REGIME BOUNDARY ALIGNMENT FOR EVIDENCE-GATED QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented language models are expected to answer from the retrieved evidence, but in practice they often keep answering when that evidence is missing. We trace this behavior to the training signal: answer-focused fine-tuning assigns no target to unsupported contexts, so it cannot distinguish a reader that abstains from one that guesses, and unsupported answering stays near 100% even as supported accuracy improves. We introduce Regime Boundary Alignment (RBA), which trains a single reader on matched variants of the same question and gold answer. The reader is trained to produce the gold answer when the context supports it, including when conflicting evidence is also present, and to abstain when the correct support is removed; inference is ordinary decoding, with no verifier, threshold, or regime label. On three multi-hop QA datasets across three seeds, RBA reduces the unsupported-answer rate by more than sixty percentage points relative to conflict-focused training while matching its supported accuracy. On a held-out TriviaQA retrieval-miss slice, the same reader reduces unsupported answering from 100% to below 1% while also improving supported accuracy. These results indicate that evidence-gated answering must be learned on both sides of the support boundary.

1 INTRODUCTION

Retrieval-augmented generation (RAG) is intended to ground answers in retrieved evidence (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021). Retrieval, however, may omit the answer-bearing passage or surface related distractors. Under an evidence-grounded contract, a reader should not silently fill such gaps from memory. Parametric knowledge may help interpret the context, but it does not by itself license an answer.

Existing approaches address parts of this contract. Prompting and uncertainty estimation rely on instructions, confidence estimates, or tuned thresholds (Ren et al., 2023; Kuhn et al., 2023); conflict-supervised training preserves a supported answer amid distractors; capability- and knowledge-aware post-training draws the decision boundary from estimated model knowledge (Franzmeyer et al., 2026; Sun et al., 2025); refusal-only supervision can minimize unsupported answering by refusing indiscriminately. What they share is that supervision attaches to answers: none of them specifies a target action for the case in which correct support is absent.

The missing action defines a boundary. For a fixed question and gold answer, the reader should answer when the retrieved context contains sufficient correct support and abstain otherwise. Correct support plus a distractor remains answerable, since disagreement alone is not a reason to refuse (Figure 1). Standard answer-focused fine-tuning under-specifies this boundary: it rewards supported answering, and the same supported-answer loss is consistent with a reader that abstains on unsupported input and with a reader that guesses.

This gap is both formal and empirical. On finite observed regimes, answer-only likelihood leaves the no-support conditional unidentified, while a sequence loss with a matched abstention term controls a weighted upper bound on the pairwise answer–abstain boundary violations (Propositions 1 and 2). The empirical counterpart matches: on four held-out datasets, answer-focused conflict training reaches 66.13% supported accuracy and still answers every no-support example.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

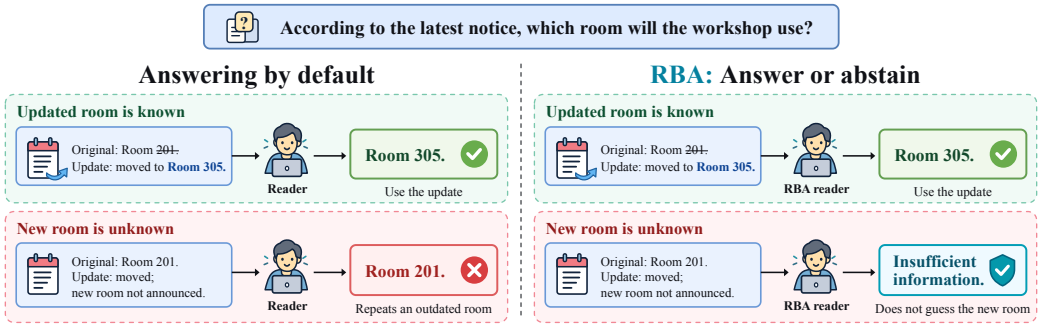

Figure 1: The answerability boundary on one example. A notice that names the new room supports answering with it; a notice that confirms only the move does not, and repeating the old room is then unsupported. A reader trained only to answer (left) answers in both cases, while the RBA reader (right) answers with support and abstains without it.

We introduce Regime Boundary Alignment (RBA), which trains a single reader on matched evidence conditions. For a fixed (q, g), C^+ contains correct support, $C^\pm$ adds conflicting evidence without removing that support, and C^0 removes the correct support while keeping the distractor. The reader learns the gold answer on C^+ and $C^\pm$ and a fixed abstention on C^0 under one joint likelihood objective, and at inference it sees only the question and context, with no verifier, tuned threshold, or regime label.

Trained this way, the reader reduces the no-support unsupported-answer rate (UAR) from 100% to 0% while Supported Gold stays at 66.13%; a paired bootstrap estimates the Supported Gold change at -0.58 percentage points with a 95% interval of $[-1.84, 0.67]$. Controls attribute the separation to the matched supervision rather than a generic refusal prior: replacing the query-aligned C^0 with unrelated contexts leaves UAR at 66.58%, and removing the answer anchor collapses Supported Gold to zero. The same -100-point UAR change repeats on Llama-3.1-8B and Mistral-7B readers, and the learned boundary transfers to dense-retrieval misses.

Contributions.

1. We formulate evidence-gated QA through a sequence-level answer–abstain boundary and construct matched retrieval regimes that vary support while holding the question and gold answer fixed.
2. We introduce RBA, a single joint likelihood objective, and show that it identifies the observed no-support action and upper-bounds pairwise boundary violations, without an inference-time verifier, threshold, or regime label.
3. Across four datasets, three seeds, and three reader families, RBA eliminates unsupported answering while preserving supported accuracy, and paired, component, and distribution-shift experiments attribute the effect to the matched two-sided supervision.

2 Related Work

Evidence quality and acquisition. RAG and retrieval-augmented pretraining use external documents to supplement parametric knowledge (Lewis et al., 2020; Guu et al., 2020). One line of follow-up work measures how readers respond to imperfect evidence, covering irrelevant passages, evidence position, and retrieval noise (Yoran et al., 2024; Liu et al., 2024; Cuconasu et al., 2024; Shen et al., 2024); RGB explicitly evaluates negative rejection alongside noise robustness, information integration, and counterfactual robustness (Chen et al., 2024), and RAGLens uses sparse-autoencoder features to detect unfaithful generations for post-hoc revision (Xiong et al., 2026). A second line modifies or extends the evidence itself: Self-RAG learns retrieval and critique actions (Asai et al., 2024), CRAG evaluates retrieved evidence and triggers correction (Yan et al., 2024), RECOMP compresses context and can omit unhelpful augmentation (Xu et al., 2024), Chain-of-Note assesses retrieved documents sequentially (Yu et al., 2024), and KnowGuard grounds clinical

abstention in iterative knowledge-graph exploration across conversation turns (Dang et al., 2026). Both lines treat the evidence as the moving part, whether to evaluate it, repair it, or extend it. RBA instead holds the supplied context fixed and trains what the reader does with it: answer when the context supports the gold answer, and abstain when it does not.

Contextual and parametric knowledge. Controlled substitutions expose how readers arbitrate between retrieved statements and memorized facts (Longpre et al., 2021; Neeman et al., 2023; Xie et al., 2024). Contrastive and adaptive decoding adjust their relative influence during generation (Shi et al., 2024; Kim et al., 2024; Wang et al., 2025b; Yuan et al., 2024). FaithfulRAG, Astute RAG, and MADAM-RAG resolve fact-level or document-level conflicts (Zhang et al., 2025; Wang et al., 2025a;c). Knowledgeable-R1 uses reinforcement learning to preserve parametric knowledge as a fallback under misleading retrieval (Lin et al., 2026). These methods decide which source to trust when memory and evidence compete. RBA studies the complementary case in which the required evidence is absent entirely: conflict with surviving support remains answerable, while a context without correct support receives an abstention target during training.

Abstention and selective prediction. Unanswerable QA and selective QA study withholding answers under missing evidence or distribution shift (Rajpurkar et al., 2018; Kamath et al., 2020). LLM uncertainty methods estimate correctness through self-evaluation, verbalized confidence, or semantic uncertainty (Ren et al., 2023; Lin et al., 2022; Kuhn et al., 2023; Kadavath et al., 2022); the Refusal Index measures whether refusal probability tracks error probability (Pan et al., 2026), and SAFER calibrates sampling and filtering to control miscoverage risk in open-ended QA (Wang et al., 2026). Abstention can also be trained directly: DTA partitions examples by the availability of internal and retrieved knowledge and constructs quadrant-specific DPO preferences, including abstention when neither source suffices (Sun et al., 2025), and HALT constructs capability-aligned fine-tuning responses by removing incorrect factual fragments or inserting uncertainty markers (Franzmeyer et al., 2026). RBA differs from the first group in that abstention is a generated response learned in training, not a post-hoc threshold on confidence; it differs from the second in how the action is defined—the question and gold answer are held fixed, answers are supervised under supported and mixed contexts, and abstention is supervised on the matched context with correct support removed.

3 Problem Formulation

3.1 Evidence-grounded action boundary

Let q be a question, $g = (g_1, \ldots, g_{|g|})$ a gold answer sequence, and $C = \{d_1, \ldots, d_k\}$ the retrieved context. We reserve $a_0 =$ "Insufficient information." as the abstention sequence and assume $g \neq a_0$ by construction. An autoregressive reader assigns every candidate sequence a the score

$$S_\theta(a \mid q, C) = \sum_{t=1}^{|a|+1} \log p_\theta(a_t \mid q, C, a_{<t}), \qquad a_{|a|+1} = \text{EOS}. \tag{1}$$

Thus S_θ is a full sequence log probability, not a first-token score. The answer–abstain log-odds for a labeled pair (q, g, C) are

$$m_\theta(q, g, C) = S_\theta(g \mid q, C) - S_\theta(a_0 \mid q, C) = \log \frac{p_\theta(g \mid q, C)}{p_\theta(a_0 \mid q, C)}. \tag{2}$$

A positive margin prefers the gold answer to the fixed abstention; a negative margin reverses that preference. This margin is an analytical device that uses g during training and analysis; it is not computed as an inference-time gate.

Let $s(C, g) \in \{0, 1\}$ indicate whether C contains sufficient correct support for g under the task contract. The desired evidence-gated action and the corresponding pairwise sign condition are

$$a^\star(q, g, C) = \begin{cases} g, & s(C, g) = 1, \\ a_0, & s(C, g) = 0, \end{cases} \qquad (2s(C, g) - 1) m_\theta(q, g, C) > 0. \tag{3}$$

Here support concerns the supplied evidence, not whether the reader can recall g parametrically. Parametric knowledge may help interpret a passage, but it does not license an unsupported answer in this setting.

3.2 MATCHED CONTEXT INTERVENTIONS

For a fixed (q, g), let S denote answer-bearing support documents and D denote query-related distractors. We construct one matched triplet $z = (q, g, C^+, C^\pm, C^0)$ through

$$C^+ = S, \qquad C^\pm = S \cup D, \qquad C^0 = D. \tag{4}$$

Consequently, $s(C^+, g) = s(C^\pm, g) = 1$ and $s(C^0, g) = 0$. The contrast $C^+ \to C^\pm$ adds conflict without removing evidence, whereas $C^\pm \to C^0$ removes the answer-bearing support while retaining the distractor. Question identity, answer identity, and distractor source are held fixed. This matching prevents the target action from being explained solely by different question populations. The three desired inequalities can be summarized by the triplet boundary clearance

$$B_\theta(z) = \min\left\{m_\theta(q, g, C^+),\ m_\theta(q, g, C^\pm),\ -m_\theta(q, g, C^0)\right\}. \tag{5}$$

$B_\theta(z) > 0$ means that one reader places this observed triplet on the correct sides of the answerability boundary.

Both C^+ and $C^\pm$ target g: the clean context preserves ordinary evidence-conditioned answering, and the mixed context prevents disagreement alone from becoming a refusal cue. Only C^0 targets a_0. We instantiate these triplets from 2WikiMultihopQA, HotpotQA, MuSiQue, and TriviaQA (Ho et al., 2020; Yang et al., 2018; Trivedi et al., 2022; Joshi et al., 2017). Benchmark-annotated supporting documents form C^+; adding the paired adversarial documents yields $C^\pm$; retaining those distractor documents while removing all annotated support yields C^0. We discard rows for which either side cannot be constructed. Because the three contexts share (q, g) and the same distractor source, the main change is the availability of correct support rather than the question distribution.

4 REGIME BOUNDARY ALIGNMENT

Section 3 leaves the no-support action without a training signal. RBA supplies it directly, fitting the three target actions of Section 3 with one conditional generator: answering is anchored on both supported regimes and abstention on the matched no-support regime, with no separate classifier, router, or verifier.

4.1 SEQUENCE-LEVEL JOINT OBJECTIVE

For target a, define the teacher-forced sequence loss

$$\ell_\theta(a; q, C) = -S_\theta(a \mid q, C) = -\sum_{t=1}^{|a|+1} \log p_\theta(a_t \mid q, C, a_{<t}). \tag{6}$$

Prompt tokens are masked and EOS is included. We sum rather than length-average target-token NLLs.

For one matched triplet, RBA minimizes

$$\mathcal{L}_{\text{RBA}}(\theta; z) = \underbrace{\ell_\theta(g; q, C^+) + \alpha\, \ell_\theta(g; q, C^\pm)}_{\mathcal{L}_{\text{emit}}} + \underbrace{\gamma\, \ell_\theta(a_0; q, C^0)}_{\mathcal{L}_{\text{withhold}}}. \tag{7}$$

The clean term preserves ordinary evidence-conditioned answering. The mixed term preserves that action in the presence of conflict, so conflict alone does not become a refusal cue. The withhold term supplies the otherwise missing action at the matched retrieval failure. The main setting uses $\alpha = 1$ and $\gamma = 0.5$.

For training triplets $\mathcal{D} = \{z_i\}_{i=1}^N$, including the appended hard-replay triplets, the optimized empirical risk is

$$\widehat{\mathcal{R}}_{\text{RBA}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{RBA}}(\theta; z_i). \tag{8}$$

Hard replay changes which triplets occur in $\mathcal{D}$, not the objective. The decisive Conflict-SFT baseline is exactly the same answer anchor with the withhold coefficient removed:

$$\mathcal{L}_{\text{conflict}}(\theta; z) = \mathcal{L}_{\text{RBA}}(\theta; z)\big|_{\gamma=0} = \mathcal{L}_{\text{emit}}. \tag{9}$$

The RBA–Conflict-SFT comparison therefore isolates whether observing the no-support action changes the boundary while retaining the same two answer targets.

4.2 IDENTIFICATION AND BOUNDARY CONTROL

Proposition 1. On a finite set of observed triplets, suppose targets are deterministic, no identical input is assigned conflicting targets, and the conditional sequence distributions are unrestricted at each observed input. Minimizing $\sum_i \mathcal{L}_{\text{emit}}(\theta; z_i)$ leaves every observed $p_\theta(\cdot \mid q, C^0)$ unidentified. For $\alpha, \gamma > 0$, the pointwise minimizer of $\sum_i \mathcal{L}_{\text{RBA}}(\theta; z_i)$ assigns unit mass to g on $C^+, C^\pm$ and to a_0 on C^0.

Proof. $\sum_i \mathcal{L}_{\text{emit}}(\theta; z_i)$ contains no term evaluated at C^0, so two conditional distributions that agree on C^+ and $C^\pm$ have identical loss regardless of their C^0 actions. In the summed RBA objective, every observed regime has a positive-weight negative log likelihood. Over the probability simplex, each term is minimized by assigning unit mass to its deterministic target. □

Proposition 1 is an identification statement, not a claim that a finite neural optimizer reaches the simplex optimum. The next result connects the implemented sequence loss to the boundary signs in Equation 5. Let $\mathcal{K} = \{+, \pm, 0\}$ index the regimes, with $y_+ = y_\pm = 1$, $y_0 = -1$ and weights $w_+ = 1, w_\pm = \alpha, w_0 = \gamma$. Write m_r for the corresponding margin from Equation 2 and define the weighted violation count

$$V_\theta(z) = \sum_{r \in \mathcal{K}} w_r \, \mathbf{1}[y_r m_r \le 0] \,. \tag{10}$$

Proposition 2. For any autoregressive sequence distribution satisfying $p_\theta(g \mid q, C_r) > 0$ and $p_\theta(a_0 \mid q, C_r) > 0$ for every $r \in \mathcal{K}$, and any $\alpha, \gamma \ge 0$,

$$V_\theta(z) \le \frac{\mathcal{L}_{\text{RBA}}(\theta; z)}{\log 2}. \tag{11}$$

Proof. For regime r, normalize the model probabilities of only g and a_0. The target's binary negative log probability is $\log(1 + \exp(-y_r m_r))$. It is no larger than the full sequence NLL because $p_\theta(g \mid q, C_r) + p_\theta(a_0 \mid q, C_r) \le 1$. Whenever $y_r m_r \le 0$, this binary loss is at least $\log 2$. Multiplying by w_r and summing gives Equation 11. □

Thus each term has a distinct geometric job: raise the two supported margins and reverse the no-support margin. The bound shows that the implemented sequence NLL is a valid surrogate for the three observed pairwise boundary decisions.

Algorithm 1 One Regime Boundary Alignment update.

Require: Batch $\mathcal{B}$ of $(q, g, C^+, C^\pm, C^0)$
$\mathcal{J} \leftarrow 0$
for $(q, g, C^+, C^\pm, C^0) \in \mathcal{B}$ **do**
  $\ell_+ \leftarrow \ell_\theta(g; q, C^+)$; $\ell_\pm \leftarrow \ell_\theta(g; q, C^\pm)$
  $\ell_0 \leftarrow \ell_\theta(a_0; q, C^0)$
  $\mathcal{J} \leftarrow \mathcal{J} + \ell_+ + \alpha \ell_\pm + \gamma \ell_0$
end for
$\theta \leftarrow \text{OptimizerStep}(\theta, \nabla_\theta \mathcal{J} / |\mathcal{B}|)$

Why the no-support context is query-aligned. For triplet i, compare the aligned and shuffled constructions

$$C_i^0 = D_i, \qquad \widetilde{C^0}_i = D_{\sigma(i)}, \quad \sigma(i) \ne i. \tag{12}$$

RBA uses the distractors D_i paired with q_i. The unaligned alternative instead uses a permutation across questions. Its loss constrains $p_\theta(a_0 \mid q_i, \widetilde{C^0}_i)$, not $p_\theta(a_0 \mid q_i, C_i^0)$. Without assumptions linking those inputs, low unaligned empirical risk does not identify the desired matched action. Query alignment therefore determines where the abstention anchor is placed; RQ2 tests whether parameter sharing transports a generic refusal association to the matched boundary.

4.3 TRAINING ALGORITHM AND COMPLEXITY

Algorithm 1 gives one optimization step. The three contexts for an example remain together in the same minibatch, so the answer and abstention signals update the same conditional generator under the same (q, g). Sequence log probabilities are computed with teacher forcing and include the EOS token.

Let n_+, $n_\pm$, and n_0 denote the prompt-plus-target lengths for the three regimes, and let $T_\theta(n)$ be the cost of scoring a length-n sequence. RBA scores three sequences per triplet, for $O(T_\theta(n_+) + T_\theta(n_\pm) + T_\theta(n_0))$ work; concatenated along the batch dimension, this is arithmetic work rather than three serial model calls. For a dense L-layer Transformer of width d, $T_\theta(n) = O(L(n^2 d + n d^2))$; RBA changes only the constant number of scored sequences, not this asymptotic dependence. With LoRA rank r, trainable parameters scale as $O(r \sum_j (d_{j,\text{in}} + d_{j,\text{out}}))$ over adapted matrices. At inference, RBA performs one ordinary reader generation and adds no retrieval pass, verifier, or routing model.

Table 1: **Main result and mechanism controls.** Each dataset cell and Macro report Supported Gold/UAR (G/U); abstention F1, AURC, and C@10 pool paired supported/no-support rows. Values are seed medians and percentages except AURC. Base and Prompt are deterministic single-checkpoint evaluations; all trained rows use three seeds. The best displayed value within each individual metric is bold (lower is better for UAR and AURC); displayed ties are included.

Training condition	2Wiki G/U	HotpotQA G/U	MuSiQue G/U	TriviaQA G/U	Macro G/U	Abst.F1↑	AURC↓	C@10↑
Reference and answer-focused baselines								
Base	0.78/100.00	2.03/100.00	0.00/100.00	8.59/100.00	2.85/100.00	0.00	.980	0.00
Prompt	27.34/8.59	46.70/17.26	9.09/7.14	37.11/99.61	30.06/33.15	63.24	.304	11.36
DTA	27.73/100.00	45.69/100.00	12.34/100.00	49.61/100.00	33.74/100.00	0.00	.707	0.00
Conflict-SFT (no C^0)	**72.27**/100.00	**64.97**/100.00	70.78/100.00	55.47/100.00	**66.13**/100.00	0.00	.486	1.33
Boundary-supervision controls								
No answer anchor	0.00/**0.00**	0.00/**0.00**	0.00/**0.00**	0.00/**0.00**	0.00/**0.00**	66.67	.680	0.00
Unaligned C^0	**72.27**/57.81	63.45/58.88	70.13/50.00	**55.86**/99.22	65.62/66.58	47.13	.284	16.86
No hard replay	70.70/**0.00**	63.96/**0.00**	69.48/**0.00**	52.34/**0.00**	64.64/**0.00**	99.83	.039	86.73
RBA	71.09/**0.00**	63.45/**0.00**	**71.43/0.00**	54.30/**0.00**	**66.13/0.00**	**100.00**	**.037**	**86.96**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Registered questions and data. The frozen campaign asks whether RBA changes the boundary (RQ1), which supervision is load-bearing (RQ2), where it generalizes (RQ3), and whether the direction repeats across readers (RQ4). Training uses 2,000 natural BM25 triplets plus 147 hard TriviaQA triplets, with no evaluation-question overlap. The paired test set contains 256 2Wiki, 197 HotpotQA, 154 MuSiQue, and 256 TriviaQA questions, each with supported and BM25 no-support views. RQ3 freezes the RQ1 checkpoints and adds paired dense-retrieval views (95, 157, 232, and 256 questions for the four datasets), 256 partial-evidence examples per multi-hop dataset, and 1,239 controlled seen, unseen-template, and semantic-negation examples. Partial contexts retain one supporting document but lack the remaining multi-hop chain.

Models and comparisons. The primary reader is Qwen3-8B in non-thinking mode. Base is unadapted, Prompt adds an abstention instruction, Conflict-SFT uses Equation 9, and DTA is a matched-capacity quadrant-preference baseline (Sun et al., 2025). RQ2 removes or misaligns individual RBA components; RQ4 repeats the decisive contrast with Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. All readers are rank-16 LoRA adapters (scale 32, dropout 0.05) trained for six epochs with AdamW at learning rate 10^{-4}, with an effective batch size of 8 over 4 ranks. Training sequences are left-truncated at 3,072 tokens, and padded positions carry zero loss weight in every compared objective, so all rows supervise the same tokens.

Evaluation and aggregation. Generation is greedy with the training prompt, 3,000 input tokens, and at most 48 new tokens. Fixed lexical rules identify abstentions. Supported Gold is normalized exact match; UAR is the answered fraction of no-support or partial examples. Abstention F1 treats

no-support as positive. Selective confidence is the generated-sequence geometric-mean probability; lower AURC and higher 10%-risk coverage are better. We macro-average datasets within each seed, then report median and range over seeds 42, 137, and 271; Base and Prompt are deterministic. F1, AURC, and coverage instead pool matched positive/negative rows within seed. The decisive contrast uses 20,000 question-stratified paired-bootstrap resamples, averaging over seeds and macro-averaging datasets.

5.2 RQ1: Learning the support boundary

Table 1 shows the per-dataset consistency and joint trade-off. Conflict-SFT reaches 66.13% macro Gold but answers every retrieval miss; RBA retains 66.13% Gold and answers none of them. The UAR change from 100% to 0% occurs on every dataset, and the minimum RBA abstention F1 over seeds is 99.88%. The identical macro medians hide a small per-dataset pattern: RBA gains 0.65 Gold points on MuSiQue and gives up between 1.17 and 1.52 on the other three datasets, which is why the medians coincide while the paired bootstrap below estimates a small negative mean shift. Pooled F1 rises from 0% to 100%, AURC falls from .486 to .037, and 10%-risk coverage rises from 1.33% to 86.96%. No other evaluated row keeps supported accuracy while abstaining on retrieval misses.

The paired bootstrap quantifies the accuracy side of this trade-off. Across the four datasets, RBA minus Conflict-SFT is -0.58 percentage points in supported Gold (95% interval $[-1.84, 0.67]$) and -100 points in UAR (interval $[-100, -100]$): supported accuracy is statistically indistinguishable between the two readers, while the unsupported-answer rates are completely separated.

The remaining rows fail in instructively different ways. Prompting lowers UAR on the three multi-hop datasets (7.14–17.26%) but barely moves it on TriviaQA (99.61%), where questions are single-hop factoids and a memorized answer is most available; the abstention instruction loses exactly where the model has an answer in memory. DTA fails on both axes at once: it answers every no-support example and reaches only 33.74% macro Gold, so quadrant-specific preferences neither fit the abstention action nor preserve reading accuracy in this setting. The no-answer-anchor row marks the opposite extreme, and its abstention F1 of 66.67% is a reminder that F1 alone does not certify the boundary: on balanced paired views, a reader that abstains on everything attains two-thirds F1, which is why we read F1 together with Gold and UAR. Finally, the selective metrics show that the learned boundary is reflected in confidence, not only in the emitted action. Conflict-SFT's sequence confidence separates supported from no-support inputs only weakly (.486 AURC, 1.33% coverage at 10% risk), while RBA's separates them almost perfectly (.037 AURC, 86.96% coverage), so a practitioner could threshold RBA's confidence and recover most of the risk–coverage trade-off without retraining.

5.3 RQ2: Which supervision is load-bearing?

The lower block of Table 1 exposes opposite degenerate policies under the two one-sided objectives. No-answer-anchor training always abstains, producing 0% Gold and 0% UAR; removing C^0 always answers, producing 66.13% Gold and 100% UAR. Both targets are therefore necessary to identify the desired action boundary.

The context attached to the abstention target matters as well. Replacing query-aligned C^0 with a random wrong document preserves 65.62% Gold but leaves UAR at 66.58%: the reader learns a generic refusal association, not the question-conditioned boundary. The failure is again uneven across datasets: unaligned UAR stays between 50.00% and 58.88% on the three multi-hop datasets but reaches 99.22% on TriviaQA, the same dataset where prompting fails. An instruction and a misaligned refusal target thus break down at the same place, where the question most strongly invites a memorized answer. Removing the 147 hard TriviaQA replay examples keeps median UAR at 0% (at most 0.26% across seeds) and lowers Gold to 64.64%, so hard replay contributes about 1.5 points of supported accuracy but does not carry the abstention effect. The drop concentrates where the replay triplets come from: TriviaQA loses 1.96 Gold points and MuSiQue 1.95, while HotpotQA is unchanged within half a point.

Boundary geometry. Figure 2 scores the fixed gold and abstention sequences on the paired held-out views, using the per-question margin of Equation 2 averaged over the three frozen seeds.

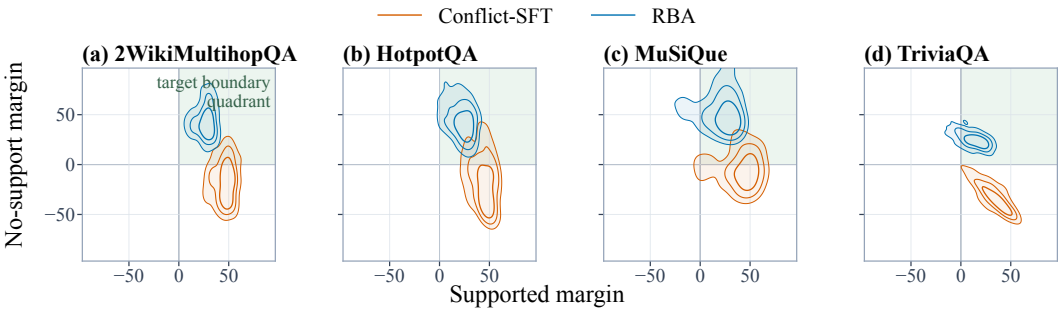

Figure 2: **Evidence-boundary margin distributions.** Each contour is a Gaussian KDE over held-out (q, g) pairs after averaging exact teacher-forced sequence margins within question across the three frozen seeds. The horizontal axis is $S(g \mid q, C^+) - S(\text{IDK} \mid q, C^+)$; the vertical axis is $S(\text{IDK} \mid q, C^0) - S(g \mid q, C^0)$. The shaded upper-right quadrant therefore represents the desired joint preference.

Conflict-SFT's supported margin is positive, so it does prefer the gold answer when support is present; its no-support margin, however, is predominantly negative, meaning the reader still prefers the gold string over abstention after support has been removed. Unsupported answering under Conflict-SFT is therefore an active preference learned from a signal that never scored the alternative, not a near-tie resolved by surface priors. RBA shifts the density into the positive-positive quadrant on all four datasets, and because both margins are per question, the quadrant mass reflects jointly correct decisions rather than two separately correct averages.

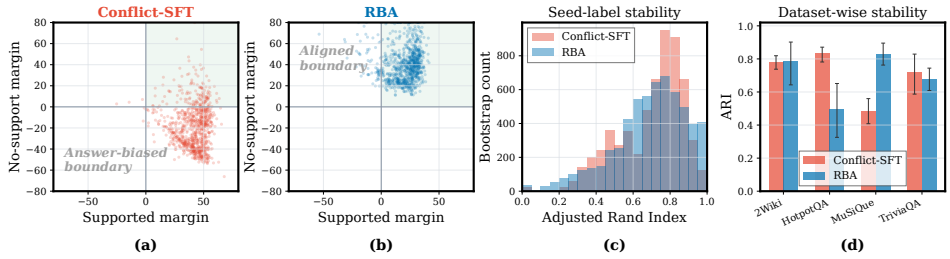

Figure 3: **Boundary-state geometry and seed stability.** Panels (a–b) show seed-averaged paired margin states for Conflict-SFT and RBA. Panels (c–d) compute adjusted Rand index (ARI) between boundary-quadrant labelings of the same held-out questions across seed pairs, with bootstrap resampling for the histogram and interquartile bars.

Seed stability. Figure 3 checks that these boundaries are properties of the trained readers rather than of a lucky seed: the quadrant labelings of the same held-out questions are stable across seed pairs for both readers, overall and per dataset. Conflict-SFT's answer-biased boundary is as stable as RBA's evidence-gated one, so the two readers differ in where the boundary sits, not in whether training converges.

5.4 RQ3: Where does the boundary generalize?

The boundary transfers across retrieval algorithms: dense-miss UAR is 0.10% for RBA versus 100% for Conflict-SFT, with Gold at 62.10% versus 61.70%. Because the dense views come from a different retriever, this transfer indicates that the reader keys on the absence of support rather than on surface properties of BM25 retrieval failures. It does not transfer broadly across semantic forms of insufficiency. RBA UAR is 83.07% with partial evidence and 97.62%, 76.50%, and 99.30% on controlled seen, unseen-template, and negation shifts. All four conditions stay far above the BM25-miss level: the reader answers most semantic-insufficiency variants regardless of phrasing, and partial evidence — one genuine supporting document with the rest of the chain missing — is answered 83.07% of the time even though the full chain is absent. The unseen-template condition

Table 2: RQ3 frozen-checkpoint transfer. Entries are dataset-macro percentages and seed medians. Paired rows report G/U; no-support-only rows report U. Controlled conditions cover three multi-hop datasets.

Evidence shift	Metric	Base	Conflict-SFT	RBA
BM25 miss	G / U	2.85 / 100.00	66.13 / 100.00	**66.13 / 0.00**
Dense miss	G / U	3.12 / 100.00	61.70 / 100.00	**62.10 / 0.10**
Partial evidence	U	100.00	100.00	**83.07**
Controlled seen	U	100.00	100.00	**97.62**
Unseen template	U	100.00	100.00	**76.50**
Semantic negation	U	100.00	100.00	**99.30**

shows the widest seed dispersion of any RQ3 row (66.25% to 95.36%), so what partial transfer exists there is real but unstable. The learned boundary is therefore robust to the retrieval algorithm but specific to retrieval misses: partially supported and unseen phrasings of insufficiency are still answered a substantial fraction of the time.

5.5 RQ4: Does the direction repeat across readers?

Table 3: RQ4 cross-family comparison. Values are dataset-macro percentages and medians over three seeds. Deltas are RBA minus Conflict-SFT.

	Conflict-SFT		RBA		Change	
Reader	G	U	G	U	ΔG	ΔU
Qwen3-8B	66.13	100.00	66.13	0.00	+0.01	-100.00
Llama-3.1-8B	67.89	100.00	67.69	0.00	-0.19	-100.00
Mistral-7B	66.76	100.00	66.71	0.00	-0.05	-100.00

All three Conflict-SFT readers answer every no-support example; all RBA readers attain 0% median UAR, with Gold changes within 0.2 points. Two aspects of the table matter beyond the main effect. First, the Gold deltas are small and centered on zero (-0.19, -0.05, $+0.01$), each well inside the Qwen bootstrap interval of $[-1.84, 0.67]$ from RQ1: the cost of the boundary does not grow across families within this size range. Second, the defect being repaired is identical across families — three readers with different pretraining corpora and instruction tuning all arrive at a 100% UAR boundary after answer-focused training. That uniformity supports the paper's central attribution: the missing action is a property of the training signal these readers share, not of any one model family.

6 Conclusion

We studied why retrieval-augmented readers keep answering when the retrieved evidence does not support an answer. The cause is structural: training signals that attach only to answers specify no action for unsupported contexts, so unsupported answering is never penalized, regardless of how accurate the reader becomes on supported questions.

RBA closes this gap by training one reader on matched supported, conflicting, and no-support views of each question, anchoring answering and abstaining in the same likelihood. Across four datasets, three seeds, and three reader families, the reader abstains on retrieval misses while its supported accuracy remains statistically indistinguishable from answer-only conflict training. Controls attribute the effect to the two-sided, question-aligned supervision rather than to a generic refusal association or to the additional training examples.

More broadly, the results suggest that evidence-gated behavior must be supervised on both sides of the support boundary: a reader cannot learn when not to answer from examples in which it always should. The boundary learned here is specific to retrieval misses; extending it to semantic forms of insufficiency, where the reader must judge whether partially relevant evidence suffices, remains open.

AI USE STATEMENT

Generative AI tools were used to edit and polish the writing and to search for and discover related work. The authors take responsibility for the final manuscript, including its citations and claims.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, 2024. doi: 10.1609/aaai.v38i16.29728. URL https://ojs.aaai.org/index.php/AAAI/article/view/29728.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, pp. 719–729. ACM, July 2024. doi: 10.1145/3626772.3657834. URL https://doi.org/10.1145/3626772.3657834.

Xilin Dang, Kexin Chen, Xiaorui Su, Ayush Noori, Iñaki Arango, Lucas Vittor, Xinyi Long, Yuyang Du, Marinka Zitnik, and Pheng-Ann Heng. KnowGuard: Knowledge-driven abstention for multi-round clinical reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gQRefH8upx.

Tim Franzmeyer, Archie Sravankumar, Lijuan Liu, Yuning Mao, Rui Hou, Sinong Wang, Jakob Nicolaus Foerster, Luke Zettlemoyer, and Madian Khabsa. High accuracy, less talk (HALT): Reliable LLMs through capability-aligned finetuning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LYqBnNVaXD.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna

Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2421–2431, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.136. URL https://aclanthology.org/2024.findings-emnlp.136/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2302.09664. URL https://arxiv.org/abs/2302.09664.

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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Chenyu Lin, Yilin Wen, Du Su, Hexiang Tan, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lv. Resisting contextual interference in RAG via parametric-knowledge reinforcement. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6Qc6sO1jh9.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, October 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.559. URL https://aclanthology.org/2023.acl-long.559/.

Wenbo Pan, Jie Xu, Qiguang Chen, Junhao Dong, Libo Qin, Xinfeng Li, Yu Haining, and Xiaohua Jia. Can LLMs refuse questions they do not know? measuring knowledge-aware refusal in factual tasks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9gJBhkLRat.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In Javier Antorán, Arno Blaas, Kelly Buchanan, Fan Feng, Vincent Fortuin, Sahra Ghalebikesabi, Andreas Kriegler, Ian Mason, David Rohde, Francisco J. R. Ruiz, Tobias Uelwer, Yubin Xie, and Rui Yang (eds.), *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pp. 49–64. PMLR, 16 Dec 2023. URL https://proceedings.mlr.press/v239/ren23a.html.

Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. Assessing "implicit" retrieval robustness of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8988–9003, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.507. URL https://aclanthology.org/2024.emnlp-main.507/.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. Divide-then-align: Honest alignment based on the knowledge boundary of RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11461–11480, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.561. URL https://aclanthology.org/2025.acl-long.561/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 11636–11652, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.581. URL https://aclanthology.org/2025.naacl-long.581/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. In *Second Conference on Language Modeling*, 2025c. URL https://openreview.net/forum?id=z1MHB2m3V9.

Qingni Wang, Yue Fan, and Xin Eric Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kJmLmOvwLC.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. Toward faithful retrieval-augmented generation with sparse autoencoders. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hgBZP67BkP.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation, 2024. URL https://arxiv.org/abs/2401.15884.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14672–14685, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.813. URL https://aclanthology.org/2024.emnlp-main.813/.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3903–3922, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.234. URL https://aclanthology.org/2024.findings-acl.234/.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21863–21882, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1062. URL https://aclanthology.org/2025.acl-long.1062/.