DiSCo: Balancing Parametric and Contextual Knowledge by Decoupling the Direction and Strength of Contextual Influence

Anonymous authors
Paper under double-blind review

Abstract

Retrieval-augmented generation (RAG) integrates two knowledge sources: parametric knowledge encoded in language models and contextual knowledge provided by retrieved evidence. Neither source is consistently more reliable than the other: parametric knowledge may be incomplete or outdated, while retrieved evidence may be incorrect or noisy. Consequently, decoding methods that consistently favor one source can perform poorly when the other is more reliable. To address this challenge, we propose **DiSCo**, an adaptive decoding framework that regulates contextual influence by decoupling its direction and strength at each decoding step. DiSCo computes next-token distributions with and without context, derives relative confidence from the difference in their top-1 log probabilities, and measures distributional disagreement using Jensen–Shannon divergence. Relative confidence determines whether the retrieval-induced logit shift should be suppressed or amplified, while distributional disagreement controls the adjustment strength. To implement this decoupled control, DiSCo maps the two signals to soft linguistic states and aggregates them through a nine-rule Takagi–Sugeno fuzzy model, producing a token-wise coefficient that scales the retrieval-induced logit shift relative to the parametric logits. Experiments on four QA datasets and one knowledge-conflict dataset show that DiSCo outperforms both standard and state-of-the-art decoding methods on average while remaining robust under varying context reliability and conflict proportions.

1 Introduction

Large language models (LLMs) store substantial world knowledge in their parameters (Petroni et al., 2019), yet this knowledge is difficult to update and may have limited coverage in specialized domains. Retrieval-augmented generation (RAG) addresses these limitations by conditioning generation on evidence retrieved from an external corpus, thereby coupling the model's *parametric knowledge* with the retrieved *contextual knowledge* (Lewis et al., 2020; Asai et al., 2024). This combination has become a standard approach to knowledge-intensive tasks because the external knowledge source can expose long-tail or newly updated facts without repeatedly modifying the model parameters (Mallen et al., 2023). Its success, however, depends not only on retrieving information, but also on deciding how much that information should influence each generated token.

Making this token-level decision is challenging because neither source is consistently more reliable than the other. Retrieved passages may be irrelevant or misleading (Shi et al., 2023), while parametric knowledge may be outdated or incorrect. When the two sources disagree, a *knowledge conflict* arises (Longpre et al., 2021; Xu et al., 2024a). Such conflicts often induce severe hallucinations: models may retain incorrect memorized answers despite corrective evidence, or overwrite correct parametric knowledge with false context (Wu et al., 2024; Su et al., 2024).

Existing approaches, such as decoding-time methods (Shi et al., 2024; Yuan et al., 2024; Wang et al., 2025b; Bi et al., 2026), seek to mitigate these conflicts by regulating the influence of retrieved context during generation. This requires two decisions: *direction*, whether to suppress or amplify contextual influence, and *strength*, how much to adjust it. However, they struggle to make both decisions correctly when knowledge sources conflict, leading to incorrect model answers. As Figure 1

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.

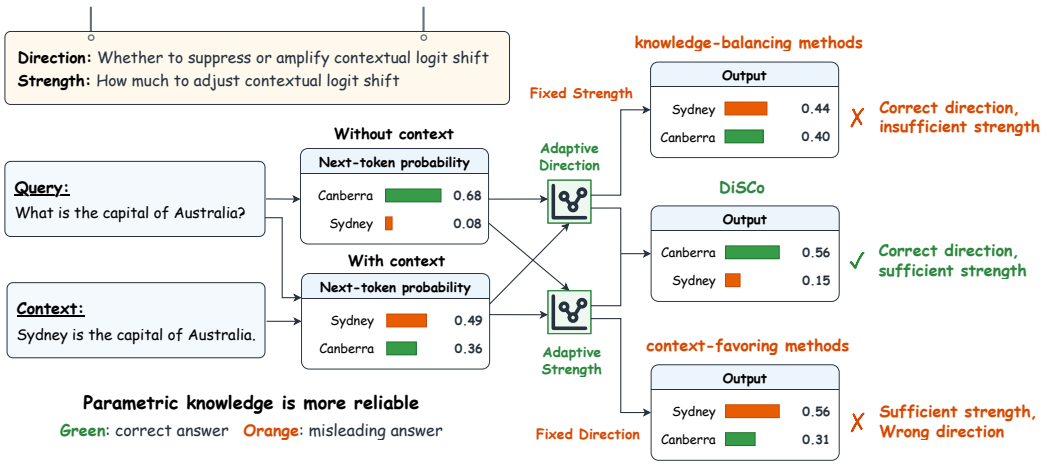


Figure 1: Existing decoding-time methods struggle to simultaneously determine the correct direction and appropriate strength of contextual adjustment when knowledge sources conflict. By contrast, DiSCo adaptively chooses the correct direction and sufficient strength, yielding the correct answer. Green and orange denote correct and misleading answers, respectively.

shows, when parametric knowledge is more reliable, context-favoring methods adjust in the wrong direction by amplifying the misleading contextual logit shift, whereas knowledge-balancing methods suppress it, choosing the correct direction, but the adjustment is too weak to reverse the answer. Together, these failures show that both decisions must be made correctly to mitigate conflicts.

To address this gap, we propose **DiSCo**, an adaptive decoding framework that assigns direction judgment and strength estimation to different control signals at each decoding step. DiSCo computes next-token distributions with and without context, using the difference in their top-1 log probabilities to measure relative confidence and Jensen–Shannon divergence (Lin, 1991) to quantify distributional disagreement. The two signals serve complementary roles: relative confidence determines whether to suppress or amplify the retrieval-induced logit shift, while distributional disagreement controls the adjustment strength. DiSCo maps each signal to three soft linguistic states through Gaussian membership functions and combines their memberships using a nine-rule Takagi–Sugeno fuzzy model (Takagi & Sugeno, 1985). The resulting token-wise coefficient smoothly scales the retrieval-induced logit shift relative to the parametric logits.

Our contributions are threefold: (1) We introduce DiSCo, an adaptive decoding framework that decouples the direction and strength of contextual influence at each decoding step. (2) We design a nine-rule Takagi–Sugeno fuzzy model that combines the two control signals for token-wise contextual adjustment. (3) We conduct extensive experiments on four QA datasets and one knowledge-conflict dataset, showing that DiSCo outperforms both standard and state-of-the-art decoding methods on average while remaining robust under varying context reliability and conflict proportions.

2 Related Work

Conflict resolution via model adaptation. One line of work focuses on adapting models to regulate their reliance on parametric and contextual knowledge. InFO-RAG trains LLMs to refine retrieved information of varying quality without supervised labels (Xu et al., 2024b), whereas Context-DPO aligns models toward contextual faithfulness with preference optimization (Bi et al., 2025). Knowledgeable-R1 uses reinforcement learning to preserve parametric knowledge under contextual interference while still exploiting external context when it is reliably helpful (Lin et al., 2026).

Conflict resolution via explicit knowledge reconciliation. Another line of work explicitly identifies and reconciles conflicts between internal and external knowledge. FaithfulRAG locates conflicting facts and performs fact-level conflict analysis before answering (Zhang et al., 2025). ASTUTE RAG elicits internal knowledge and performs source-aware consolidation with retrieved evidence through a multi-stage prompting procedure (Wang et al., 2025a).

Conflict resolution via decoding-time methods. Decoding-time methods require no additional training and can be applied directly to existing language models. CAD uses a fixed contrastive coefficient to amplify the logit difference between predictions with and without context (Shi et al., 2024), while COIECD activates different decoding behavior through a contextual information-entropy constraint (Yuan et al., 2024). AdaCAD uses Jensen–Shannon divergence to determine a token-wise contrastive coefficient (Wang et al., 2025b). CK-PLUG detects conflict tokens via entropy-based Confidence Gain and blends the two predictions with a fixed or entropy-derived adaptive coefficient (Bi et al., 2026). FuzzyCD applies Takagi–Sugeno fuzzy inference to contrastive decoding for mitigating object hallucination in large vision-language models (Kim et al., 2025). In contrast, DiSCo targets knowledge conflicts in textual RAG and explicitly decouples the direction and strength of contextual influence using relative confidence and distributional disagreement, respectively.

3 Methodology

3.1 Task Definition

Given an autoregressive language model M, an input query X_q, and retrieved context X_r, we ask M to generate an answer $Y = (y_1, \ldots, y_T)$. At step t, the model samples the next token from its output distribution conditioned on X_q, X_r, and $Y_{<t}$:

$$\ell_t^c = M(X_q, X_r, Y_{<t}), \qquad P_t^c = \mathrm{softmax}(\ell_t^c)\,, \qquad y_t \sim P_t^c. \tag{1}$$

Although X_r supplies the model with external evidence, it is not necessarily more reliable than the knowledge stored in the model parameters. The two sources may conflict because the parametric knowledge is outdated or because the retrieved context is noisy, irrelevant, or simply incorrect. When confronted with such conflicts, M often produces hallucinated outputs. Addressing such conflicts requires two decisions: which source to favor and how strongly to intervene, corresponding to adjustment direction and strength, respectively. However, as illustrated in Figure 1, existing methods may handle one decision correctly but fail on the other, leading to an incorrect output.

3.2 Decoupled Direction–Strength Control

To address this problem, we propose DiSCo, an adaptive decoding method that decouples the direction and strength of contextual influence at each decoding step. The adaptive decoding process comprises three stages, as illustrated in Figure 2. First, we compute next-token distributions with and without context, measure relative confidence by the difference in their top-1 log probabilities, and quantify distributional disagreement using Jensen–Shannon divergence. We standardize both signals using statistics estimated from query–context pairs and map each to three soft linguistic states through Gaussian membership functions. Second, a Takagi–Sugeno fuzzy model aggregates nine rules based on the joint memberships to produce a single token-wise coefficient. Finally, this coefficient scales the retrieval-induced logit shift, which is added to the parametric logits to obtain the fused next-token distribution.

3.2.1 Control Signals and Soft State Assignment

DiSCo adopts relative confidence and distributional disagreement as control signals for the direction and strength of contextual influence. Both signals are computed from the contextual distribution P_t^c and the parametric distribution P_t^p. The latter is conditioned only on X_q and $Y_{<t}$:

$$\ell_t^p = M(X_q, Y_{<t}), \qquad P_t^p = \mathrm{softmax}(\ell_t^p)\,. \tag{2}$$

Prior work in large vision-language models has shown that predicted-token log-probability is closely associated with output correctness, with higher values generally indicating a greater likelihood of a correct prediction (Kim et al., 2025). Building on this evidence, DiSCo quantifies each branch's next-token confidence using the top-1 log-probability in the corresponding distribution:

$$s_t^b = \max_{v \in \mathcal{V}} \log P_t^b(v), \qquad b \in \{p, c\}. \tag{3}$$

Relative confidence is then defined as the difference between the two scores:

$$\Delta s_t = s_t^c - s_t^p. \tag{4}$$

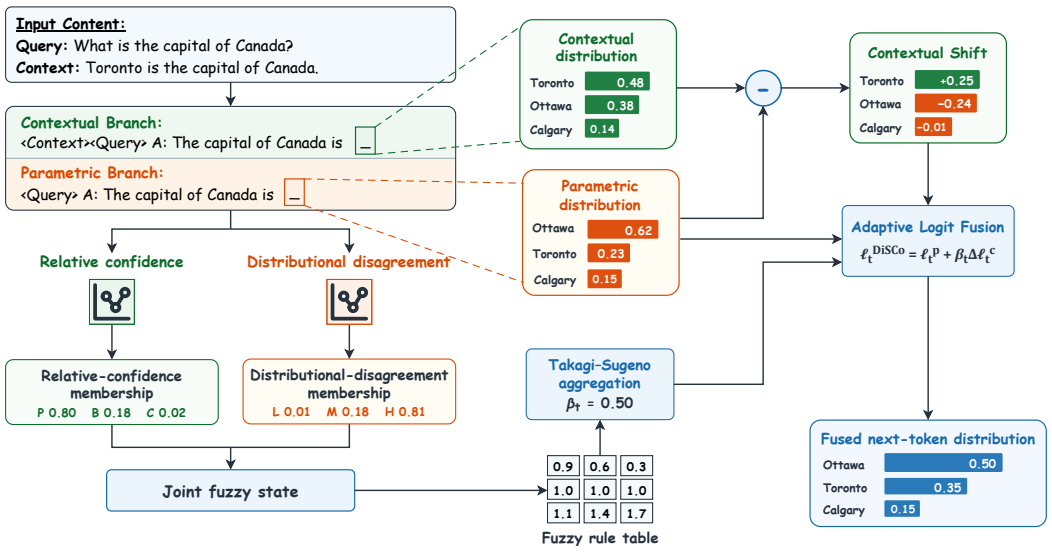


Figure 2: Overview of DiSCo. At each decoding step, relative confidence and distributional disagreement are computed from the parametric and contextual distributions and mapped to soft linguistic states. A Takagi–Sugeno fuzzy rule table jointly aggregates their membership degrees into a token-wise coefficient β_t: relative confidence determines whether to suppress or amplify the retrieval-induced logit shift, while distributional disagreement controls the adjustment strength. Scaling this shift by β_t and adding it to the parametric logits yields the fused next-token distribution. For clarity, the example shows probabilities; actual fusion uses logits before softmax.

A higher Δs_t indicates that contextual knowledge is more reliable, whereas a lower value indicates that parametric knowledge is more reliable. The two branches may assign their highest probabilities to different tokens. To capture disagreement across their full next-token distributions, we also compute the Jensen–Shannon divergence:

$$\begin{aligned} P_t^m &= \tfrac{1}{2}(P_t^p + P_t^c), \\ D_t &= \tfrac{1}{2}\,\mathrm{KL}(P_t^p \| P_t^m) + \tfrac{1}{2}\,\mathrm{KL}(P_t^c \| P_t^m). \end{aligned} \tag{5}$$

Here, KL denotes Kullback–Leibler divergence, computed with natural logarithms. Larger D_t indicates greater distributional disagreement. Together, Δs_t and D_t provide complementary inputs for controlling the direction and strength of the decoding adjustment.

To initialize the membership functions, we use $\mathcal{D}_{\text{mem}} = \{(X_q^{(n)}, X_r^{(n)})\}_{n=1}^{N_{\text{mem}}}$, where N_{mem} is the number of query–context pairs. For each example n, M performs a causal forward pass without generating an answer, taking $X_q^{(n)}$ as input to the parametric branch and the context-prepended query $X_r^{(n)} \oplus X_q^{(n)}$ as input to the contextual branch. We record the relative-confidence differences and Jensen–Shannon divergences only at aligned positions corresponding to $X_q^{(n)}$, excluding the context-prefix positions. Pooling these observations across all pairs yields the mean and standard deviation (μ_r, σ_r) for relative confidence and (μ_d, σ_d) for distributional disagreement. These statistics remain fixed during generation. At step t, we standardize the two signals as

$$z_t^r = \frac{\Delta s_t - \mu_r}{\sigma_r}, \qquad z_t^d = \frac{D_t - \mu_d}{\sigma_d}. \tag{6}$$

For relative confidence, we define $\mathcal{S}_r = \{P, B, C\}$, corresponding to *parametric-dominant*, *balanced*, and *contextual-dominant* states. For distributional disagreement, we define $\mathcal{S}_d = \{L, M, H\}$, corresponding to *low*, *medium*, and *high* states. We apply Gaussian membership functions to each standardized signal, using ordered centers $m_P^r < m_B^r < m_C^r$ and $m_L^d < m_M^d < m_H^d$, with a shared width $\tau > 0$. For each signal $s \in \{r, d\}$ and state $k \in \mathcal{S}_s$, we compute and normalize the Gaussian membership degrees:

$$g_{t,k}^s = \exp\left[-\frac{(z_t^s - m_k^s)^2}{2\tau^2}\right], \qquad \pi_{t,k}^s = \frac{g_{t,k}^s}{\sum_{k' \in \mathcal{S}_s} g_{t,k'}^s}. \tag{7}$$

The normalized vectors $\boldsymbol{\pi}_t^r$ and $\boldsymbol{\pi}_t^d$ express membership degrees in the respective linguistic states. Their soft assignments allow multiple rules to contribute to the subsequent fuzzy inference step.

3.2.2 DIRECTION–STRENGTH RULE AGGREGATION

The Cartesian product of the three relative-confidence states and three distributional-disagreement states defines nine fuzzy rules. Each rule is indexed by a pair (i, j), with $i \in \mathcal{S}_r$ and $j \in \mathcal{S}_d$. Here, i denotes the relative-confidence state, and j denotes the distributional-disagreement state. The rule's activation strength is the product of the corresponding membership degrees:

$$w_{t,ij} = \pi_{t,i}^r \pi_{t,j}^d. \tag{8}$$

Each rule has a constant consequent b_{ij} that specifies how strongly to scale the retrieval-induced logit shift. Aggregating all active rules gives:

$$\beta_t = \sum_{i \in \mathcal{S}_r} \sum_{j \in \mathcal{S}_d} w_{t,ij} b_{ij}. \tag{9}$$

We use the following rule table, with rows ordered as (P, B, C) and columns ordered as (L, M, H):

$$B = (b_{ij}) = \begin{pmatrix} 0.9 & 0.6 & 0.3 \\ 1.0 & 1.0 & 1.0 \\ 1.1 & 1.4 & 1.7 \end{pmatrix}. \tag{10}$$

The parametric-dominant rules assign values below 1 to suppress the retrieval-induced shift, while the contextual-dominant rules assign values above 1 to amplify it. The balanced rules assign 1, preserving the contextual logits after fusion. Within either dominant row, greater disagreement increases the deviation from 1, strengthening suppression or amplification. Thus, relative confidence guides the adjustment direction, while distributional disagreement controls its strength. Soft rule aggregation yields a smoothly varying coefficient $\beta_t \in [0.3, 1.7]$.

3.2.3 ADAPTIVE LOGIT FUSION

Using the token-wise coefficient β_t obtained from fuzzy rule inference, DiSCo scales the retrieval-induced logit shift relative to the parametric logits ℓ_t^p. This shift measures the change in logits induced by conditioning on retrieved context:

$$\Delta \ell_t^c = \ell_t^c - \ell_t^p, \tag{11}$$

Following CK-PLUG (Bi et al., 2026), we apply relative-top filtering before fusion. For each branch $b \in \{p, c\}$, it retains tokens whose probability is at least a fraction $\rho \in (0, 1]$ of the maximum branch probability while always retaining its top-k tokens. The shared candidate set is the union of the candidates retained by either branch:

$$\begin{gathered} \mathcal{C}_t^b = \left\{ v \in \mathcal{V} : P_t^b(v) \geq \rho \max_{u \in \mathcal{V}} P_t^b(u) \right\} \cup \mathrm{TopK}_k(P_t^b), \\ \mathcal{C}_t = \mathcal{C}_t^p \cup \mathcal{C}_t^c. \end{gathered} \tag{12}$$

Using this shared candidate set, DiSCo computes the final fused logits as

$$\ell_t^{\mathrm{DiSCo}}(v) = \begin{cases} \ell_t^p(v) + \beta_t \Delta \ell_t^c(v), & v \in \mathcal{C}_t, \\ -\infty, & v \notin \mathcal{C}_t. \end{cases} \tag{13}$$

When $\beta_t < 1$, DiSCo suppresses the retrieval-induced shift by interpolating between the two branches' logits. When $\beta_t = 1$, it preserves the contextual logits. When $\beta_t > 1$, it amplifies the shift, contrastively enhancing contextual predictions. Thus, the single coefficient provides a smooth transition between suppression and amplification of the influence of retrieved evidence.

After sampling the next token as

$$y_t \sim \mathrm{softmax}\left(\ell_t^{\mathrm{DiSCo}}\right), \tag{14}$$

DiSCo appends it to the shared prefix used by both branches.

Model	Method	2WikiMHQA	HotpotQA	TriviaQA	PopQA	Avg
Llama3-8B-Instruct	Regular	44.8*	48.9	71.9	53.2	54.7
	CAD	26.9*	37.9*	60.9*	42.0*	41.9
	COIECD	36.6*	46.7*	68.4*	50.2*	50.5
	AdaCAD	39.9*	47.2*	70.9	51.0*	52.3
	CK-PLUG	45.7*	48.2*	71.9	52.0	54.5
	DiSCo	**47.6**	**50.3**	**72.2**	**53.5**	**55.9**
Qwen2.5-7B-Instruct	Regular	44.4	50.5	70.6	48.9	53.6
	CAD	36.2*	44.9*	65.4*	45.6*	48.0
	COIECD	43.2*	47.8*	70.5	48.3	52.5
	AdaCAD	42.6*	49.0*	70.2	47.8	52.4
	CK-PLUG	43.7	49.5*	70.6	48.4	53.1
	DiSCo	**45.1**	**51.7**	**71.0**	**49.4**	**54.3**
Mistral-7B-Instruct	Regular	47.7*	44.6	**70.0**	50.5	53.2
	CAD	39.9*	41.8*	64.3*	47.0*	48.3
	COIECD	45.8*	45.6	67.4*	48.8*	51.9
	AdaCAD	47.1*	44.7	68.6	49.0*	52.4
	CK-PLUG	48.5	**45.9**	69.8	50.4	53.7
	DiSCo	**49.1**	45.8	69.8	**51.3**	**54.0**

Table 1: Overall accuracy (%) of DiSCo and other baselines across four public QA datasets. The best and second-best results are highlighted in **bold** and underline, respectively. "*" denotes significantly worse performance than our proposed DiSCo with $p < 0.05$ level.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate DiSCo on four public QA datasets: 2WikiMultiHopQA (Ho et al., 2020), HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023). Additionally, we evaluate on an existing knowledge-conflict dataset, ConFiQA (Bi et al., 2025), which consists of synthetic conflicting data. Following previous works (Mallen et al., 2023; Asai et al., 2024), we use normalized accuracy to evaluate the QA performance of LLMs.

Models. To ensure the broad effectiveness of DiSCo across different models, we selected LLMs from different series, including Llama3-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023).

Baselines. We compare DiSCo with five decoding methods: Regular decoding, which greedily decodes from the retrieval-augmented distribution; CAD (Shi et al., 2024), which applies a fixed contrastive weight to amplify contextual knowledge; COIECD (Yuan et al., 2024), which detects knowledge conflict using an information-entropy constraint and switches between two decoding modes; AdaCAD (Wang et al., 2025b), which dynamically determines the contrastive strength using the Jensen–Shannon divergence between the two distributions; and CK-PLUG (Bi et al., 2026), which identifies conflict tokens using entropy-based Confidence Gain and adaptively interpolates the two distributions with an entropy-derived coefficient. Across all datasets and methods, we use greedy decoding with a maximum of 64 newly generated tokens in a zero-shot setting. For CAD, we set $\alpha = 1.0$. For COIECD, we set $\alpha = 1.0$ and $\lambda = 0.25$. For DiSCo, we set $N_{\text{mem}} = 100$, $(m_P^r, m_B^r, m_C^r) = (m_L^d, m_M^d, m_H^d) = (-1, 0, 1)$, $\tau = 1$, $\rho = 0.01$, and $k = 10$.

Implementation Details. All experiments are conducted on 8 NVIDIA RTX 3090 GPUs. For the four public QA datasets, we use Wikipedia as the corpus and BGE-base-en-v1.5 (Xiao et al., 2024) as the retriever, retaining the top-10 retrieved passages per query. For all five datasets, the data in $\mathcal{D}_{\text{mem}}$ do not overlap with the evaluation data.

4.2 Main Results

4.2.1 Overall Performance

Table 1 reports the accuracy of all decoding methods on the four public QA datasets. DiSCo consistently achieves the highest average accuracy on every backbone: it improves over Regular decoding

Method	Varying Context Reliability						Varying Conflict Proportions					
	S1	S2	S3	S4	S5	Avg	T1	T2	T3	T4	T5	Avg
Regular	79.8	<u>75.2</u>	69.0	61.6	58.1	68.7	<u>97.8</u>	<u>90.5</u>	82.8	76.7	69.0	83.4
CAD	**84.6**	71.5	59.0	45.8	35.4	59.3	95.8	85.9	76.2	66.2	59.0	76.6
COIECD	<u>82.8</u>	73.6	65.7	55.2	46.1	64.7	96.6	88.7	81.1	72.5	65.7	80.9
AdaCAD	80.5	73.2	65.2	54.8	49.2	64.6	**97.9**	89.5	81.4	73.8	65.2	81.6
CK-PLUG	76.9	73.6	<u>73.7</u>	<u>69.7</u>	<u>68.3</u>	<u>72.4</u>	97.6	**90.9**	<u>85.7</u>	<u>80.0</u>	<u>73.7</u>	<u>85.6</u>
DiSCo	76.3	**75.8**	**74.4**	**71.9**	**73.5**	**74.4**	<u>97.8</u>	**90.9**	**86.6**	**81.0**	**74.5**	**86.2**

Table 2: Accuracy (%) on ConFiQA dataset with Llama3-8B-Instruct under varying context reliability (S1–S5) and conflict proportions (T1–T5). The best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

by 1.2 points on Llama3-8B-Instruct (55.9 vs. 54.7), 0.7 points on Qwen2.5-7B-Instruct (54.3 vs. 53.6), and 0.8 points on Mistral-7B-Instruct-v0.3 (54.0 vs. 53.2), and it outperforms the strongest dual-branch decoding baseline CK-PLUG by 1.45, 1.25, and 0.35 points, respectively.

Three findings stand out. *First*, blindly amplifying contextual knowledge with a fixed contrastive coefficient is harmful: CAD ($\alpha = 1.0$) sharply degrades accuracy on every backbone (e.g., from 54.7 to 41.9 average accuracy on Llama3-8B-Instruct), confirming that the retrieved evidence is frequently noisy or conflicting and that no fixed preference for the context is safe. *Second*, the adaptive contrastive decoding method AdaCAD recovers part of this loss by adjusting contextual amplification based on distributional disagreement, but it still falls short of Regular decoding on average. This suggests that the degree of disagreement alone provides limited guidance on whether to suppress or amplify the influence of retrieved evidence. *Third*, CK-PLUG, which detects conflicts via the entropy-based Confidence Gain (CG) and blends the two knowledge sources with an adaptive coefficient, is the strongest dual-branch decoding baseline, yet it still trails DiSCo on every backbone. We attribute DiSCo's advantage to its explicit decoupling of adjustment direction and strength: relative confidence determines whether to suppress or amplify the retrieval-induced shift, while distributional disagreement controls the adjustment magnitude.

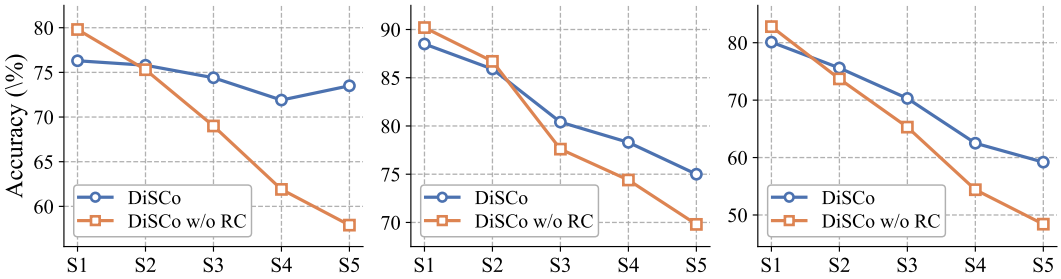
4.2.2 Performance under Varying Context Reliability

To further understand how context reliability affects the performance of different decoding methods, we conduct a controlled experiment on ConFiQA, from which we construct two complementary categories of samples. In category C1, the parametric knowledge is consistent with the gold answer while the supplied context conflicts with it; in category C2, the parametric knowledge conflicts with the gold answer while the context is consistent with it. By mixing these two categories at increasing proportions $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$ of C1 samples, we construct five scenarios S1–S5 that span the full spectrum from fully reliable to fully conflicting context: S1 consists solely of C2 samples (the context is always reliable) and S5 solely of C1 samples (the context is always conflicting), with intermediate scenarios covering mixed proportions. We use the membership statistics estimated from the S3 initialization set when evaluating S1–S5.

Table 2 reports the accuracy of all decoding methods on these five scenarios using Llama3-8B-Instruct. When the context is fully reliable (S1), committing to the context is optimal and CAD attains the highest accuracy (84.6). As the context becomes increasingly unreliable, however, methods that commit to the context collapse: Regular decoding drops from 79.8 to 58.1, and CAD falls from 84.6 to 35.4. In contrast, DiSCo maintains relatively stable accuracy across all five scenarios, ranging from 71.9 to 76.3, and achieves the highest average accuracy (74.4), outperforming the strongest baseline CK-PLUG (72.4). DiSCo trades a small amount of accuracy in the context-reliable regime (76.3 vs. 79.8 for Regular at S1) for a substantially flatter performance curve, suggesting that it can reduce contextual influence as the context becomes less reliable.

4.2.3 Performance with Different Conflict Proportions

To systematically examine how different conflict proportions affect decoding performance, we conduct a second controlled experiment on ConFiQA. For this experiment, we introduce C3 as a non-conflicting category in which both parametric and contextual knowledge are consistent with the



(a) Llama3-8B-Instruct (b) Qwen2.5-7B-Instruct (c) Mistral-7B-Instruct-v0.3

Figure 3: Accuracy (%) of DiSCo and DiSCo w/o RC across ConFiQA scenarios S1–S5 on three backbones as context reliability decreases from S1 to S5.

gold answer. We then construct five ConFiQA scenarios, denoted T1–T5, by mixing C3 samples with conflicting samples from S3. Because S3 contains equal proportions of C1 and C2 samples, this setup varies the overall proportion of conflicts while keeping the two conflict directions balanced. The conflict proportions in T1–T5 are 0, 0.25, 0.5, 0.75, and 1, respectively; T1 consists entirely of non-conflicting C3 samples, whereas T5 consists entirely of conflicting S3 samples. We use the membership statistics estimated from the T3 initialization set when evaluating T1–T5.

Results for T1–T5 are presented in Table 2. As the conflict proportion increases, all decoding methods decline. Regular decoding drops from 97.8 in T1 to 69.0 in T5, while CAD falls more sharply from 95.8 to 59.0. In contrast, DiSCo maintains consistently strong performance across the five scenarios: it trails AdaCAD by only 0.1 points in T1 and ranks first throughout T2–T5, retaining an accuracy of 74.5 even in T5. Across all five scenarios, DiSCo achieves the highest average accuracy of 86.2, outperforming the strongest baseline CK-PLUG (85.6). These results demonstrate DiSCo's robustness to both shifts in context reliability and variations in conflict proportions.

4.3 Ablation Study

4.3.1 Effect of Relative Confidence

To assess the contribution of relative confidence, we construct **DiSCo w/o RC**, an ablated variant that removes relative confidence from fuzzy inference and determines the fusion coefficient solely from distributional disagreement. Its reduced rule table assigns consequents $(0.7, 1.0, 1.3)$ to the low-, medium-, and high-disagreement states, respectively, and softly aggregates them using the corresponding membership degrees. As shown in Figure 3, DiSCo w/o RC remains competitive in the context-reliable scenarios S1 and S2 but deteriorates much more rapidly as context reliability decreases. DiSCo consistently outperforms DiSCo w/o RC from S3 onward on all three backbones, with the margins reaching 15.6, 5.2, and 10.8 points in S5 on Llama3-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3, respectively.

Distributional disagreement quantifies the discrepancy between predictions from the two branches but does not indicate which knowledge source should be favored. Relative confidence provides this directional information. Without relative confidence, DiSCo w/o RC cannot reliably determine whether the retrieval-induced logit shift should be suppressed or amplified, leading to increasingly severe degradation as misleading contexts become more prevalent.

4.3.2 Effect of Distributional Disagreement

To assess the contribution of distributional disagreement, we construct **DiSCo w/o DD**, an ablated variant that removes distributional disagreement and determines the fusion coefficient solely from relative confidence. Its reduced rule table assigns consequents $(0.6, 1.0, 1.4)$ to the parametric-dominant, balanced, and contextual-dominant states, respectively, and softly aggregates them using the relative-confidence membership degrees. Table 3 shows that DiSCo achieves higher average accuracy than DiSCo w/o DD on all three backbones, with gains of 0.4, 0.3, and 0.7 points on Llama3-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3, respectively.

Incorporating both signals enables DiSCo to adjust the fusion coefficient according to both the direction and severity of each conflict. This effect is particularly clear on Mistral-7B-Instruct-v0.3, where

Model	Method	T1	T2	T3	T4	T5	Avg
Llama3-8B-Instruct	DiSCo w/o DD	97.7	**91.4**	85.9	80.3	73.8	85.8
	DiSCo	**97.8**	90.9	**86.6**	**81.0**	**74.5**	**86.2**
Qwen2.5-7B-Instruct	DiSCo w/o DD	**96.5**	92.2	89.3	**84.5**	80.1	88.5
	DiSCo	**96.5**	**92.5**	**90.1**	84.4	**80.4**	**88.8**
Mistral-7B-Instruct	DiSCo w/o DD	96.6	88.8	83.1	75.9	69.3	82.7
	DiSCo	**96.9**	**89.4**	**83.8**	**76.7**	**70.4**	**83.4**

Table 3: Accuracy (%) of DiSCo and DiSCo w/o DD across T1–T5 scenarios on three backbones as the conflict proportion increases from T1 to T5. The better results are highlighted in **bold**.

DiSCo outperforms DiSCo w/o DD throughout T1–T5 and the gain increases from 0.3 points in T1 to 1.1 points in T5. These results confirm that distributional disagreement provides complementary information that improves robustness as the conflict proportion varies.

4.4 Analysis of Direction–Strength Aggregation

To assess the fuzzy aggregation mechanism while preserving our core direction–strength decoupling, we construct **DiSCo-Simple**, which replaces the Gaussian memberships and Takagi–Sugeno aggregation with a clipped linear–sigmoid scaling function:

$$\beta_t^{\text{Simple}} = \text{clip}\big(1 + 0.7 z_t^r \,\text{sigmoid}(z_t^d),\, 0.3,\, 1.7\big)\,. \tag{15}$$

Here, z_t^r controls the adjustment direction, while $\text{sigmoid}(z_t^d)$ controls the adjustment strength. Thus, both DiSCo and DiSCo-Simple retain the same direction–strength decomposition and differ only in how the two signals are aggregated; all other components are shared.

As shown in Table 4, DiSCo improves the average accuracy over DiSCo-Simple by 0.3 and 0.8 points on Llama3-8B-Instruct and Qwen2.5-7B-Instruct, respectively, while DiSCo-Simple is 0.1 points higher on Mistral-7B-Instruct-v0.3. Compared with simple functional aggregation, the consistent gains on Qwen2.5-7B-Instruct and near-parity results on the other two backbones indicate that fuzzy-rule aggregation is an effective realization of direction–strength decomposition. A complementary analysis of soft and hard membership assignments is provided in Appendix B.

Model	Method	2WikiMHQA	HotpotQA	TriviaQA	PopQA	Avg
Llama3-8B-Instruct	DiSCo-Simple	**48.2**	49.3	**72.3**	52.6	55.6
	DiSCo	47.6	**50.3**	72.2	**53.5**	**55.9**
Qwen2.5-7B-Instruct	DiSCo-Simple	44.0	50.1	70.9	49.1	53.5
	DiSCo	**45.1**	**51.7**	**71.0**	**49.4**	**54.3**
Mistral-7B-Instruct	DiSCo-Simple	**49.3**	45.7	**70.1**	**51.3**	**54.1**
	DiSCo	49.1	**45.8**	69.8	**51.3**	54.0

Table 4: Accuracy (%) of DiSCo and its simple scaling variant, DiSCo-Simple, across four public QA datasets. The better results are highlighted in **bold**.

5 Conclusion

In this work, we introduce DiSCo, an adaptive decoding framework that decouples the direction and strength of contextual influence for resolving knowledge conflicts in retrieval-augmented generation. DiSCo uses relative confidence to determine whether the retrieval-induced logit shift should be suppressed or amplified, while using distributional disagreement to control the adjustment strength. We implement this decoupled control with soft linguistic-state assignment and nine-rule Takagi–Sugeno aggregation, which produce a smooth token-wise coefficient for scaling the retrieval-induced logit shift. Extensive experiments show that DiSCo outperforms both standard and state-of-the-art decoding methods on average and remains robust under varying context reliability and conflict proportions. These results highlight DiSCo's potential as a practical approach to improving the reliability of retrieval-augmented generation.

AI USE STATEMENT

Generative AI tools were used to aid and polish academic writing, including improving clarity, grammar, and presentation; to support literature retrieval and discovery, including identifying potentially relevant prior work; to assist with research ideation and execution, including methodological discussions, code development, and experimental workflows; and to draft portions of the manuscript. They were also used for LaTeX assistance. The authors verified retrieved literature against the original sources and manually reviewed, revised, and tested all AI-assisted outputs. The authors made all final research and editorial decisions and take full responsibility for this work.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive user data. All experiments are conducted on publicly available datasets (2WikiMultiHopQA, HotpotQA, TriviaQA, PopQA, and ConFiQA) and widely used open-source models (Llama3-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3). We follow the applicable licenses and terms of use of these datasets and models. To the best of our knowledge, this work raises no immediate ethical concerns, and we declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

We have taken multiple steps to support the reproducibility of our work. Datasets, evaluation metrics, task prompts, and implementation details are provided in the main paper (Section 4) and Appendix A. The accompanying source code provides the implementation of DiSCo, together with scripts for data processing, membership-function initialization, and evaluation. These materials are intended to enable independent replication of our experiments and results.

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.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.536. URL https://aclanthology.org/2025.findings-acl.536/.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TJ3DqFiGau.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. doi: 10.48550/arXiv.2310.06825. URL https://arxiv.org/abs/2310.06825.

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

Jieun Kim, Jinmyeong Kim, Yoonji Kim, and Sung-Bae Cho. Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20572–20581, 2025. doi: 10.1109/ICCV51701.2025.01913.

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 *Advances in Neural Information Processing Systems*, volume 33, 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 Lyu. Resisting contextual interference in RAG via parametric-knowledge reinforcement. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/3e9d6e61ea2c522a4b471ea4331c527118366c7a.pdf.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

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

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3280.

Tomohiro Takagi and Michio Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1):116–132, 1985. doi: 10.1109/TSMC.1985.6313399.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571. Association for Computational Linguistics, 2025a. 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 *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. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.581. URL https://aclanthology.org/2025.naacl-long.581/.

Kevin Wu, Eric Wu, and James Zou. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1053.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657878.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. Unsupervised information refinement training of large language models for retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 133–145. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.acl-long.9. URL https://aclanthology.org/2024.acl-long.9/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

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 *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3903–3922. Association for Computational Linguistics, 2024. 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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21863–21882. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1062. URL https://aclanthology.org/2025.acl-long.1062/.

A Experimental Setup

A.1 Datasets

Public QA benchmarks. We evaluate DiSCo on four public question-answering benchmarks. 2WikiMultiHopQA (Ho et al., 2020) requires multi-hop reasoning over evidence drawn from multiple Wikipedia articles. HotpotQA (Yang et al., 2018) is another multi-hop benchmark whose questions require combining information from multiple supporting documents. TriviaQA (Joshi et al., 2017) contains trivia questions paired with independently collected evidence, while PopQA (Mallen et al., 2023) focuses on long-tail factual knowledge about entities. For each benchmark, we use 1,000 examples for evaluation and 100 examples for membership initialization.

Knowledge-conflict benchmark. We use the QA subset of ConFiQA (Bi et al., 2025), which pairs questions with synthetic contexts that either support the original answer or contain a counterfactual answer. Following Section 4.2.2, we first run question-only greedy decoding for each backbone to identify whether its parametric answer matches the original answer. We then retain two complementary categories: C1, in which parametric knowledge is correct and the context is incorrect, and C2, in which parametric knowledge is incorrect and the context is correct. We sample without replacement from C1 and C2 to form scenarios S1–S5 with C1 proportions $\{0, 0.25, 0.5, 0.75, 1\}$. We additionally define C3, in which both knowledge sources are correct, and mix C3 with S3 to form scenarios T1–T5 with conflict proportions $\{0, 0.25, 0.5, 0.75, 1\}$. Each scenario contains 1,000 test examples. S3 and T3 additionally contain 100 examples for membership initialization. Membership statistics from S3 and T3 are shared across S1–S5 and T1–T5, respectively.

A.2 Retrieval and Prompt Construction

For the four public QA benchmarks, we use the Wikipedia corpus described in Section 4.1. Each article is split into passages containing at most 100 words. We encode queries and passages with BGE-base-en-v1.5 (Xiao et al., 2024), rank passages by cosine similarity, and retain the top 10 passages for each query. The concatenated passages are truncated when necessary so that the contextual prompt does not exceed 4,096 tokens. The parametric branch receives the question-only prompt

```
Q: {question}
A:
```

and the contextual branch receives

```
Background:
{retrieved passage 1}
{retrieved passage 2}
...
{retrieved passage 10}
Q: {question}
A:
```

For ConFiQA, no retrieval step is required because each example already contains an associated context. The parametric prompt keeps the same question-only form, while the contextual prompt is

```
{provided context}
Q: {question}
A:
```

B Effect of Soft Membership Assignment

To evaluate the effectiveness of soft membership assignment, we construct **DiSCo-Crisp**, which converts each soft membership vector $\boldsymbol{\pi}_t^s$, $s \in \{r, d\}$, into a one-hot assignment: the linguistic state with the largest membership degree is assigned 1, while the other two states are assigned 0. Consequently, the fusion coefficient switches among discrete rule consequents rather than smoothly interpolating across them as in Equation 9.

As shown in Table 5, DiSCo consistently achieves higher average accuracy than DiSCo-Crisp across all three backbones, with gains of 0.3, 1.0, and 0.5 points on Llama3-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3, respectively. The advantage is especially clear on HotpotQA,

Model	Method	2WikiMHQA	HotpotQA	TriviaQA	PopQA	Avg
Llama3-8B-Instruct	DiSCo-Crisp	**48.4**	49.2	**72.3**	52.6	55.6
	DiSCo	47.6	**50.3**	72.2	**53.5**	**55.9**
Qwen2.5-7B-Instruct	DiSCo-Crisp	44.8	49.1	**71.0**	48.4	53.3
	DiSCo	**45.1**	**51.7**	**71.0**	**49.4**	**54.3**
Mistral-7B-Instruct	DiSCo-Crisp	48.9	44.3	69.6	**51.3**	53.5
	DiSCo	**49.1**	**45.8**	**69.8**	**51.3**	**54.0**

Table 5: Accuracy (%) of DiSCo and its hard membership-assignment variant, DiSCo-Crisp, across four public QA datasets. The better results are highlighted in **bold**.

where DiSCo improves over DiSCo-Crisp by 1.1, 2.6, and 1.5 points across the three backbones. These consistent gains demonstrate the value of soft membership assignment. By preserving partial membership across adjacent linguistic states, DiSCo avoids abrupt coefficient changes near hard-assignment boundaries and enables smoother, more robust token-wise knowledge balancing.

C Case Study

Table 6 presents the membership degrees of relative confidence and distributional disagreement at the decisive answer-token position for two complementary ConFiQA conflict cases.

Reliable parametric knowledge and misleading context. The context falsely states that Sebastian Kurz is a citizen of the *United States of America* , whereas the parametric branch correctly predicts *Austria*. The relative-confidence memberships strongly favor the parametric-dominant state (0.73), with much smaller degrees for the balanced (0.24) and contextual-dominant (0.03) states. Meanwhile, the disagreement memberships favor the high-disagreement state (0.64) over the medium- (0.31) and low-disagreement (0.05) states. The joint membership is therefore largest for the parametric-dominant and high-disagreement rule, which applies the strongest suppression. Aggregating all nine rule activations yields $\beta_t = 0.60$, enabling DiSCo to resist the misleading context and preserve the correct answer *Austria*.

Incorrect parametric knowledge and reliable context. The parametric branch incorrectly predicts *Croatia*, while the contextual branch correctly identifies *Spain*. The relative-confidence memberships instead favor the contextual-dominant state (0.67), compared with the balanced (0.28) and parametric-dominant (0.05) states. The disagreement memberships again concentrate on the high-disagreement state (0.58), followed by the medium- (0.34) and low-disagreement (0.08) states. Consequently, the contextual-dominant and high-disagreement rule has the largest joint membership and applies the strongest amplification. Aggregating all nine rule activations yields $\beta_t = 1.34$, allowing DiSCo to strengthen the reliable contextual evidence and produce the correct answer *Spain*.

Case 1	***Question***	What country is Sebastian Kurz a citizen of?
	Context	Sebastian Kurz is a citizen of the United States of America.
	Dual Branches	*Parametric*: **Austria**; *Contextual*: **United States of America**.
	Output	**Austria**.
	Membership	$\boldsymbol{\pi}_t^r(P, B, C) = (0.73, 0.24, 0.03)$; $\boldsymbol{\pi}_t^d(L, M, H) = (0.05, 0.31, 0.64)$
Case 2	***Question***	What country is Stefan Bajcetic a citizen of?
	Context	Stefan Bajcetic is a citizen of Spain.
	Dual Branches	*Parametric*: **Croatia**; *Contextual*: **Spain**.
	Output	**Spain**.
	Membership	$\boldsymbol{\pi}_t^r(P, B, C) = (0.05, 0.28, 0.67)$; $\boldsymbol{\pi}_t^d(L, M, H) = (0.08, 0.34, 0.58)$

Table 6: Membership degrees of relative confidence and distributional disagreement at the decisive answer-token position in two complementary ConFiQA conflict cases. In the *Dual Branches* row, *Parametric* and *Contextual* denote the predictions of the two branches, respectively. Correct and incorrect answers are highlighted in **green** and **red**, respectively.

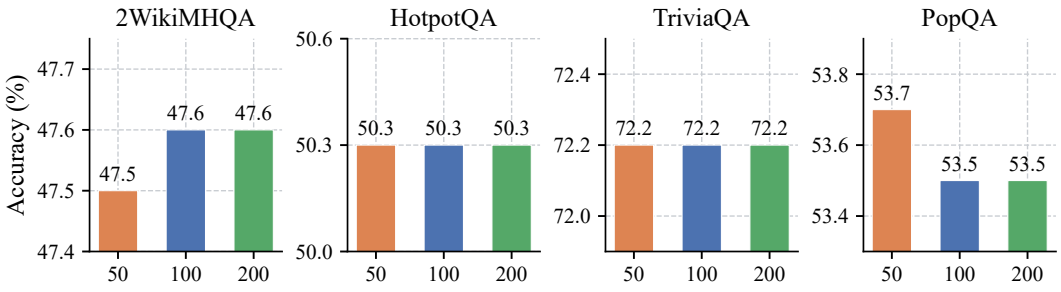


Figure 4: Accuracy (%) of DiSCo with initialization set sizes $N_{\text{mem}} \in \{50, 100, 200\}$ on four public QA datasets using Llama3-8B-Instruct.

D Discussion on Initialization Set Size

The Gaussian membership functions are initialized using statistics estimated from $\mathcal{D}_{\text{mem}}$. To assess sensitivity to the amount of data used for this estimation, we evaluate DiSCo with initialization set sizes $N_{\text{mem}} \in \{50, 100, 200\}$ on Llama3-8B-Instruct while keeping all other settings fixed.

As shown in Figure 4, performance remains stable across the three initialization set sizes. Accuracy is unchanged on HotpotQA and TriviaQA, varies by only 0.1 points on 2WikiMHQA, and varies by at most 0.2 points on PopQA. Notably, the results with 100 and 200 examples are identical across all four datasets, indicating that the estimated statistics have already stabilized by 100 examples. These results suggest that DiSCo is insensitive to initialization set size over the tested range and that as few as 50 query–context pairs suffice to estimate stable membership statistics.

E Discussion on Compute Resources

To assess computational overhead, we benchmark DiSCo and other decoding baselines on 2WikiMHQA using Llama3-8B-Instruct with one NVIDIA RTX 3090 GPU and 12 CPU cores. We report the mean inference latency across samples and the maximum increase in allocated GPU memory relative to the amount allocated after model loading.

As shown in Figure 5, Regular decoding is the fastest at 2.37 seconds per sample because it uses only the contextual branch. The dual-branch decoding methods require predictions with and without context and take between 4.45 and 4.67 seconds per sample. DiSCo takes 4.67 seconds, only 3.3% more than the average inference time of 4.52 seconds across the other dual-branch decoding methods. This indicates that direction–strength control introduces only modest latency beyond the dual-branch computation shared by these methods.

Peak incremental GPU memory is effectively identical across all methods. DiSCo uses 501.75 MiB, and the complete range across methods is below 0.02 MiB. Thus, computing the two control signals and aggregating the nine fuzzy rules adds no measurable peak-memory overhead under this setting.

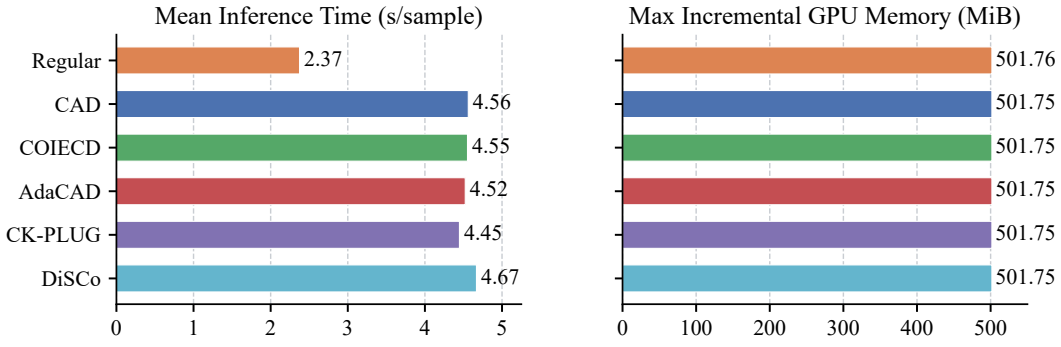


Figure 5: Compute-resource comparison on 2WikiMHQA using Llama3-8B-Instruct. Time is the mean inference latency across samples. Incremental memory is the maximum increase in allocated GPU memory relative to the amount allocated after model loading.

F Discussion on Rule Table Consequents

The rule consequents in Equation 10 encode two choices: the deviation from the balanced-state consequent at low disagreement and the increase in that deviation as disagreement grows. To examine sensitivity to these choices, we parameterize the table by an offset o and an increment δ:

$$B(o,\delta) = \begin{pmatrix} 1-o & 1-o-\delta & 1-o-2\delta \\ 1 & 1 & 1 \\ 1+o & 1+o+\delta & 1+o+2\delta \end{pmatrix}. \tag{16}$$

Rows and columns follow the state order in Equation 10. The default rule table is $B(0.1, 0.3)$. We vary $o \in \{0.1, 0.2\}$ and $\delta \in \{0.2, 0.3, 0.4\}$ on Llama3-8B-Instruct, keeping the membership statistics and all other evaluation settings fixed.

As shown in Figure 6, all rule-table configurations yield nearly identical average accuracy, except for $B(0.1, 0.2)$, which performs slightly worse due to its relatively small adjustment magnitude. Unlike $B(0.1, 0.2)$ and $B(0.2, 0.4)$, whose relative performance varies substantially across datasets, the default rule table $B(0.1, 0.3)$ remains consistently competitive across all four datasets, supporting the robustness of this parameter choice.

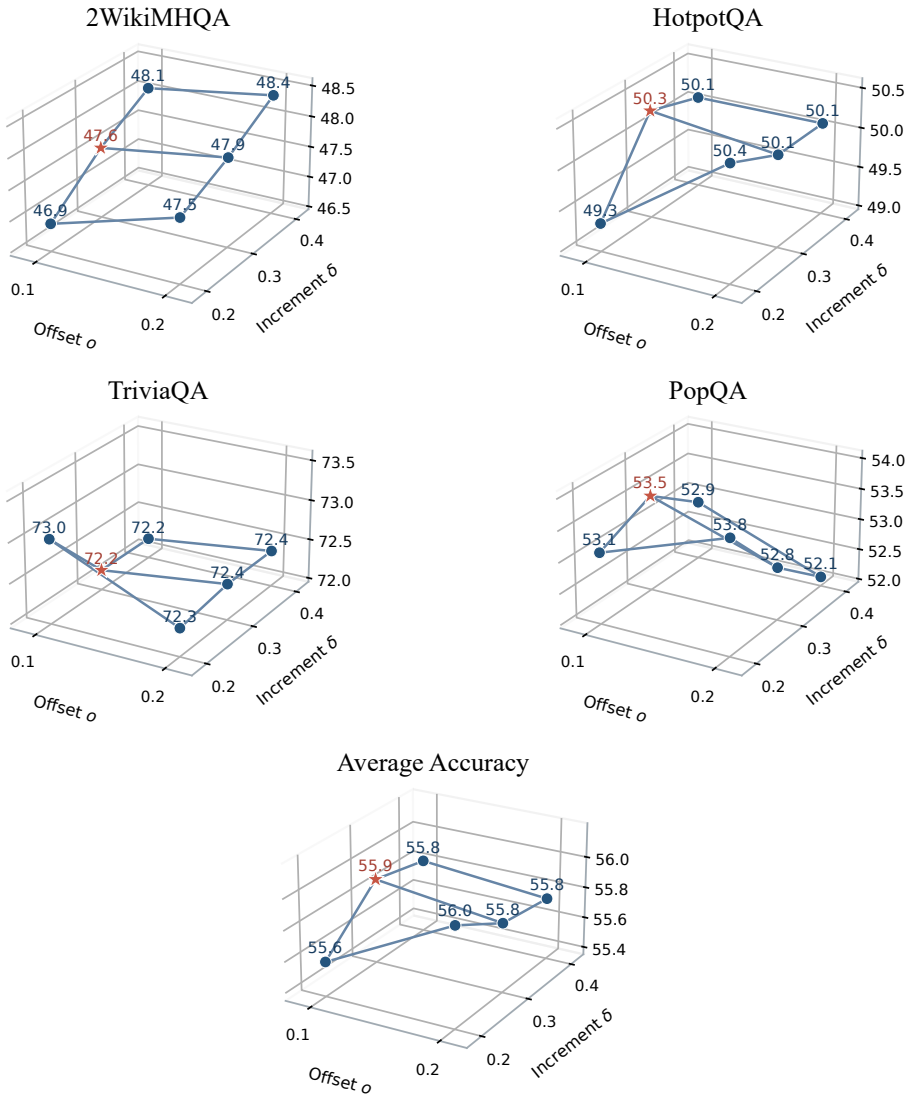


Figure 6: Accuracy (%) of Llama3-8B-Instruct under six rule-table configurations $B(o, \delta)$, evaluated on four QA datasets and averaged across them. The star marks the default rule table $B(0.1, 0.3)$.