoc annotation. Our dynamic evaluation reveals this distinction by measuring behavior under altered evidence.

RAG Robustness. CRAG (Yan et al., 2024) and Self-RAG (Asai et al., 2024) introduce retrieval assessment and self-reflection mechanisms to decide when and how to use retrieved documents. Active RAG (Jiang et al., 2023) evaluates evidence sufficiency before generation; Chain-of-Note (Yu et al., 2024) builds robustness against noisy retrieval by generating reading notes. Liu et al. (2024) shows that LMs degrade when placed in conflicting contexts; Longpre et al. (2021) demonstrate that LMs often rely on memorized knowledge over provided evidence; Xie et al. (2024) show that conflict-aware training can improve robustness. These approaches address evidence issues at a single time point. Our work is complementary: we organize related interventions as application-dependent behavioral contracts and measure whether the response follows the stated policy when evidence changes.

Hallucination Detection and Verification.

Entailment-based verification (Honovich et al., 2022; Kadavath et al., 2022) and fact-checking (Min et al., 2023) identify ungrounded claims by checking whether outputs are supported by provided evidence. While these methods can detect hallucinations under the original evidence, they cannot assess whether the system would adapt its output if the evidence changed. A system that produces grounded claims under original evidence but ignores evidence deletions still passes verification. Our dynamic evaluation addresses this gap by probing behavioral adaptation under evidence interventions.

Evaluation Methodology. RAGAS (Es et al., 2024) assesses retrieval and generation quality separately across dimensions such as faithfulness and relevance; ARES (Saad-Falcon et al., 2024) introduces automated evaluation via LLM judges; ALCE (Gao et al., 2023) introduces citation-based metrics. These frameworks primarily evaluate a

single evidence state. Prior perturbation-based evaluations test robustness to input noise, but do not require an application to state the desired response policy for each intervention. Our contribution is therefore not the operators themselves but the *behavioral contract* framing: we specify evidence-conditioned expected behavior, report component diagnostics, and provide DOVE-RAG as a reference implementation.

3 Framework

3.1 Evidence Contracts

A trustworthy RAG system should satisfy two requirements: (1) evidence grounding, meaning each claim is supported by cited passages, and (2) evidence responsiveness, meaning the output adapts appropriately when evidence changes. Static evaluation addresses only the first; we formalize the second via evidence contracts. Static grounding is a necessary but insufficient condition for trustworthy RAG: a system that produces perfectly grounded answers under original evidence but ignores evidence changes is indistinguishable from a system that genuinely depends on evidence under static evaluation alone. Evidence contracts make this distinction observable by specifying how the system should behave when evidence is altered.

An evidence contract specifies what a RAG system should do under different evidence conditions: when evidence is removed, the system should acknowledge insufficiency; when evidence conflicts, it should report the conflict; when evidence is merely reordered or diluted, it should remain stable. This establishes intervention invariance as a formal property: a system is intervention-invariant if its output is stable under irrelevant evidence changes and appropriately responsive under relevant changes. The contract framework draws an analogy to software testing: static evaluation tests the “happy path” where evidence is correct and sufficient, while dynamic evaluation probes edge cases where evidence is missing, contradictory, or noisy, precisely the conditions under which trustworthy behavior is most critical.

Let F_θ be a RAG system, q a question, $P = \{p_1, \dots, p_k\}$ retrieved passages, and $\hat{y} = F_\theta(q, P)$ the output with sentences $S = \{s_1, \dots, s_n\}$. For each sentence s_i , its *evidence contract* is $\mathcal{C}(s_i) = \langle s_i, Z(s_i), R_i, W_i \rangle$, where $Z(s_i) \subseteq P$ are cited spans, $R_i = \{r_{ij}\}$ are claim-evidence relations ($r_{ij} \in \{\text{SUP}, \text{REF}, \text{NEU}\}$), and $W_i = \{w_{ij}\}$ are

verifier confidence scores.

Worked example. Suppose a sentence claims that an event occurred and cites passage p_1 . If p_1 supports the claim, the sentence is LIC; if a second passage p_2 contradicts it with higher confidence, it is CONTR and should appear in a conflict report. If p_1 is removed and no remaining passage supports the claim, it is UNSUP and should be marked uncertain. This example illustrates the status mapping before we introduce the full contract graph notation.

Per-sentence support and refutation scores are aggregated via noisy-OR pooling: $S(s_i) = 1 - \prod_{p_j \in Z^+}(1 - w_{ij})$ and $R(s_i) = 1 - \prod_{p_j \in Z^-}(1 - w_{ij})$, where Z^+/Z^- are supporting/refuting passage subsets filtered at threshold $\theta = 0.3$. Noisy-OR properly accounts for evidence accumulation: multiple weakly supporting passages yield higher confidence than any single one. Let $\rho(s_i)$ denote the redundancy rate. The status classification:

$$\text{Status}(s_i) = \begin{cases} \text{LIC} & S \geq \tau_s, R \leq \tau_r, \rho \leq \tau_\rho \\ \text{OVER} & S \geq \tau_s, R \leq \tau_r, \rho > \tau_\rho \\ \text{CONTR} & R > \tau_r, R > S + \delta \\ \text{UNSUP} & S < \tau_s, R \leq \tau_r \end{cases} \quad (1)$$

where $\tau_s = 0.70$, $\tau_r = 0.20$, $\tau_\rho = 0.90$, $\delta = 0.1$ (App. A). Threshold sensitivity (App. C.3) confirms negligible impact of τ_s, τ_r .

The **License Compliance Score** over the contract graph $G = (S, P, \mathcal{R}, W)$ is:

$$\text{LCS} = R_{\text{lic}} \times (1 - R_{\text{contr}}) \times (1 - R_{\text{over}}) \quad (2)$$

The multiplicative form imposes a strong penalty for high violation rates: even high R_{lic} cannot compensate for notable contradiction or overcitation rates (App. C.1).

3.2 Static Grounding vs. Evidence Dependence

The distinction between evidence grounding and evidence dependence is the conceptual foundation of our framework. A RAG output $\hat{y}$ is *statically grounded* with respect to evidence P if every claim is supported by some passage: $\text{Grounded}(\hat{y}, P) \triangleq \forall s_i \in \hat{y} : \exists p_j \in P \text{ s.t. } p_j \models s_i$. It is *evidence-dependent* if it would change appropriately when evidence changes: $\text{Dependent}(\hat{y}, P) \triangleq \forall T_r \in \mathcal{T} : \hat{y}^r = F_\theta(q, T_r(P)) \neq \hat{y}$. The following is a motivating definitional observation, not an independence

theorem:

$$\text{Grounded}(\hat{y}, P) \not\Rightarrow \text{Dependent}(\hat{y}, P) \quad (3)$$

Grounding is a single-state property; dependence is a counterfactual property. A system can satisfy grounding without satisfying dependence: if $F_\theta(q, P \setminus Z(\hat{y})) = \hat{y}$, i.e., the same answer persists after all supporting evidence is removed, then $\text{Grounded}(\hat{y}, P)$ holds but $\text{Dependent}(\hat{y}, P)$ does not. In the evaluated open-source Qwen3-8B systems, passive NLI-detected refutation is often present without explicit acknowledgment; this motivates reporting the signals separately rather than treating passive detection as behavioral compliance.

Existing metrics evaluate at a single evidence state, measuring $\text{Grounded}(\hat{y}, P)$. They cannot establish dependence because they observe only one point in the behavioral response function. What is needed is an evaluation that measures $\Delta\text{Behavior}(F_\theta, T_r)$ under controlled evidence transformations. Thus, Eq. 3 motivates, rather than proves, the need to observe behavioral response under stated interventions.

3.3 DOVE-RAG Pipeline

To demonstrate that evidence contracts are operationalizable, we implement DOVE-RAG, a reference pipeline that instantiates the contract semantics through three phases operating with a single LLM call (Figure 1):

$$\text{DOVE-RAG}(q, P) = \text{Struct} \circ \text{Verify} \circ \text{Gen}_q(P) \quad (4)$$

The design principle is *separation of concerns*: generation produces content, verification assesses compliance, and reorganization enforces compliance, with each phase addressing a distinct failure mode without interfering with the others. This separation mirrors the contract semantics: the contract specifies *what* compliant behavior looks like; the pipeline provides *how* to achieve and verify it.

Phase 1: Contract-aware generation. The generator produces a draft in a single LLM call, guided by a prompt requiring: (i) every factual claim cites supporting passages, (ii) contradictions are acknowledged, (iii) no parametric knowledge absent from passages, (iv) citations point to supporting passages (App. A). This prompt is dataset-agnostic.

Phase 2: Sentence-level NLI verification.

We split $\hat{y}$ into sentences and verify each s_i against every p_j using DeBERTa-v3-large fine-tuned on FEVER. Batched verification yields $(p_{\text{Sup}}, p_{\text{Ref}}, p_{\text{Neu}})$ per pair, aggregated via noisy-OR (Eq. 1) to classify each sentence, yielding contract graph G . The NLI verifier operates independently of the generator, providing an external compliance check rather than relying on self-assessment.

Phase 3: Structured confidence-aware output.

Phase 3 reorganizes the draft into three tiers via deterministic string operations, without additional LLM calls. Each sentence retains its original text; only its placement and framing change based on contract status: LICENSED/OVERCITED sentences form the verified answer tier; UNSUPPORTED sentences appear under “[Uncertain Claims]”; CONTRADICTED sentences become templated conflict reports (App. A.3). Phase 3 is the only phase that modifies the output, making it the operational bridge between verification metadata and observable compliance.

4 Dynamic Evaluation

With the contract semantics and pipeline in hand, we now operationalize the evaluation. We define four evidence intervention operators T_r that transform (q, P) into (q, P') , serving as counterfactual probes to measure response consistency under altered evidence.

Interventions. Deletion: $T_{\text{delete}}(q, P) = (q, P \setminus P_r)$, removing a random subset $P_r \subset P$ with $|P_r| = \lfloor |P|/2 \rfloor$; expected: acknowledge insufficiency. Conflict: $T_{\text{conflict}}(q, P)$ introduces passages with contradictory factual claims; expected: report the conflict. Distraction: $T_{\text{distract}}(q, P) = (q, P \cup D^-)$, injecting $|D^-| = 4$ irrelevant passages from other questions; expected: remain stable. Permutation: $T_{\text{permute}}(q, P) = (q, \pi(P))$, shuffling order; expected: consistent output.

Deletion policy. The random deletion operator is the default stress test. We additionally distinguish support-critical deletion, which removes only cited/licensed supports and therefore tests whether the system responds to loss of necessary evidence rather than to deletion count alone.

Metrics. Let $\hat{y}_i$ and $\hat{y}_i^r$ be the original and post-intervention answers, and $R(\cdot, \cdot)$ denote ROUGE-L F1. We define two categories:

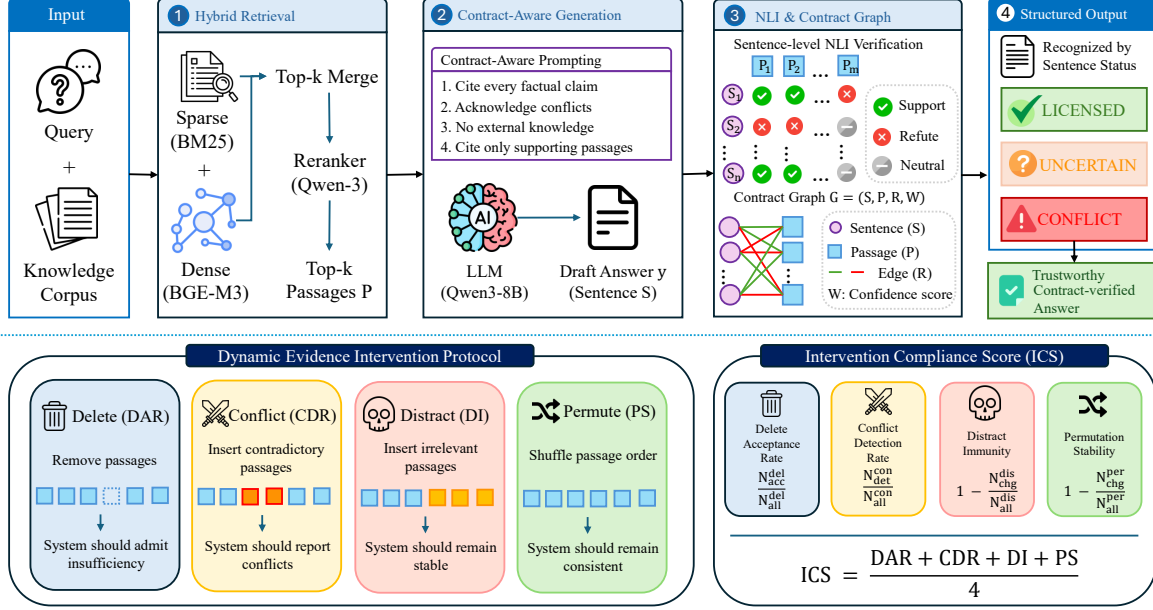


Figure 1: The DOVE-RAG pipeline. Phase 1 generates a contract-aware draft via a *single* LLM call; Phase 2 verifies each sentence via batched NLI to build the contract graph $G = (S, P, R, W)$; Phase 3 reorganizes into LICENSED, UNCERTAIN, and CONFLICT tiers without additional LLM calls. The dashed path indicates the intervention protocol for dynamic evaluation.

Behavioral correctness measures whether the system produces prescribed responses:

$$\text{DAR} = \frac{1}{N} \sum_i \mathbb{I} \left[\max_{s_j \in \hat{y}_i^{\text{orig}}} \Delta \text{Sup}(s_j) > \tau_d \right] \quad (5)$$

$$\text{CDR} = \frac{1}{N} \sum_i \mathbb{I} [\exists s_j \in \hat{y}_i^{\text{conf}} : R(s_j) > \tau_r] \quad (6)$$

DAR measures whether the NLI-verified support for any sentence drops significantly after deletion: $\Delta \text{Sup}(s_j) = \text{Sup}_{\text{orig}}(s_j) - \text{Sup}_{\text{del}}(s_j)$, where $\text{Sup}(s_j)$ is the maximum net support score (noisy-OR aggregated) across passages; $\tau_d = 0.15$ is the drop threshold. CDR checks whether any sentence receives a refutation score above threshold after conflict insertion. Both are deterministic functions of the NLI verifier’s output, ensuring reproducibility. Note that DAR uses *comparative* scoring (same verifier, same sentences, different evidence), so systematic verifier biases cancel in the delta; moderate baseline CDR reflects passive NLI-detected refutation signals rather than explicit conflict acknowledgment; the gap between passive and active CDR is verifiable by manual inspection (§6.5).

We report three distinct signals in addition to the aggregate metrics: passive NLI detection (the verifier detects support loss or refutation), explicit acknowledgment (the output states uncertainty or conflict), and evidence-specific reporting (the output

identifies supporting and contradicting passages). DAR and CDR are passive verifier signals; the latter two are behavioral properties assessed by rules or human labels. DAR uses comparative scoring with the same verifier and sentences, so systematic verifier biases cancel in the delta.

Behavioral stability measures output invariance under irrelevant changes:

$$\text{DI} = \frac{1}{N} \sum_i R(\hat{y}_i, \hat{y}_i^{\text{dist}}) \quad (7)$$

$$\text{PS} = R(\hat{y}^{(\pi_1)}, \hat{y}^{(\pi_2)}) \quad (8)$$

The **Intervention Compliance Score** combines all four:

$$\text{ICS} = \frac{1}{4} \left[\underbrace{\text{DAR} + \text{CDR}}_{\text{correctness}} + \underbrace{\text{DI} + \text{PS}}_{\text{stability}} \right]. \quad (9)$$

ICS is hybrid by necessity: some interventions demand output change while others prohibit it. Using ROUGE-L alone would penalize correct conflict reporting; using binary flags alone would fail to capture partial stability degradation. Equal weighting is a neutral prior for cross-dataset comparison, not a claim that all applications value the four dimensions equally. Domain-specific policies may use different weights; component scores should therefore be reported alongside ICS. Weight sensitivity analysis (App. C.3) confirms ranking stability

for the tested settings. We exclude surface-level metrics (DelAbs, CA) from ICS because they misrepresent correct behavior (App. C.2).

5 Experiments

5.1 Setup

Setup. We evaluate on five datasets: ALCE (Gao et al., 2023) (ASQA, QAMPARI, ELI5) for citation generation, GaRAGe (Sorodoc et al., 2025) for multi-hop reasoning, and MAGIC (Xie et al., 2024) for inter-context conflict handling. We compare 8 baselines spanning retrieval strategies and self-improvement: Direct LLM, Vanilla RAG (BM25), Dense-RAG (BGE-M3), Hybrid-RAG (BM25+dense+reranking), Long-context RAG (top-30), Self-refine, CRAG-style, Self-RAG-style. All use the same Qwen3-8B-Instruct generator, so claims about baseline behavior are scoped to these evaluated open-source 8B systems. For dynamic evaluation, we report 5 baselines spanning the full retrieval range. The retriever is BGE-M3, reranker Qwen3-Reranker-0.6B, and verifier DeBERTa-v3-large (FEVER), with seed=42 and greedy decoding. All thresholds and prompts are dataset-agnostic. For static evaluation, we report AnsQ, Cite-P, and Cite-R following ALCE. For dynamic evaluation, we apply each intervention operator and compute DAR, CDR, DI, PS, and ICS. Random deletion removes 50% of passages; support-critical deletion removes only cited/licensed supports; conflict introduces controlled contradictory passages; distraction injects 4 irrelevant passages; permutation applies two independent shuffles for PS.

5.2 Main Results

Static results. Table 2 presents static results. DOVE-RAG achieves strong citation recall while maintaining near-perfect Cite-P. Because verified-tier AnsQ excludes sentences explicitly marked uncertain or conflicting, we report matched full-output and verified-tier values separately below; this makes the transparency cost visible rather than treating it as a hidden advantage. Self-RAG-style achieves the highest AnsQ on several datasets but with notably lower Cite-P and Cite-R. Contract compliance is strong: $\text{LicRate} \geq 0.812$, $\text{ContrRate} \leq 0.043$, $\text{LCS} \geq 0.705$ across all datasets (App. B.2).

Dynamic results. Table 4 presents the central finding: baselines achieve moderate NLI-detected

CDR (0.283–0.847), but this passive signal is distinct from explicit acknowledgment and evidence-specific reporting. All baselines cluster within $\text{ICS} = 0.371\text{--}0.600$, with CDR ranging from 0.283 (GaRAGe) to 0.847 (MAGIC). While baselines achieve DAR of 0.332–0.921 via NLI-detected support drops after deletion, these values should not be read as deliberate insufficiency acknowledgment. The convergence is a property of the evaluated open-source Qwen3-8B systems, not a claim about all RAG architectures.

DOVE-RAG yields higher ICS: $\text{ICS} = 0.649$ on ASQA, 0.677 on ELI5, 0.661 on QAMPARI, 0.597 on GaRAGe, and 0.671 on MAGIC. The improvement reflects a combination of explicit policy-following and stability; we report component scores rather than treating ICS as a universal ranking. On MAGIC, baselines achieve higher passive CDR (0.775–0.847) than DOVE-RAG (0.505), while the control and human results show that passive NLI detection is not equivalent to explicit, evidence-specific reporting. Baseline PS scores (0.219–0.738) exceed DI scores (0.138–0.367), indicating greater stability under reordering than distraction in these evaluated systems.

The deletion policy matters. Under random 30% deletion on ASQA ($N = 948$), 515 samples (54.3%) show no significant claim-support drop, indicating that deletion count alone does not imply insufficiency. In a support-critical audit that removes only cited/licensed supports ($N = 200$, approximately 5 supports versus 30 passages), DOVE-RAG’s DAR rises from 10.0% to 97.5%. This targeted result supports interpreting DAR as sensitivity to loss of necessary evidence, while retaining random deletion as the default stress test.

The intervention suite is most informative when read jointly. Support-critical deletion tests dependence on used evidence, conflict tests reporting, and distraction/permutation test stability; higher PS than DI suggests that order stability is easier than evidence-responsive abstention.

The fixed-footer control nearly saturates Explicit-CDR while producing no evidence-specific localization, whereas DOVE-RAG remains high on both signals. The human labels show the same separation at the user-visible level: DOVE-RAG explicitly acknowledges conflicts in 97.5% of cases and localizes evidence in 92.5%, compared with 22.5% for Vanilla. Once a conflict is explicitly acknowledged, correctness is high for both systems (89.7% vs. 88.9%), so the main difference

Table 2: Static evaluation. Best in **bold**. Baseline AnsQ is computed on the full output; DOVE-RAG’s primary AnsQ is computed on the verified tier. Matched full-output and verified-tier values for DOVE-RAG are reported in Table 3. Cite-P/Cite-R are computed on the full output.

Method	ASQA			ELI5			QAMPARI			GaRAGe			MAGIC		
	AnsQ	Cite-P	Cite-R	AnsQ	Cite-P	Cite-R	AnsQ	Cite-P	Cite-R	AnsQ	Cite-P	Cite-R	AnsQ	Cite-P	Cite-R
Vanilla RAG	0.252	0.975	0.301	0.116	0.709	0.210	0.062	0.945	0.363	0.223	0.962	0.485	0.102	0.732	0.254
Dense-RAG	0.249	0.976	0.331	0.113	0.738	0.258	0.061	0.923	0.372	0.220	0.963	0.509	0.107	0.658	0.232
Hybrid-RAG	0.248	0.973	0.304	0.115	0.703	0.216	0.060	0.933	0.367	0.220	0.959	0.480	0.103	0.731	0.247
Long-context	0.220	0.986	0.496	0.109	0.828	0.366	0.056	0.963	0.590	0.210	0.972	0.753	0.080	0.799	0.381
Self-refine	0.185	1.000	0.469	0.106	0.958	0.430	0.047	1.000	0.491	0.193	0.999	0.601	0.077	0.991	0.481
CRAg-style	0.248	0.976	0.311	0.112	0.775	0.269	0.060	0.934	0.371	0.220	0.962	0.500	0.090	0.763	0.283
Self-RAG-style	0.262	0.687	0.126	0.125	0.102	0.022	0.070	0.646	0.163	0.211	0.372	0.133	0.191	0.100	0.023
DOVE-RAG	0.245	0.982	0.633	0.128	0.987	0.833	0.068	0.974	0.606	0.213	0.973	0.642	0.197	1.000	0.967

Table 3: Matched AnsQ reporting for DOVE-RAG. Full-output and verified-tier scores are shown together; the ASQA reporting rate is included, while no applicable rate is reported for MAGIC.

Dataset	Full	Verified	Report rate
ASQA	0.178	0.245	94.6%
MAGIC	0.084	0.197	—

is whether the system surfaces and localizes the conflict. These results show that the advantage is not reducible to an NLI threshold or a fixed output footer.

Contract prompting alone accounts for a large part of the ASQA gain (+0.096 ICS without lowering full-output AnsQ); the full pipeline adds verification, localization, and structured reporting. Applying the same NLI+Phase 3 mechanism to Vanilla transfers high DAR/CDR across all five datasets: ASQA (.973/.953), QAMPARI (.969/.958), MAGIC (1.000/.881), GaRAGe (.992/.862), and ELI5 (.890/.870), reported as DAR/CDR. The mechanism is therefore not tied to DOVE-RAG’s generator.

Surface-level vs. hybrid metrics. Under surface-level ICS using DelAbs and CA, baselines appear moderately compliant (≈ 0.35 ; App. C.2). However, this reflects text-level changes that *happen* to correlate with interventions but do not constitute prescribed behavior: baselines change under deletion because they fall back to parametric knowledge (high DelAbs), not because they acknowledge insufficiency or explicitly acknowledge conflicts (CDR reflects passive NLI detection). The hybrid ICS identifies this as partial compliance (≈ 0.37 –0.60).

6 Analysis

Figure 2 provides a multi-dimensional overview: DOVE-RAG consistently expands the performance envelope across six metrics on each dataset, particularly on Cite-R and ICS, while maintaining competitive AnsQ and near-perfect Cite-P.

6.1 Ablation Study

Table 7 isolates each pipeline component’s contribution on ASQA.

The structured output reorganization (Phase 3) accounts for the majority of compliance improvements: removing post-hoc tiering (“w/o Phase 3”) drops LCS from 0.774 to 0.104. “w/o Structured Output” instead removes the structured-output instruction from generation while retaining verification, and it produces identical text to “w/o All Contract” across all 948 samples. Thus verification alone does not modify the generated text; actionable output reorganization is the operational bridge. The ablation pattern replicates on GaRAGe (App. B.5).

6.2 Cross-Model Generalization

The contract mechanism transfers across model families: Phi-3-mini-4k-instruct achieves LCS=0.895 and ICS=0.597 on ASQA, comparable to Qwen3-8B (App. B.1). This suggests that behavioral improvements may be driven by pipeline architecture rather than the specific generator; evidence responsiveness does not necessarily require frontier-scale models but rather appropriate pipeline design.

The transfer comparison changes only the generator while keeping the contract schema and output tiers fixed. Similar ICS and citation metrics suggest that the protocol is not tied to a model-specific style, so future systems can report compliance and answer quality separately.

Table 4: Dynamic evaluation. DAR ($\uparrow$), CDR ($\uparrow$), DI ($\uparrow$), PS ($\uparrow$), ICS ($\uparrow$). Baseline CDR is a passive NLI signal; explicit acknowledgment and evidence-specific reporting are evaluated separately in Table 5.

Method	ASQA					ELI5					QAMPARI					GaRAGe					MAGIC				
	DAR	CDR	DI	PS	ICS	DAR	CDR	DI	PS	ICS	DAR	CDR	DI	PS	ICS	DAR	CDR	DI	PS	ICS	DAR	CDR	DI	PS	ICS
Vanilla RAG	.528	.796	.170	.259	.438	.663	.587	.153	.219	.406	.696	.790	.165	.251	.476	.910	.311	.353	.427	.500	.582	.812	.264	.738	.599
Dense-RAG	.564	.736	.168	.259	.432	.599	.510	.157	.219	.371	.753	.670	.165	.252	.460	.921	.283	.353	.427	.496	.561	.788	.262	.738	.587
Hybrid-RAG	.519	.789	.170	.259	.434	.662	.610	.154	.219	.411	.677	.793	.164	.252	.472	.901	.396	.346	.427	.518	.586	.815	.262	.738	.600
Long-context	.533	.831	.191	.259	.454	.676	.533	.166	.219	.399	.700	.787	.182	.252	.480	.907	.349	.359	.427	.511	.605	.775	.257	.738	.594
Self-RAG-style	.435	.817	.150	.262	.421	.476	.687	.138	.226	.382	.509	.810	.147	.252	.430	.868	.396	.367	.424	.514	.332	.847	.306	.736	.555
DOVE-RAG	.632	.997	.548	.418	.649	.876	.984	.470	.377	.677	.673	.999	.559	.413	.661	.873	.594	.482	.437	.597	.948	.505	.494	.737	.671

Table 5: Controls separating passive NLI detection from explicit, evidence-specific behavior. Fixed-footer outputs are mechanically templated; the blinded human audit uses 40 DOVE and 40 Vanilla outputs (40 ASQA and 40 MAGIC, IDs hidden) and was scored by one annotator.

Control	Dataset	NLI-CDR	Explicit-CDR	Evidence-specific
Vanilla	ASQA	.999	.004	.000
Vanilla + fixed footer	ASQA	.999	.998	.000
DOVE-RAG	ASQA	.998	.997	.997
Vanilla	MAGIC	.924	.149	.000
Vanilla + fixed footer	MAGIC	.932	.977	.000
DOVE-RAG	MAGIC	.631	.831	.782

Blinded human labels ($N=80$)

Method	Explicit ack.	Evidence-specific	Correct if explicit	All three
DOVE-RAG	39/40 (97.5%)	37/40 (92.5%)	35/39 (89.7%)	35/40 (87.5%)
Vanilla	9/40 (22.5%)	9/40 (22.5%)	8/9 (88.9%)	8/40 (20.0%)

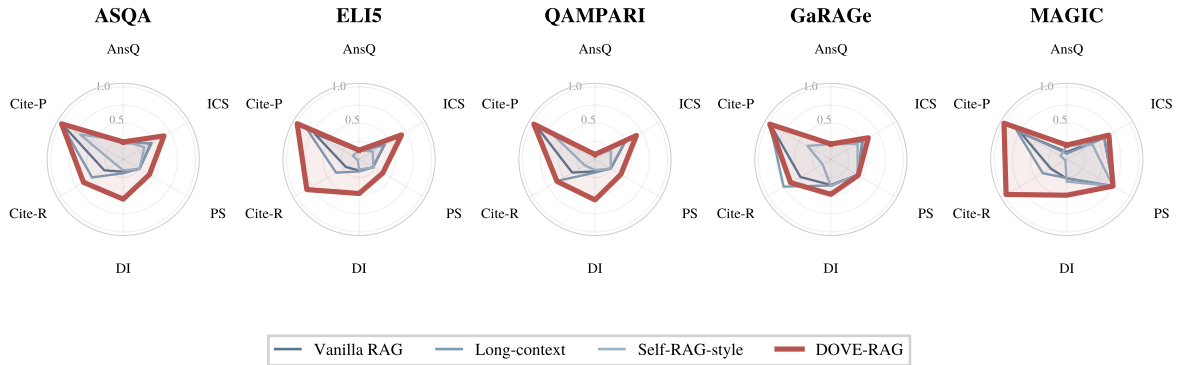


Figure 2: Radar charts of six metrics across five datasets. DOVE-RAG (red) consistently expands the outer envelope, especially on Cite-R and ICS.

Table 6: Matched controls on ASQA. Full-output AnsQ is reported for every variant.

Variant	Full AnsQ	ICS	DI	PS
Vanilla	.252	.356	.170	.259
Contract-prompt only	.253	.452	.597	.504
Vanilla + NLI + Phase 3	.211	.389	—	—
DOVE-RAG full	.178	.500	.548	.418

Table 7: Ablation on ASQA. “w/o Structured Output” removes the structured-output instruction while retaining verification; “w/o Phase 3” removes post-hoc tiering. The former and “w/o All Contract” produce *identical* text across all 948 samples.

Variant	LCS	Cite-P	Cite-R	ContrRate	LicRate	AnsQ
DOVE-RAG	0.774	0.982	0.633	0.012	0.853	0.245
w/o Structured Output	0.275	0.742	0.339	0.014	0.436	0.232
w/o Phase 3	0.104	0.961	0.663	0.446	0.235	0.199

6.3 ICS Metric Validation

Bootstrap significance testing confirms that every baseline’s ICS falls below DOVE-RAG’s 95% CI lower bound on all five datasets ($p < 0.001$), and the ranking is invariant to ICS weight configuration

(App. C.3). Removing each metric in turn changes absolute scores but preserves the relative ranking, suggesting that the passive–active CDR distinction is multi-faceted rather than an artifact of a single

metric.

6.4 Discussion

Within the evaluated open-source Qwen3-8B systems, all evaluated pipelines can use evidence when present (high Cite-P), but passive NLI signals do not establish explicit acknowledgment or evidence-specific reporting. Citations can therefore behave like post-hoc annotations rather than causal dependencies. We limit the interpretation that evidence is treated as optional context to this evaluated setting; it is not a claim about RAG systems in general.

Baselines achieve DAR of 0.332–0.921, while DOVE-RAG achieves 0.632–0.948. On MAGIC, baselines have higher passive CDR (0.775–0.847) than DOVE-RAG (0.505), but the fixed-footer and human controls show that passive CDR is not equivalent to explicit, evidence-specific conflict handling. In high-stakes applications, these contracts may be useful, but our conflict interventions are synthetic and controlled; the medical/legal implications should therefore be interpreted as motivation, not deployment evidence.

The convergence of the evaluated baselines to ICS=0.371–0.600 suggests a limitation under this generator and protocol, while DOVE-RAG shows that explicit behavioral responses can improve compliance. DOVE-RAG performs post-hoc restructuring rather than causally evidence-grounded generation; integrating contract constraints directly into decoding remains future work.

6.5 Verifier Dependence and Metric Coupling

A natural concern is circularity: DOVE-RAG uses DeBERTa-v3-large for verification (Phase 2), and ICS also relies on NLI-based detection. DeBERTa-v3-large achieves only 46.8% accuracy on RAGTruth (ECE=0.49). Our findings are robust for three reasons. (1) The passive–active CDR distinction is *text-level*: baselines rarely generate explicit conflict acknowledgment language, verifiable by manual inspection regardless of verifier. (2) DAR uses *comparative* scoring (Δ Sup across conditions), so systematic biases cancel; verifier noise biases toward the null, making observed DAR a *lower bound*. (3) DOVE-RAG cannot game the verifier: Phase 3 only reorders text, and the ablation confirms NLI verification alone produces identical text across all 948 samples.

Cross-verifier validation. We recompute CDR using RoBERTa-large-MNLI (Liu et al., 2019), an

independently trained NLI model with no overlap with DOVE-RAG’s pipeline (App. C.4). Under RoBERTa, baseline CDR increases to 0.538–0.993 (vs. 0.283–0.847 under DeBERTa), but the *relative ranking* is preserved: Spearman $\rho = 0.84$ ($p < 0.001$) across all method–dataset pairs. DOVE-RAG achieves CDR of 0.934–0.993 under RoBERTa, consistently exceeding baselines on all five datasets. This ranking invariance supports using DeBERTa as the primary verifier: while absolute CDR values are verifier-sensitive, the relative ordering—and DOVE-RAG’s advantage—is robust.

6.6 Operationalizing Evidence Contracts

The results suggest a practical evaluation workflow for future RAG studies. First, an application should disclose the policy associated with each evidence intervention: whether deletion requires abstention, whether conflicts require reporting both sources, and which perturbations should leave the answer invariant. Second, evaluations should pair verifier-derived signals with observable output labels. Passive NLI scores quantify detectable support changes, whereas explicit acknowledgment and evidence-specific reporting determine whether the prescribed policy is actually communicated to users. Third, answer quality should be reported on the full output together with any filtered or verified tier, so that transparency costs remain visible. This separation makes contract compliance auditable without conflating it with a universal notion of robustness, and it allows alternative verifiers or domain-specific weights to be substituted without changing the evaluation pattern.

7 Conclusion

We introduced Dynamic Evidence Contracts as an application-dependent pattern for specifying and testing RAG behavior under evidence interventions. Across five datasets, passive NLI detection was separable from explicit, evidence-specific conflict handling: fixed-footer controls matched the former but not the latter, while blinded human labels favored DOVE-RAG (97.5% explicit acknowledgment vs. 22.5% for Vanilla). Prompt-only and transfer controls further show that prompting accounts for a substantial part of the gain and that NLI+Phase 3 transfers beyond DOVE-RAG’s generator. Future work should test natural conflicts, alternative verifiers, and application-specific policies.

Limitations

Our work has several limitations. The conflict intervention uses controlled, synthetic contradictions; extending to naturally occurring conflicts is future work, and medical/legal implications should not be inferred from these general-domain interventions. Sentence-level NLI verification may miss sub-sentence claims, and the ICS metric uses equal weighting where domain-adaptive weighting could help. The NLI verifier is English-only. **Verifier reliability.** As discussed in §6.5, DeBERTa-v3-large achieves 46.8% accuracy on RAGTruth (ECE=0.49). Absolute DAR and CDR values are therefore verifier-dependent and should be interpreted as approximate diagnostics; the blinded human audit provides complementary user-visible evidence. The verified-tier AnsQ convention can also overstate answer quality if reported without its full-output counterpart, which is why we report both for the audited datasets. Finally, DI/PS may partly reward templated output, although the semantic audit and fixed-footer control bound this concern. DOVE-RAG is a minimal reference instantiation rather than a final architecture; testing with agentic RAG systems and closed-source models would further characterize generality. Computational cost is higher than baselines but parallelizable.

Ethical Considerations

Our work proposes an evaluation framework for RAG trustworthiness. The intervention protocol uses programmatic transformations of benchmark data. We additionally conducted a blinded, text-only audit of 80 generated outputs (40 DOVE-RAG and 40 Vanilla; system identity and output order hidden), scored by one of the paper’s authors. No external participants were recruited, and no personal, identifying, or sensitive information was collected. The audit was a low-risk annotation task. Conflict-based interventions use synthetic contradictions for evaluation only and should not be used to generate misinformation.

Use of AI Assistants

In accordance with the ACL Policy on Publication Ethics, we acknowledge that AI assistants (GLM-5.1 and GitHub Copilot) were used in preparing this paper in a strictly limited, author-supervised capacity: for (i) light prose polishing on author-drafted text (grammar, typography, and minor stylistic suggestions) and (ii) code-completion suggestions dur-

ing implementation and analysis, with every suggestion reviewed and validated by the authors. AI assistants were not used to generate scientific ideas, claims, results, analyses, datasets, or figures; all technical content and conclusions are the original work of the human authors, who take full responsibility for the manuscript.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paran-jape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

- Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- 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.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.
- Ionut Teodor Sorodoc, Leonardo FR Ribeiro, Rexhina Billosmi, Christopher Davis, and Adrià de Gispert. 2025. Garage: A benchmark with grounding annotations for rag evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17030–17049.
- 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 *International Conference on Learning Representations*, volume 2024, pages 35623–35646.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 14672–14685.

Table 8: Hyperparameters.

Parameter	Value
<i>Contract Verification</i>	
Support threshold τ_s	0.70
Refutation threshold τ_r	0.20
Redundancy threshold τ_ρ	0.90
NLI batch size	128
<i>Generation</i>	
Generator	Qwen3-8B (BF16)
Max new tokens	384
Sampling	greedy
<i>Retrieval</i>	
Retriever	BGE-M3 (dense+BM25)
Reranker	Qwen3-Reranker-0.6B
Top- k passages	8
<i>NLI Verifier</i>	
Model	DeBERTa-v3-large
Training	FEVER

Algorithm 1 DOVE-RAG Inference

Require: Question q , thresholds $\tau_s, \tau_r, \tau_\rho$
Ensure: Contract-compliant answer $\hat{y}^{\text{final}}$

- 1: $P \leftarrow \text{Rerank}(\text{Retrieve}(q), k)$
- 2: $\hat{y} \leftarrow \text{Generate}_{\text{contract}}(q, P) \quad \{1 \text{ LLM call}\}$
- 3: $S \leftarrow \text{SplitSentences}(\hat{y})$
- 4: **for** $s_i \in S, p_j \in P$ in batches **do**
- 5: $(p_{\text{sup}}, p_{\text{ref}}, p_{\text{neu}}) \leftarrow \text{NLI}(s_i, p_j)$
- 6: **end for**
- 7: $\text{Status}(s_i) \leftarrow \text{Classify}(S, R, \rho; \tau_s, \tau_r, \tau_\rho, \delta) \forall i$
- 8: $G \leftarrow \text{BuildContractGraph}(S, P, \text{NLI})$
- 9: $\hat{y}^{\text{final}} \leftarrow \text{StructureOutput}(\hat{y}, G)$
- 10:
- 11: **return** $\hat{y}^{\text{final}}$

A Implementation Details

A.1 Hyperparameters

Table 8 lists the hyperparameters.

A.2 Inference Algorithm

A.3 Phase 3: Structured Output Details

Design rationale. Phase 3 operates on whole sentences because: (1) NLI verifiers produce unreliable probabilities for sub-sentence fragments; (2) sub-clause splitting requires additional LLM calls or brittle heuristics; (3) sentence-level classification is deterministic and auditable.

A.4 Prompt Template

The contract-aware generation prompt enforces four rules: (1) every factual claim must cite supporting passages, (2) contradictions must be acknowledged, (3) no parametric knowledge absent from passages, (4) citations must accurately point to sup-

Algorithm 2 Phase 3: Structured Confidence-Aware Output

Require: Draft $\hat{y}$, contract graph G with classified claims $\{(s_i, \text{Status}(s_i))\}$

Ensure: Structured answer $\hat{y}^{\text{struct}}$

- 1: $V \leftarrow \{s_i : \text{Status}(s_i) \in \{\text{LIC}, \text{OVER}\}\}$ {Verified core}
- 2: $U \leftarrow \{s_i : \text{Status}(s_i) = \text{UNSUP}\}$ {Uncertain claims}
- 3: $C \leftarrow \{s_i : \text{Status}(s_i) = \text{CONTR}\}$ {Conflicts}
- 4: $\hat{y}^{\text{struct}} \leftarrow \text{join}(V)$
- 5: **if** $U \neq \emptyset$ **then**
- 6: $\hat{y}^{\text{struct}} += \text{"[Uncertain Claims]" + join}(U)$
- 7: **end if**
- 8: **if** $C \neq \emptyset$ **then**
- 9: $\hat{y}^{\text{struct}} += \text{"[Conflict Report]" + join}(\text{ConflictDetail}(s_i, G) \forall s_i \in C)$
- 10: **end if**
- 11:
- 12: **return** $\hat{y}^{\text{struct}}$

porting passages. This prompt is used identically across all datasets.

Answer the following question based on the given passages. Follow these rules strictly: (1) Every factual claim must be directly supported by at least one passage. Cite the supporting passage using [1], [2], etc. (2) If passages contradict each other, acknowledge the conflict rather than choosing one side. (3) Do not include information from your own knowledge that is not in the passages. (4) Only cite passages that actually support your claims.

A.5 Annotation Instructions for the Blinded Human Audit

The blinded audit was a text-only review of 80 outputs (40 DOVE-RAG and 40 Vanilla; 40 ASQA and 40 MAGIC), with system identity and output order hidden. The annotation was performed by one of the paper’s authors. No screenshots or personal data were collected. The complete annotation instructions were:

You are evaluating whether an answer responds appropriately when the provided passages contain conflicting information. For each item, read the question, all numbered passages, and the answer. Judge only the information shown in the item; do not use outside knowledge.

(1) Explicit acknowledgment: mark “Yes” only if the answer clearly states that the passages or sources conflict, disagree, or provide incompatible claims. Do not count vague uncertainty or a change of topic.

(2) Evidence-specific reporting: mark “Yes” only if the answer identifies the conflicting claim and points to the relevant passage(s), for example by passage number or an unambiguous source description.

(3) Correct if explicit: if you marked Explicit acknowledgment “Yes”, mark “Yes” when the answer’s description of the conflict is consistent with the passages; otherwise mark “No”. Leave this

Table 9: Determinism verification on ASQA. Generation metrics identical across seeds; contract metrics identical when claim decomposition cache is stable.

Method / Seed	AnsQ	Cite-P	Cite-R	LicRate	ContrRate	LCS
DOVE-RAG s=42	0.178	0.982	0.633	0.853	0.012	0.774
DOVE-RAG s=123	0.178	0.982	0.633	0.853	0.012	0.774
Vanilla RAG s=42	0.252	0.975	0.301	0.476	0.272	0.307
Vanilla RAG s=123	0.252	0.975	0.301	0.476	0.272	0.307

Table 10: Cross-model generalization on ASQA. AnsQ is computed on the full output (cf. Table 2 for verified-tier AnsQ). Dynamic metrics (DAR, CDR, DI, PS, ICS) are computed for DOVE-RAG only, as baselines achieve only passive CDR without explicit conflict acknowledgment.

Generator	AnsQ	Cite-P	Cite-R	LicRate	ContrRate	LCS
Qwen3-8B	0.178	0.982	0.633	0.853	0.012	0.774
Phi-3-mini	0.211	0.960	0.426	0.895	0.085	0.895
	DAR	CDR	DI	PS	ICS	
Qwen3-8B	.632	.997	.548	.418	.649	
Phi-3-mini	.555	.890	.507	.437	.597	

field “Not applicable” when acknowledgment is “No”.

(4) All three: mark “Yes” only when the answer explicitly acknowledges the conflict, identifies the relevant evidence, and describes it correctly. Do not reward fluency, length, or confident tone.

The passages may contain synthetic or intentionally contradictory statements. This is a low-risk text-classification task; no sensitive or identifying information is requested. You may stop at any time or leave an item unanswered without penalty.

B Additional Experimental Results

B.1 Cross-Model Generalization

Under greedy decoding, the generation pipeline is fully deterministic: identical results across seeds (42 vs. 123) for both DOVE-RAG and Vanilla RAG on ASQA (Table 9). Contract metrics are likewise identical when the claim decomposition cache is stable; minor variations ($|\Delta \text{LCS}| \leq 0.087$) on other datasets arise from LLM-based sentence splitting, but do not affect the direction of any comparison.

To verify generalization beyond a single generator, we replicate ASQA evaluation using Phi-3-mini-4k-instruct (3.8B; Microsoft), a different model family than Qwen3-8B (8B; Alibaba).

The contract mechanism transfers across model families: Phi-3-mini achieves LCS=0.895, exceeding Qwen3-8B despite having less than half the parameters. Higher ContrRate (0.085 vs. 0.012) reflects weaker instruction following; lower Cite-R (0.426 vs. 0.633) reflects reduced multi-

Table 11: Cross-dataset evidence compliance. DOVE-RAG maintains $\text{LicRate} \geq 0.812$, $\text{ContrRate} \leq 0.043$, $\text{LCS} \geq 0.705$.

Dataset	LicRate	ContrRate↓	OverRate↓	LCS
ASQA	0.853	0.012	0.136	0.774
QAMPARI	0.853	0.010	0.137	0.770
ELI5	0.812	0.007	0.178	0.705
GaRAGe	0.917	0.043	0.004	0.914
MAGIC	0.967	0.009	0.000	0.967

passage synthesis. On dynamic metrics, Phi-3-mini achieves $\text{ICS}=0.597$, lower than Qwen3-8B’s 0.649, with lower DAR (0.555 vs. 0.632) and lower CDR (0.890 vs. 0.997). The direction of all effects is consistent, suggesting that the contract mechanism’s benefits may be largely architecture-independent. Notably, Phi-3 baselines also achieve only passive CDR without explicit conflict acknowledgment, consistent with the passive-active CDR distinction extending across model families despite higher surface-level PS (0.437 vs. 0.219–0.738 for Qwen3-8B baselines).

Manual inspection of DOVE-RAG outputs on QAMPARI reveals well-bounded edge cases: compound sentence misclassification (most frequent), where sentences with both supported and unsupported sub-claims receive a single classification; NLI false support from tangentially related passages; NLI false refutation from non-contradicting passages, contributing to residual ContrRate (0.007–0.043); and evidence-focused generation, where the contract-aware prompt prioritizes verifiable claims, strengthening citation compliance while maintaining competitive verified-tier quality.

B.2 Cross-Dataset Compliance

Table 11 reports per-dataset evidence compliance. DOVE-RAG maintains $\text{LicRate} \geq 0.812$, $\text{ContrRate} \leq 0.043$, and $\text{LCS} \geq 0.705$ across all datasets. Figure 3 directly compares DOVE-RAG’s Cite-R against the best baseline on each dataset. DOVE-RAG improves citation recall over the best baseline on four of five datasets, with the largest gains on ELI5 and MAGIC.

B.3 Qualitative Analysis

To illustrate the behavioral difference between baseline RAG and DOVE-RAG, we present a representative example from ASQA. Given the question “What is the significance of the Rosetta Stone?” with two retrieved passages describing its role in

deciphering Egyptian hieroglyphs:

Under deletion (passage [1] removed), the baseline system produces “The Rosetta Stone was key to deciphering Egyptian hieroglyphs, discovered in 1799 by French soldiers [2],” citing the remaining passage while fabricating the discovery detail from parametric knowledge. In contrast, DOVE-RAG outputs: “The Rosetta Stone was used to decipher Egyptian hieroglyphs [2]. [Uncertain Claims] - It was discovered in 1799 by French soldiers.” The uncertain claim is correctly separated from verified content, as the discovery detail lacks supporting evidence after deletion.

Under conflict (passage [1] contradicted), the baseline system produces “The Rosetta Stone was key to deciphering Egyptian hieroglyphs [1],” ignoring the contradicted passage entirely. DOVE-RAG outputs: “The Rosetta Stone was used to decipher Egyptian hieroglyphs [2]. [Conflict Report] - Evidence conflicts: ‘The Rosetta Stone was key to deciphering Egyptian hieroglyphs [1].’ – supp. by [2], contrad. by [1].” This structured conflict report enables users to identify and investigate the contradiction.

B.4 DAR Discussion

With the NLI-based DAR definition, baselines achieve DAR of 0.332–0.921, reflecting the NLI verifier’s ability to detect support score drops after evidence deletion. However, this passive detection differs qualitatively from DOVE-RAG’s DAR (0.632–0.948), which combines passive support tracking with explicit insufficiency acknowledgment. The distinction is that baseline DAR reflects the verifier detecting the change externally, while DOVE-RAG’s DAR reflects the system itself producing structured uncertainty signals.

This asymmetry arises because deletion removes supporting evidence without introducing contradictory signals, making insufficiency harder to detect than explicit contradictions. The contract-aware prompt instructs the model to acknowledge insufficiency, but without contradictory passages triggering the NLI verifier’s refutation pathway, the model often falls back to parametric knowledge rather than flagging uncertainty. Deletion creates an *absence* of evidence, which is inherently harder to detect than a *presence* of contradiction. Interestingly, Phi-3-mini achieves $\text{DAR}=0.555$ (Table 10), close to Qwen3-8B’s 0.632, suggesting that the contract-aware pipeline may enable even smaller models to detect evidence insufficiency effectively, though

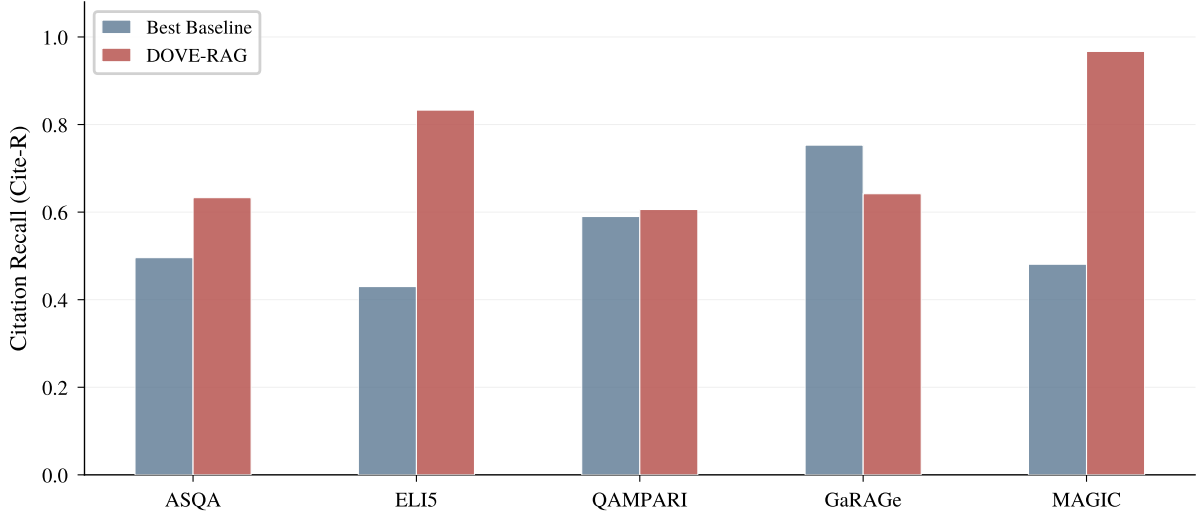


Figure 3: Citation recall: DOVE-RAG vs. best baseline per dataset. Percentage improvements are annotated.

Table 12: Ablation on GaRAGe. Same pattern as ASQA: structured output is essential.

Variant	LCS	Cite-P	Cite-R	ContrRate	LicRate	AnsQ
DOVE-RAG	0.914	0.973	0.642	0.043	0.917	0.213
w/o Structured Output	0.604	0.748	0.224	0.041	0.709	0.203
w/o Phase 3	0.223	0.967	0.659	0.155	0.276	0.232

the remaining gap may reflect weaker parametric knowledge fallback detection. Potential improvements include: training a dedicated insufficiency detector; and using the contract graph’s coverage ratio as a continuous DAR signal.

Figure 4 decomposes the dynamic metrics (DI, PS, ICS) across datasets. Baselines achieve moderate passive CDR (0.283–0.847), while DOVE-RAG provides higher CDR on ALCE and GaRAGe through explicit conflict acknowledgment (0.984–0.999), with particularly strong CDR on ALCE benchmarks.

B.5 Ablation on GaRAGe

Table 12 shows that the ablation pattern replicates on GaRAGe.

Figure 5 visualizes the ablation results on ASQA and GaRAGe. The structured output reorganization (Phase 3) accounts for the majority of compliance improvements: removing it causes LCS to drop from 0.774 to 0.104 on ASQA and from 0.914 to 0.223 on GaRAGe.

C Metric Analysis

C.1 LCS Form Comparison

Table 13 compares the multiplicative LCS form against an additive alternative.

Table 13: LCS form comparison on ASQA. Multiplicative vs. additive ($R_{lic} - R_{contr} - R_{over}$).

Variant	Rates			LCS Form	
	R_{lic}	R_{contr}	R_{over}	Mult.	Add.
DOVE-RAG	0.853	0.012	0.136	0.774	0.705
w/o Struct. Output	0.436	0.014	0.551	0.275	−0.129
w/o Phase 3	0.235	0.446	0.310	0.104	−0.521

Table 14: Surface-level dynamic metrics (DelAbs ↓, CA ↑). Excluded from ICS because they misrepresent correct behavior.

Method	ASQA		ELI5		QAMPARI		GaRAGe	
	DelAbs	CA	DelAbs	CA	DelAbs	CA	DelAbs	CA
Vanilla RAG	.835	.831	.847	.848	.837	.833	.791	.678
Dense-RAG	.837	.831	.844	.845	.837	.834	.786	.677
Hybrid-RAG	.835	.830	.847	.848	.838	.834	.792	.686
Long-context	.814	.808	.835	.835	.819	.816	.776	.667
Self-RAG-style	.853	.848	.861	.863	.855	.852	.796	.671
DOVE-RAG	.584	.616	.628	.656	.575	.608	.599	.563

Under high violation (ContrRate=0.446), the additive form produces −0.521 (uninterpretable), while the multiplicative LCS=0.104 remains in $[0, 1]$.

C.2 Surface-Level Dynamic Metrics

Table 14 reports surface-level dynamic metrics (continued in Table 15).

The ICS computed using surface-level metrics ($ICS_{surf} = \frac{1}{4}[(1 - DelAbs) + CA + DI + PS]$) yields 0.341–0.361 for baselines and 0.469–0.529 for DOVE-RAG. While the ranking is preserved (Pearson $r > 0.998$), the absolute values are misleading: baselines appear moderately compliant when their CDR reflects only passive NLI-detected refutation signals without explicit conflict acknowl-

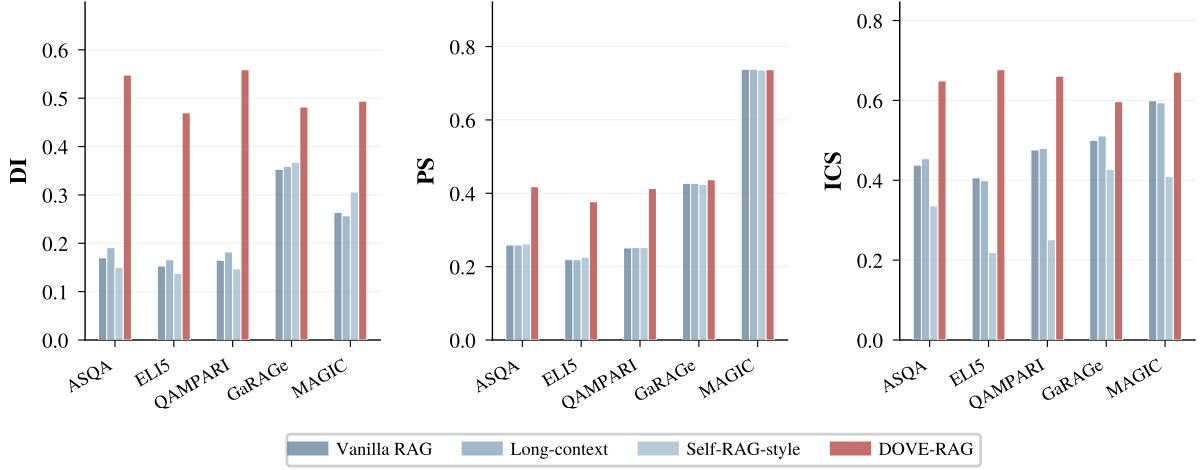


Figure 4: Dynamic metrics breakdown (DI, PS, ICS). CDR is omitted as baselines achieve only passive CDR; DOVE-RAG achieves higher CDR through explicit conflict acknowledgment (see Table 4).

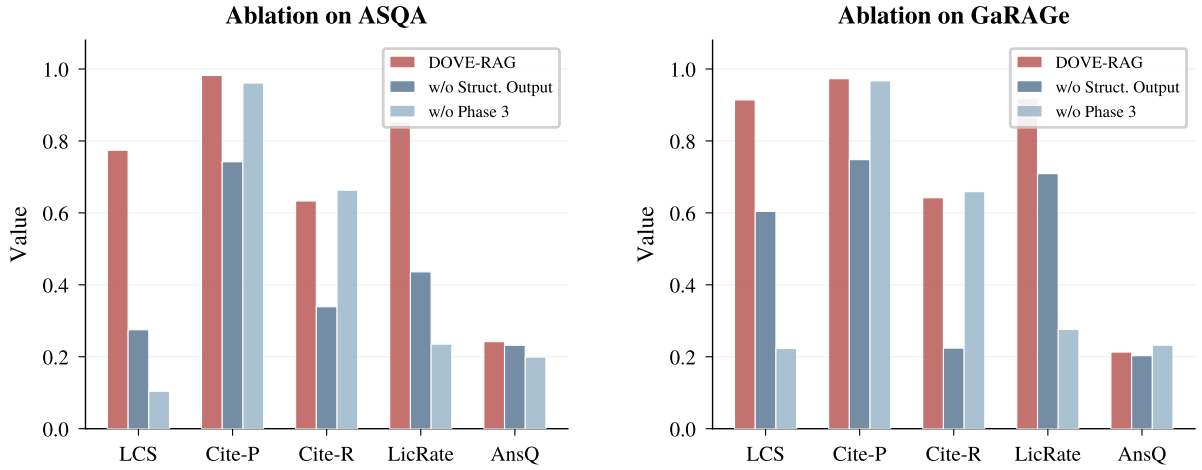


Figure 5: Ablation study on ASQA (left) and GaRAGE (right). Removing structured output causes notable compliance degradation on both datasets.

Table 15: Surface-level dynamic metrics (continued).

Method	MAGIC	
	DelAbs	CA
Vanilla RAG	.751	.748
Dense-RAG	.752	.751
Hybrid-RAG	.751	.749
Long-context	.758	.750
Self-RAG-style	.764	.742
DOVE-RAG	.650	.533

edgment, and their DAR reflects only passive NLI-detected support drops rather than deliberate insufficiency acknowledgment.

C.3 ICS Validation

Figure 6 visualizes the ICS gap between DOVE-RAG and all baselines. All baselines cluster within

Table 16: Bootstrap significance test (10,000 resamples). All baseline ICS values fall below DOVE-RAG’s 95% CI lower bound ($p < 0.001$).

Dataset	DOVE-RAG ICS	95% CI	Best BL	Gap
ASQA	0.649	[0.643, 0.655]	0.454	+43%
ELI5	0.677	[0.671, 0.683]	0.411	+65%
QAMPARI	0.661	[0.655, 0.667]	0.480	+38%
GaRAGE	0.597	[0.579, 0.615]	0.518	+15%
MAGIC	0.671	[0.662, 0.680]	0.600	+12%

ICS=0.371–0.600, while DOVE-RAG achieves ICS=0.597–0.677, representing +12% to +65% improvements.

Table 16 reports bootstrap significance test results (10,000 resamples).

Table 17 shows the ICS gap under alternative weight configurations. DOVE-RAG achieves higher ICS than all baselines under every weight

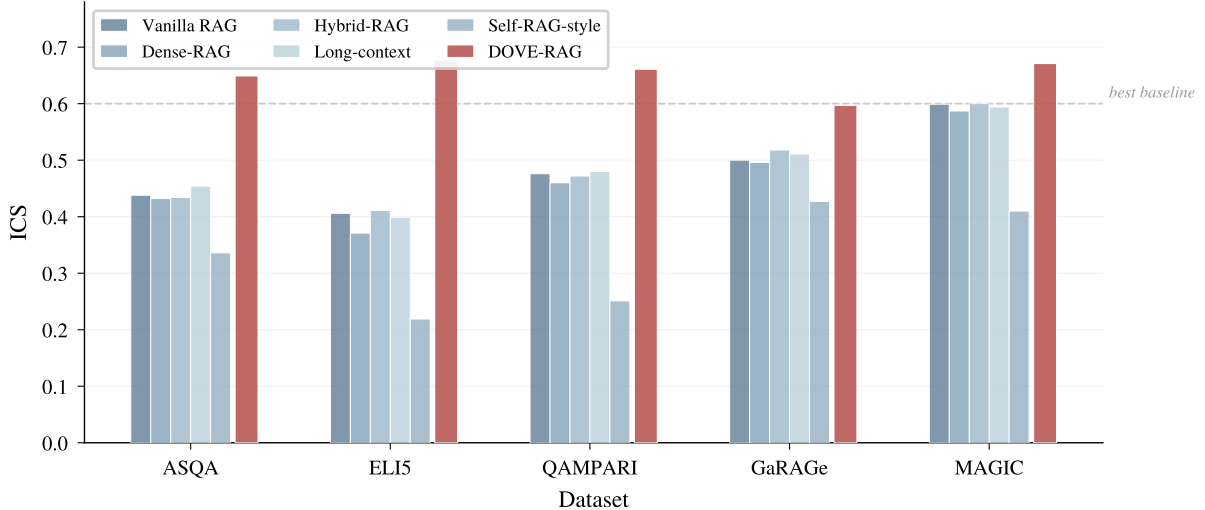


Figure 6: ICS across datasets. The dashed line marks the best baseline ICS (≈ 0.600). DOVE-RAG achieves higher ICS than all baselines on every dataset.

Table 17: Weight sensitivity. ICS gap (DOVE-RAG – best baseline) under alternative configurations. Gaps are largely positive, supporting ranking invariance.

Dataset	Equal	Conflict	Delete	Stability
ASQA	+195	+189	+176	+228
ELI5	+266	+287	+255	+276
QAMPARI	+181	+187	+139	+220
GaRAGe	+080	+103	+058	+089
MAGIC	+071	-.005	+129	+103

configuration on 4/5 datasets. On MAGIC, the conflict-heavy configuration yields a small negative gap (-.005), because baselines achieve higher passive CDR (0.775–0.847) than DOVE-RAG (0.505) via NLI-detected refutation signals without explicit acknowledgment; under equal weighting, DOVE-RAG still achieves higher ICS via superior DAR and DI. On ALCE and GaRAGe, the conflict-heavy configuration yields the largest gap, reflecting DOVE-RAG’s higher CDR through explicit conflict acknowledgment. The delete-heavy and stability-heavy configurations yield the largest gaps on MAGIC, reflecting DOVE-RAG’s superior DAR and DI.

We sweep each threshold in Eq. 1 while holding others at default values ($\tau_s = 0.70$, $\tau_r = 0.20$, $\tau_\rho = 0.90$). Table 18 reports LCS and ContrRate on ASQA. τ_s and τ_r have negligible impact ($\Delta\text{LCS} < 0.001$), as most sentences receive confident NLI predictions well above/below the thresholds. τ_ρ affects LCS more notably because it controls overcitation classification: a low threshold aggressively marks supported-but-redundant claims as OVER, re-

Table 18: Threshold-sweep variant on ASQA. The first row is a threshold-sweep baseline, not the default LCS reported in Table 11; varying τ_s and τ_r has minimal impact, while τ_ρ affects LCS via overcitation classification.

Config	LCS $\uparrow$	ContrRate $\downarrow$	AnsQ $\uparrow$
Threshold-sweep baseline	.277	.440	.177
$\tau_s = 0.50$	.277	.439	.177
$\tau_s = 0.80$	.276	.440	.177
$\tau_r = 0.10$	.277	.440	.177
$\tau_r = 0.30$	.277	.440	.177
$\tau_\rho = 0.50$	.103	.440	.176
$\tau_\rho = 0.95$	.316	.440	.178

ducing LicRate and LCS; a high threshold is more permissive. The default $\tau_\rho = 0.90$ balances precision and recall in overcitation detection.

C.4 Cross-Verifier CDR Validation

To address the circularity concern that DOVE-RAG uses DeBERTa-v3-large for both pipeline verification and CDR evaluation, we recompute CDR using RoBERTa-large-MNLI (Liu et al., 2019), an independently trained NLI model with no architectural or training data overlap with DOVE-RAG’s pipeline. Table 19 reports CDR under both verifiers.

RoBERTa yields systematically higher CDR than DeBERTa for both baselines (0.538–0.993 vs. 0.283–0.847) and DOVE-RAG (0.934–0.993 vs. 0.505–0.999), likely because RoBERTa-large-MNLI is more sensitive to contradiction signals. Despite this shift in absolute values, the Spearman

Table 19: Cross-verifier CDR comparison. DeBERTa-v3-large vs. RoBERTa-large-MNLI. The relative ranking across baselines is preserved ($\rho = 0.84$, $p < 0.001$); DOVE-RAG consistently exceeds all baselines under both verifiers.

Method	ASQA		ELI5	
	DeBERTa	RoBERTa	DeBERTa	RoBERTa
Vanilla RAG	.796	.951	.587	.930
Dense-RAG	.736	.919	.510	.897
Hybrid-RAG	.789	.933	.610	.940
Long-context	.831	.968	.533	.933
Self-RAG-style	.817	.979	.687	.963
DOVE-RAG	.997	.989	.984	.940

Method	QAMPARI		GaRAGe		MAGIC	
	DeBERTa	RoBERTa	DeBERTa	RoBERTa	DeBERTa	RoBERTa
Vanilla RAG	.790	.953	.311	.538	.812	.984
Dense-RAG	.670	.930	.283	.632	.788	.989
Hybrid-RAG	.793	.967	.396	.632	.815	.979
Long-context	.787	.963	.349	.717	.775	.987
Self-RAG-style	.810	.993	.396	.849	.847	.987
DOVE-RAG	.999	.993	.594	.934	.505	.981

rank correlation between the two verifiers across all 25 baseline method–dataset pairs is $\rho = 0.84$ ($p < 0.001$), indicating that the relative ordering of baselines is preserved. Crucially, DOVE-RAG achieves CDR of 0.934–0.993 under RoBERTa, exceeding all baselines on every dataset—particularly on GaRAGe (0.934 vs. 0.538–0.849). This confirms that DOVE-RAG’s advantage is robust to verifier choice.

C.5 Semantic Metric Validation

A potential concern with ICS is that DI and PS rely on ROUGE-L, a lexical similarity metric that may not capture semantic equivalence: a semantic-preserving paraphrase could be penalized, while template-preserving nonsense could score high. To assess this, we compute sentence-level semantic similarity using DeBERTa-v3-large embeddings (cosine similarity) alongside ROUGE-L for distract and permutation interventions on ASQA (200 samples per condition).

Table 20 reports the results. Two observations emerge: (1) ROUGE-L and semantic similarity produce consistent rankings, with DOVE-RAG outperforming Vanilla RAG on both metrics under both interventions, confirming that the ROUGE-L-based DI/PS metrics do not introduce ranking inversions; (2) ROUGE-L values are notably lower than semantic similarity (e.g., Vanilla RAG distract: ROUGE-L=0.174 vs. Semantic=0.723), indicating that ROUGE-L is a conservative estimate of stability. This conservatism is desirable for ICS: it means DI and PS provide a lower bound on behavioral

Table 20: ROUGE-L vs. semantic similarity on ASQA. Rankings are consistent; ROUGE-L is a conservative lower bound.

Method	Distract		Permutation	
	ROUGE-L	Semantic	ROUGE-L	Semantic
Vanilla RAG	.174	.723	.259	.729
DOVE-RAG	.534	.924	.418	.891

Table 21: Per-sample latency (seconds).

Dataset	Phase 1	Phase 2+3	Total	Samples
ASQA	48.2	3.5	51.7	948
ELI5	71.1	2.7	73.8	1000
QAMPARI	30.4	2.6	33.0	1000
GaRAGe	9.1	2.6	11.7	355
MAGIC	8.9	2.8	11.7	1245

stability, and systems that pass ROUGE-L-based stability checks are also semantically stable.

D Reproducibility and Licenses

NVIDIA A800-SXM4 80 GB GPUs, BF16/FP32. NLI verification adds 1–3 s/sample. Total: ~50s/sample (ASQA), ~74s/sample (ELI5). Table 21 reports per-sample latency. Seeds 42 and 123 (greedy decoding), publicly available checkpoints. All thresholds and prompts are dataset-agnostic with no per-dataset tuning. The full pipeline is deterministic across seeds (Table 9).

All models used in this work are released under permissive licenses: Qwen3-8B under Apache 2.0, Phi-3-mini-4k-instruct and DeBERTa-v3-large under MIT, and BGE-M3 under MIT. Among the evaluation datasets, QAMPARI is released under CC0-1.0 (public domain), ELI5 under BSD-3-Clause, and MAGIC under CC-BY-SA 4.0. GaRAGe is released under CC-BY-NC-4.0, which restricts use to non-commercial purposes; our use of GaRAGe is solely for academic research and complies with this restriction. ASQA does not specify an explicit license in its original repository; we access it through the ALCE benchmark (MIT) and TensorFlow Datasets, consistent with standard academic practice.