CONFIDENT HALLUCINATIONS LIVE IN THE BLIND SPOT OF ENTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time methods that detect or reduce hallucinations through entropy cannot handle confident hallucinations. These wrong answers are as low-entropy as correct ones and persist under resampling. We study where their confidence comes from in context-grounded QA. Context ablation splits an answer's log-probability into a parametric prior and the evidence added by the context. Entropy bounds only their sum and is therefore blind to how the confidence is split. The split reveals two opposite sources of confident hallucination. Context-lure errors copy a wrong span from the context and have abnormally low parametric support. Prior-dominated errors ignore the context and have abnormally high parametric support. We trace both errors to retrieval that either copies a wrong span from the context or finds no matching span. Because benchmarks mix the two types, pooled grounding signals flip sign across datasets and cancel within them. To address this problem, we propose Grounding-Profile Contrastive Decoding (GPCD). GPCD is a label-free test-time method that corrects confident hallucinations without updating any weights. It detects answers that lean too far on one source by scoring them within context-support strata. It re-decodes each flagged answer with contrastive decoding in the direction opposite to its failure. It accepts a change only when resampled candidates agree on an answer that the context supports. On five context-grounded QA benchmarks, the stratified score detects confident hallucinations better than pooled signals on every dataset, and GPCD gives the best average accuracy while global contrastive decoding and self-consistency lose up to 7 EM points on average.

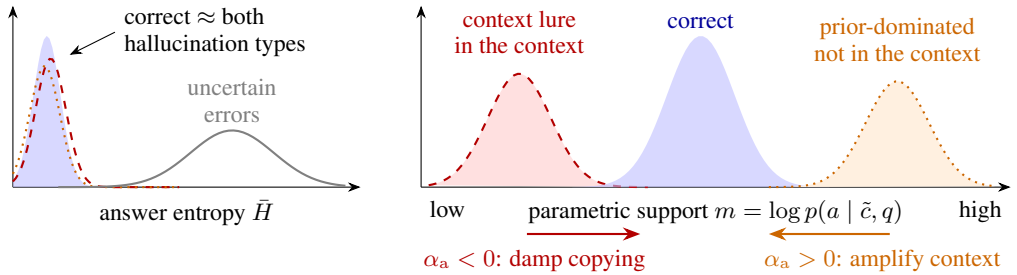
1 INTRODUCTION

Large language models (LLMs) hallucinate even when they are given the evidence they need (Ji et al., 2023; Huang et al., 2025). Most label-free remedies treat hallucination as uncertainty. Detectors flag high token or semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023), and test-time methods minimize entropy or vote (Wang et al., 2021; 2023; Zuo et al., 2025; Agarwal et al., 2025). These methods fail by construction on *confident hallucinations*. Such wrong answers have the same low entropy as correct ones (Liew et al., 2026; Yona et al., 2026). The question is therefore not whether the model is confident, but where its confidence comes from.

To answer this question, we decode each prompt with the context and with the context masked, and take the log-ratio of the two probabilities as a token-level grounding score g. One expects hallucinations to be under-grounded. This holds on HybridQA and MoreHopQA. On 2WikiMultihopQA it reverses, and on MuSiQue the two groups coincide (Figure 3).

We show that the reversal has a simple explanation. Entropy bounds the full-context term of g (Proposition 1). Among confident answers, g therefore mirrors the answer's *parametric support* m. This is the answer's log-probability without the context. A wrong answer can become confident in two opposite ways (Figure 1). A *context-lure* error copies a wrong context span that the model would not produce alone, so m is low. A *prior-dominated* error follows the parametric prior against the context, so m is high. Both failures come from retrieval. Retrieval matches surface form, so it can copy a wrong span or find no matching span and leave the prior to decide. Datasets mix the two types, so the pooled gap flips and cancels. This is a composition effect (Proposition 2). On 2WikiMultihopQA, context-supported hallucinations have $m = -5.1$ nats and unsupported ones have $m = -2.4$. Correct answers lie between them at $m = -2.7$.

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.



(a) What entropy sees **(b) What context ablation sees**

Figure 1: **Two sources of confident hallucination.** (a) Correct answers and confident hallucinations have the same low entropy, so entropy cannot separate them. (b) Parametric support m places lures and prior-dominated errors on opposite sides of correct answers. GPCD scores each side within its own context-support stratum and corrects it in the matching direction.

These findings pose two challenges for correcting confident hallucinations. First, no single direction of g or m identifies errors on every dataset, because the two error types deviate in opposite directions. Second, the two types need opposite corrections. Context-aware decoding (CAD) amplifies the context for every answer. This helps prior errors but makes lures worse.

To address these challenges, we propose Grounding-Profile Contrastive Decoding (GPCD; Figure 2). GPCD is a label-free test-time method and updates no weights. It builds on one observation. By the definition of g, the model's answer score is the sum of a parametric prior m and context evidence g. This split mirrors Bayes' rule, and CAD rescales only its second term. GPCD has three stages. (i) To detect both error types, it scores confident answers with fixed directions separately within two context-support strata. (ii) To correct each type in its own direction, it re-decodes flagged answers with CAD that damps the context for lures and amplifies it for prior errors. (iii) It accepts a change only if resampled candidates agree on an answer that the context supports. Otherwise it keeps the original answer.

Our contributions are as follows.

- We identify a failure mode of grounding-based hallucination detection. Across five context-grounded QA benchmarks, the grounding gap between confident hallucinations and correct answers flips sign, and within a dataset the two error types cancel each other out.
- We explain this failure. We show that entropy cannot see where confidence comes from, and that confident hallucinations arise in two opposite ways. Retrieval either copies a wrong span from the context or finds no matching span and leaves the answer to the parametric prior.
- We propose GPCD, a label-free test-time decoding method that detects both error types and corrects each one in its own direction. Experiments on five benchmarks show that the stratified score detects confident hallucinations more reliably than pooled signals, and that GPCD gives the best average accuracy among decoding methods while global contrastive decoding and self-consistency degrade it. An attention analysis further supports the proposed mechanism.

2 Related Work

Hallucination mitigation. Hallucination is a central obstacle to deploying LLMs, and mitigation follows several directions. Retrieval-augmented generation grounds answers in external evidence (Lewis et al., 2020), but models can still misuse the retrieved context (Shi et al., 2023; Yoran et al., 2024). Training-based methods update the model with supervised or preference data (Tian et al., 2024; Ouyang et al., 2022; Rafailov et al., 2023). Label-free test-time methods treat hallucination as uncertainty. They flag high token or semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023), vote over sampled answers (Wang et al., 2023), or minimize entropy at test time (Wang et al., 2021; Agarwal et al., 2025; Zuo et al., 2025). Such methods cannot see confident hallucinations. Recent work shows that such hallucinations are common (Kalai et al., 2025; Liew et al., 2026; Yona et al., 2026). Existing responses to confident errors need gradients or trained

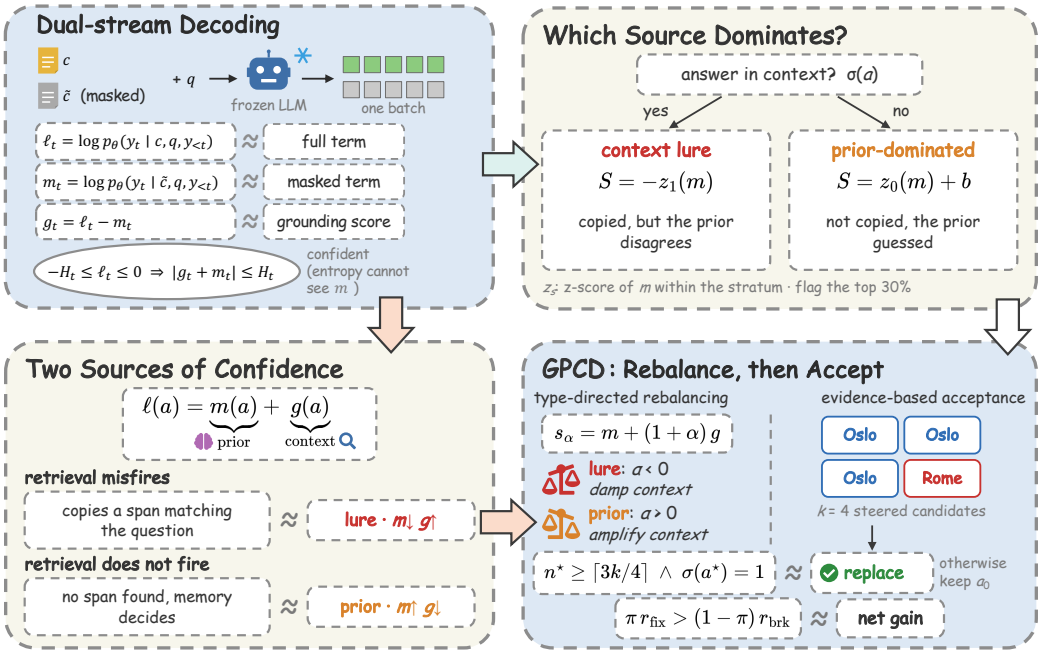


Figure 2: **Overview of GPCD.** *Top left:* dual-stream decoding splits each token's score into m_t and g_t. *Bottom left:* lures come from retrieval that copies a wrong span and prior errors from retrieval that finds no matching span. *Top right:* each stratum is scored in its own direction. *Bottom right:* flagged answers are re-decoded with the context damped or amplified and replaced only on agreement and support.

probes (Liew et al., 2026; Zhang et al., 2026). Our goal is to explain where such confidence comes from when a context is given, and to correct it at test time without labels.

Decoding-time intervention. A line of work improves faithfulness by intervening at decoding time. Contrastive decoding (Li et al., 2023b) contrasts an expert model with a weaker amateur model. DoLa (Chuang et al., 2024b) contrasts the final layer with an early layer. ITI (Li et al., 2023a) shifts activations along truthful directions of selected attention heads. NDAD (Qiu et al., 2026) masks critical heads to elicit a hallucination-leaning distribution and subtracts it. For context-grounded generation, CAD (Shi et al., 2024) amplifies the pointwise mutual information between the context and the output. AdaCAD (Wang et al., 2025) and COIECD (Yuan et al., 2024) set its strength from a measured knowledge conflict (Longpre et al., 2021; Xie et al., 2024). The RAG detectors ReDeEP (Sun et al., 2025) and Lookback Lens (Chuang et al., 2024a) also treat low context use as the failure mode. All of these methods apply one correction direction to every answer. GPCD instead chooses the direction per answer: it damps the context for lures and amplifies it for prior errors. It also changes an answer only when resampled candidates agree on a context-supported alternative. Appendix A extends this discussion.

3 Observing Confident Hallucinations

We analyze greedy traces of Llama-2-7B-Base (Touvron et al., 2023) on HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), HybridQA (Chen et al., 2020), MoreHopQA (Schnitzler et al., 2024), SQuAD 2.0 (Rajpurkar et al., 2018) and FanOutQA (Zhu et al., 2024). HotpotQA, 2WikiMultihopQA and MuSiQue share one context-QA prompt. MoreHopQA, HybridQA, FanOutQA and SQuAD 2.0 keep the prompt of their original release, and Appendix C lists all of them. Here a confident hallucination is a wrong answer with entropy below the correct-group mean. Figure 3 also requires low context-free entropy. Labels are used only for this analysis.

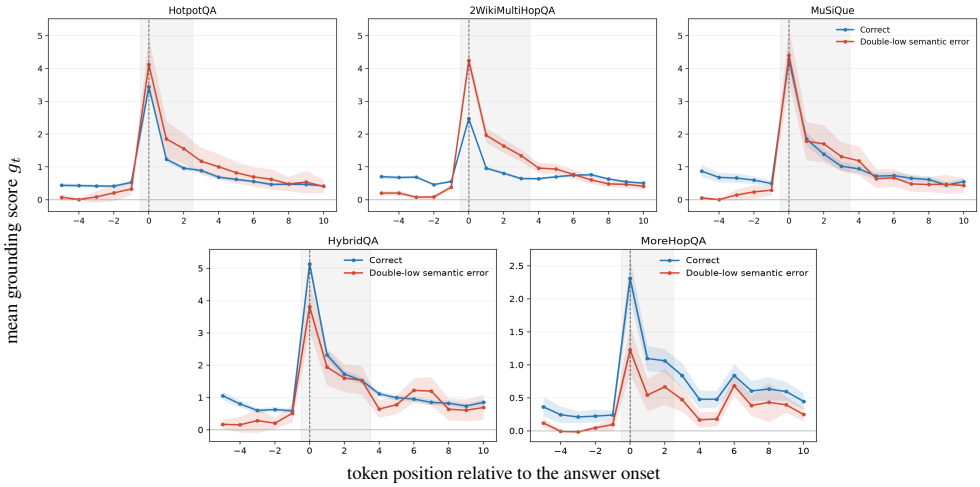


Figure 3: **The grounding gap flips sign across datasets.** Mean token-level grounding score g_t aligned at the answer onset at position 0. Shaded bands show standard errors. Blue curves show correct answers and red curves show confident hallucinations. Hallucinations peak higher on HotpotQA and 2WikiMultihopQA, match correct answers on MuSiQue, and peak lower on HybridQA and MoreHopQA in the bottom row. Red curves are labeled "double-low" errors in the legend: low entropy with and without the context.

3.1 Context ablation isolates the source of confidence

Setup. Given a context c and a question q, a model p_θ generates $y = (y_1, \ldots, y_T)$. Its answer a starts at token t_a. H_t is the entropy of $p_\theta(\cdot \mid c, q, y_{<t})$ and $\bar{H}$ its mean over the answer. An answer is *confident* if $\bar{H}$ is below the test-set median. A *confident hallucination* is a confident answer with exact match 0.

Let $\tilde{c}$ be c with every token replaced by the filler token "`.`", and for each generated token,

$$g_t = \underbrace{\log p_\theta(y_t \mid c, q, y_{<t})}_{\ell_t\text{: full term}} - \underbrace{\log p_\theta(y_t \mid \tilde{c}, q, y_{<t})}_{m_t\text{: masked term}}. \tag{1}$$

g_t is the pointwise mutual information between context and token that CAD amplifies (Shi et al., 2024). The masked term m_t is the token's *parametric support*. We decode both prompts as two rows of one batch, so each token gets ℓ_t, m_t and g_t in one pass. We read g and m at the answer onset and write g^{ans} and m^{ans}.

Proposition 1. *If* $y_t = \arg\max_v p_\theta(v \mid c, q, y_{<t})$ *as in greedy decoding, then*

$$-H_t \;\le\; \ell_t \;\le\; 0, \qquad \textit{hence} \qquad \big|\, g_t + m_t \,\big| \;\le\; H_t. \tag{2}$$

Proof. $-\ell_t$ is the min-entropy of the full-context distribution. The min-entropy never exceeds the Shannon entropy (Cover & Thomas, 2006). The rest follows from $g_t + m_t = \ell_t$. □

For confident answers, the grounding score is the negative parametric support. In Table 1, the full term stays between -0.19 and -0.12 nats in every 2WikiMultihopQA group. Entropy depends only on the full-context distribution, so two answers with equal entropy can have any m_t. Context ablation adds the coordinate that entropy cannot see.

3.2 Finding: two opposite sources flip the grounding gap

Figure 3 shows the mean grounding score aligned at the answer onset. Confident hallucinations peak lower than correct answers on HybridQA and MoreHopQA and higher on 2WikiMultihopQA and HotpotQA. On MuSiQue the means coincide. However, this flip of the group means does not imply that the grounding score is unreliable.

Table 1: **Same dataset, opposite directions.** Grounding profile at the answer onset in nats. Agree is the share of answers whose onset token is the context-free argmax, and supp. is the share supported by the context.

Dataset	Group	n	g	ℓ	m	agree %	supp. %
2WikiMultihopQA	correct	8,699	2.47	-0.19	-2.66	50.2	91.8
	conf. hal., all	971	4.31	-0.14	-4.44	23.6	75.6
	conf. hal., $\sigma = 1$	734	4.99	-0.12	-5.11	13.8	100
	conf. hal., $\sigma = 0$	237	2.20	-0.19	-2.39	54.0	0
MuSiQue	correct	1,113	4.28	-0.12	-4.40	38.8	99.4
	conf. hal., all	217	4.51	-0.24	-4.75	39.2	54.4
	conf. hal., $\sigma = 1$	118	5.81	-0.23	-6.04	23.7	100
	conf. hal., $\sigma = 0$	99	2.96	-0.25	-3.22	57.6	0
HybridQA	correct	2,670	5.1	-0.16	-5.30	27.0	88.5
	conf. hal.	213	3.8	-0.25	-4.08	36.2	60.6
MoreHopQA	correct	504	2.3	-0.21	-2.52	54.0	24.6
	conf. hal.	171	1.4	-0.34	-1.78	53.8	14.0

By Proposition 1, a difference in g between confident groups is a difference in m with the sign reversed. A wrong answer can become confident in two ways. Section 4 explains both mechanistically. A *context lure* copies a wrong context span: m is low, g is high, and the context supports the answer. A *prior-dominated* error follows the prior against the context: m is high, g is low, and the context typically does not support it. We test these opposite predictions within datasets by splitting confident hallucinations by *context support* $\sigma(a) \in \{0, 1\}$ as defined in Eq.(5). Table 1 gives two results.

1. **Both directions in one dataset.** On 2WikiMultihopQA, supported hallucinations have $m = -5.1$ nats, and only 13.8% match the context-free argmax. Unsupported ones have -2.4 nats and a 54.0% match. Correct answers lie between them with $m = -2.7$ and a 50.2% match. MuSiQue shows the same order with $m = -6.0$, -4.4 and -3.2 for supported hallucinations, correct answers and unsupported hallucinations.
2. **Prior-side datasets.** On HybridQA and MoreHopQA, hallucinations have higher m than correct answers and are less often supported. On HybridQA, m is -4.1 against -5.3 and support is 60.6% against 88.5%. On MoreHopQA, m is -1.8 against -2.5 and support is 14.0% against 24.6%. The MoreHopQA shift lies in the tail, so we compare groups in log space throughout.

On MuSiQue the two types are almost balanced, so they cancel in the pooled mean. All confident hallucinations have $g = 4.51$ against 4.28 for correct answers, although the two strata differ by 2.85 nats.

Opposite effects within strata and different stratum shares make the pooled gap uninterpretable. This is related to Simpson's paradox (Simpson, 1951), but here the strata themselves disagree in sign. Write W and C for confident hallucinations and correct confident answers, $\mu_{C,s} = \mathbb{E}[m \mid C, \sigma{=}s]$, and $\Delta_s = \mathbb{E}[m \mid W, \sigma{=}s] - \mu_{C,s}$.

Proposition 2. *The pooled gap decomposes as*

$$\mathbb{E}[m \mid W] - \mathbb{E}[m \mid C] \;=\; \underbrace{\sum_{s\in\{0,1\}} P(s \mid W)\,\Delta_s}_{\textit{mechanism}} \;+\; \underbrace{\sum_{s\in\{0,1\}} \big(P(s \mid W) - P(s \mid C)\big)\,\mu_{C,s}}_{\textit{composition}}. \qquad (3)$$

Proof. Expand $\mathbb{E}[m \mid W] = \sum_s P(s \mid W)(\mu_{C,s} + \Delta_s)$ and $\mathbb{E}[m \mid C] = \sum_s P(s \mid C)\mu_{C,s}$, then subtract. □

Because the shares sum to one, the composition term equals $(P(1 \mid W) - P(1 \mid C))(\mu_{C,1} - \mu_{C,0})$. On 2WikiMultihopQA its first factor is $0.756 - 0.918 = -0.16$. If $\Delta_1 < 0 < \Delta_0$, the mechanism term mixes two opposite effects by the dataset's type shares, and neither term has a fixed sign. A pooled score of fixed direction can therefore work on some benchmarks and fail on others. The decomposition shows that type shares can cause the flip, not that they are its only cause. The stratified statistic uses $-m$ within $\sigma{=}1$ and $+m$ within $\sigma{=}0$. It points toward error in both strata.

Deviation alone does not identify an error. On the prior side, a low g only says that the context adds nothing, which also happens when the model already knows an answer that the context confirms. A pilot that re-decoded every deviating answer lowered EM on HotpotQA from 34.0 to 28.0, as detailed in Appendix D.

4 Understanding Confident Hallucinations

Based on these findings, we further explain how the model produces the two error types and derive our method from this explanation.

How the model answers from a context. Compare candidate answers at the same position and after the same prefix. For a candidate v, the definition of g in Eq.(1) gives

$$\underbrace{\log p_\theta(v \mid c, q)}_{\ell(v)} = \underbrace{\log p_\theta(v \mid \tilde{c}, q)}_{m(v)\text{: parametric prior}} + \underbrace{g(v)}_{\text{context evidence}} . \tag{4}$$

Eq. (4) mirrors Bayes' rule $\log p(v \mid c, q) = \log p(v \mid q) + \log \frac{p(c|v,q)}{p(c|q)}$, with m estimating the parametric prior and g the pointwise evidence of the context. We mask the context instead of removing it so that lengths and positions stay fixed. The two terms are computed by different parts of the model. The prior comes from factual recall in the MLP layers (Geva et al., 2021; Meng et al., 2022). The evidence comes from retrieval heads that locate a span matching the query and copy it to the output (Olsson et al., 2022; Wu et al., 2025). For a confident token $\ell \approx 0$, so Proposition 1 gives $g \approx -m$. The two sources therefore share one fixed budget, and a confident answer can draw it mostly from retrieval, mostly from the prior, or from both. Correct answers already split this budget differently across tasks. They rely on retrieval when the answer is read off a table and on both sources when the model knows the entities. An error is thus marked not by relying on one source but by relying on it more than correct answers of the same task.

This budget view places the weak point in retrieval, which matches the surface form of the query without verifying the full relation. Retrieval therefore fails in two opposite ways. In the first, it copies a span that matches the question's words but answers a different relation, such as the wrong hop of a bridge or a distractor entity (Jia & Liang, 2017; Shi et al., 2023). The copied span receives a large g that outweighs its low prior, and the answer is confident because copying is sharp. This context lure lies in the context with $\sigma = 1$, low m and high g, as the supported errors in Table 1 show.

In the second, retrieval finds no matching span because the answer must be composed across hops, read from a table or computed. Then g is small for every candidate, and the prior decides with a sharp but unverified recall. This prior-dominated error lies outside the context with $\sigma = 0$, high m and low g, as the unsupported errors in Table 1 show.

The two failures move m in opposite directions, and their proportions depend on the task. Distractor-rich multi-hop QA invites wrong spans, while table and computation QA leaves retrieval without a match. By Proposition 2, the pooled gap averages the two in these proportions, which explains the flip in Figure 3. The attention analysis in Section 6.4 further supports this account. For lures, retrieval heads attend more to the wrong span than to the gold span. For prior errors, they attend less to the gold span than for correct answers. Their attention on the answer span also correlates positively with g.

5 Grounding-Profile Contrastive Decoding

Since a confident hallucination relies on one source more than correct answers of the same task do, we propose GPCD, which compares each answer only with answers from the same test set and the same context-support stratum. It scores each stratum in the direction of its error type and corrects flagged answers in the matching direction.

Pass 1: dual-stream decoding. We decode each prompt greedily with the full and the masked context in one batch and record g_t, ℓ_t and H_t as in Eq. (1). From each trace we read m^{ans} and $\bar{H}$. The confident pool $\mathcal{P}$ holds the answers with $\bar{H} \le Q_{0.5}(\bar{H})$.

Context support. An answer is supported if it appears in the context:

$$\sigma(a) = 1 \iff \mathrm{norm}(a) \sqsubseteq \mathrm{norm}(c) \text{ or } \max_{u \in \mathcal{N}(c,|a|)} \mathrm{F1}(a,u) \geq 0.8, \tag{5}$$

where norm is answer normalization, $\sqsubseteq$ is substring containment at word boundaries, and $\mathcal{N}(c,n)$ holds the context n'-grams with $|n'-n| \leq 1$. Yes/no answers count as supported. We call the regime *extractive* if the supported fraction h of $\mathcal{P}$ is at least 0.5.

Stage 1: stratified detection. Support reveals which source produced an answer. Among supported answers a lure shows a low m, and among unsupported answers a prior error shows a high m. In extractive QA an unsupported answer is also suspicious in itself because the gold answer is in the context. Let z_s standardize m^{ans} by its mean and standard deviation over $\{i \in \mathcal{P} : \sigma(a_i) = s\}$. The score is

$$S(x) = \begin{cases} -z_1(m^{\mathrm{ans}}) & \text{if } \sigma(a) = 1 \quad \text{candidate LURE}, \\ z_0(m^{\mathrm{ans}}) + b \cdot \mathbb{1}[h \geq 0.5] & \text{if } \sigma(a) = 0 \quad \text{candidate PRIOR}, \end{cases} \tag{6}$$

where the offset $b = 1$ encodes this suspicion. We flag the top $\rho = 30\%$ of $\mathcal{P}$ and take the stratum of a flagged answer as its type.

Stage 2: type-directed contrastive decoding. CAD reweights the two sources:

$$s_\alpha(v) \;=\; (1+\alpha)\ell(v) - \alpha\, m(v) \;=\; m(v) + (1+\alpha)\, g(v). \tag{7}$$

A wrong answer w beats the gold answer v^* when $m(w) + g(w) > m(v^*) + g(v^*)$. A lure wins through retrieval with $m(w) < m(v^*)$ and hence $g(w) > g(v^*)$, so only $\alpha < 0$ lets the prior overturn it. A prior error has $g(w) \leq g(v^*)$, so only $\alpha > 0$ lets the context overturn it. Each type thus has one correcting direction. Global CAD always amplifies the context and therefore cannot fix lures. We re-decode each flagged prompt with

$$s_t(v) \;=\; (1+\alpha_{\mathrm{a}}) \log p_\theta(v \mid c, q, y_{<t}) \;-\; \alpha_{\mathrm{a}} \log p_\theta(v \mid \tilde{c}, q, y_{<t}), \tag{8}$$

where α_{a} depends on the type with $\alpha_{\mathrm{a}}(\textsc{prior}) = +0.5$ and $\alpha_{\mathrm{a}}(\textsc{lure}) = -0.5$. We draw $k = 4$ candidates, one greedy and three sampled at $T = 0.7$.

Stage 3: evidence-based acceptance. Let π be the precision of the flagged set. A correction helps only if

$$\pi\, r_{\mathrm{fix}} > (1-\pi)\, r_{\mathrm{brk}}, \tag{9}$$

where r_{fix} and r_{brk} are the rates at which flagged errors are fixed and flagged correct answers are broken. The flagged set always contains correct answers, so the acceptance rule must keep r_{brk} low. The detection score cannot guard against it because steering moves the score by construction. The steered argmax a_α and the original a_0 satisfy $s_\alpha(a_\alpha) \geq s_\alpha(a_0)$ and $\ell(a_0) \geq \ell(a_\alpha)$, and adding the two gives $\alpha\, g(a_\alpha) \geq \alpha\, g(a_0)$ for right and wrong answers alike. GPCD therefore accepts a replacement only on evidence that steering does not create. Let $a^\star$ be the most frequent normalized candidate. We replace a_0 with $a^\star$ only if $a^\star \neq a_0$, at least $\lceil 3k/4 \rceil$ candidates agree on $a^\star$, and $\sigma(a^\star) = 1$ in extractive regimes. If each steered sample keeps a flagged correct answer with probability at least q, the agreement rule replaces it with probability at most

$$P\big(\mathrm{Bin}(k-1, 1-q) \geq \lceil 3k/4 \rceil - 1\big), \tag{10}$$

which is small when the steered model usually keeps a correct answer.

6 Experiments

We test one claim per question. **RQ1:** Does GPCD outperform greedy decoding and existing decoding methods? **RQ2:** Does the stratified score detect confident hallucinations better than pooled scores? **RQ3:** Does the attention of the model support the retrieval account?

Table 2: **Main results.** EM and F1 in percent on the full development sets of five context-grounded QA benchmarks. Best per column in **bold**.

		HotpotQA		2Wiki		MuSiQue		HybridQA		MoreHopQA		Avg.	
Model	Method	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Llama-2-7B	Greedy	32.7	43.6	24.3	30.3	7.5	14.1	11.1	15.8	9.2	10.6	17.0	22.9
	CAD (α=1)	20.4	32.6	4.2	13.4	4.9	12.4	**11.2**	15.8	9.2	10.7	10.0	17.0
	CAD (α=0.5)	28.7	40.8	16.5	25.4	7.5	15.5	**11.2**	15.8	9.2	10.7	14.6	21.6
	AdaCAD	32.1	43.5	24.2	30.2	7.5	14.1	10.6	**16.2**	**9.3**	11.1	16.7	23.0
	NDAD	32.8	43.6	24.3	30.3	7.5	14.1	11.1	15.8	**9.3**	10.6	17.0	22.9
	Self-consistency	29.0	38.9	16.6	25.1	2.1	6.9	7.4	12.4	7.4	**12.4**	12.5	19.1
	GPCD (ours)	**32.9**	**43.7**	**24.5**	**31.7**	**7.8**	**15.7**	**11.2**	15.8	**9.3**	11.1	**17.1**	**23.6**

Table 3: **Detection of confident hallucinations within the confident pool.** AUROC left of the bar. Conditional columns score within one support stratum. The last two columns give the hallucination rate among flagged answers at $\rho = 30\%$ and among all confident answers.

Dataset	$+m$	$-m$	unsupported	$-m \mid \sigma$=1	$+m \mid \sigma$=0	**Stratified**	prec.@ρ	base rate
HotpotQA	0.478	0.522	0.560	0.542	0.613	**0.593**	66.6	59.5
2Wiki	0.421	0.579	0.557	0.567	0.601	**0.611**	82.7	63.0
MuSiQue	0.504	0.496	0.680	0.544	0.825	**0.705**	92.3	89.4
HybridQA	0.474	0.526	0.609	0.565	0.662	**0.632**	87.3	83.4
MoreHopQA	0.467	0.533	0.441	0.638	0.509	**0.580**	84.0	88.4

6.1 Setup

Datasets. We use the distractor setting of HotpotQA, 2WikiMultihopQA, the answerable subset of MuSiQue, HybridQA and MoreHopQA. We evaluate on the full development set of each dataset and use all paragraphs as context.

Model and prompt. We use Llama-2-7B-Base (Touvron et al., 2023) with a direct-answer prompt given in Appendix B.

Baselines. We compare with greedy decoding, CAD (Shi et al., 2024) with $\alpha \in \{0.5, 1\}$, AdaCAD (Wang et al., 2025), NDAD (Qiu et al., 2026) and self-consistency (Wang et al., 2023) with $K = 5$ at $T = 0.7$.

Implementation and metrics. Hyperparameters are fixed across datasets and never tuned on labels: $\rho = 0.3$, $b = 1$, $\alpha_{\mathrm{a}} = \pm 0.5$, $k = 4$, $T = 0.7$, 3 votes and F1 threshold 0.8. We report EM and F1.

6.2 RQ1: Main results

Table 2 shows three results. First, methods that act on every answer hurt accuracy. Global CAD lowers the average EM from 17.0 to 10.0 at $\alpha = 1$ and to 14.6 at $\alpha = 0.5$, with the largest drops on HotpotQA and 2WikiMultihopQA, where distractor paragraphs make lures common. Self-consistency lowers it to 12.5. Second, GPCD gives the best average EM and F1, 17.1 and 23.6, and the best EM and F1 on HotpotQA, 2WikiMultihopQA and MuSiQue. It never falls below greedy decoding. Third, the gains are small for every method. GPCD improves over greedy decoding by 0.1 EM and 0.7 F1 points on average, and the adaptive methods AdaCAD and NDAD stay within 0.6 EM points of greedy decoding. Confident hallucinations are hard to repair by decoding alone, which agrees with the analysis in Section 4. The model has no uncertainty to exploit, and for lures the context evidence itself points to the wrong answer.

6.3 RQ2: Stratified versus pooled detection

Proposition 2 and Section 5 predict that pooled scores of fixed direction are unreliable across datasets, while the stratified score is reliable on all of them. Table 3 tests this within the confident pool and reports the precision π of Eq.(9).

Table 4: **Attention of retrieval heads at the first answer token.** Ctx is the attention mass on the context and Ent. its entropy over context tokens. Ans. and Gold are the mass on the generated and the gold answer span. Ptr is the share of retrieval heads whose most attended token lies in that span.

Dataset	Group	n	Ctx	Ent.	Ans.	Gold	Ptr→Ans.	Ptr→Gold
HotpotQA	correct	300	0.673	4.27	0.174	0.174	40.7%	40.4%
	lure	300	0.709	4.33	0.133	0.065	29.8%	13.7%
	prior	261	0.558	4.66	0.003	0.057	0.9%	14.5%
2WikiMultihopQA	correct	93	0.479	4.41	0.040	0.040	13.0%	13.2%
	lure	278	0.595	4.27	0.057	0.033	13.7%	8.7%
	prior	36	0.616	4.34	0.000	0.027	0.0%	6.8%

Table 3 confirms both predictions. The pooled masked term is close to chance in either direction, with AUROC between 0.42 and 0.58, and its better direction changes across datasets. Within each stratum the predicted direction holds: supported errors have lower m and unsupported errors higher m than correct answers of the same stratum. Support alone is informative on extractive datasets but falls below chance on MoreHopQA, whose answers are computed. The stratified score combines both signals and is the best detector on all five datasets, improving over support alone by 2 to 14 AUROC points.

6.4 RQ3: Attention check of the retrieval account

Section 4 attributes lures to retrieval that copies a wrong span and prior errors to retrieval that finds no matching span. We test this at the query that emits the first answer token of confident answers to extractive questions. One extra forward pass records the attention of every head on the context and on context occurrences of the generated and the gold answer. Retrieval heads are the 20 heads with the largest attention on the gold span, selected on a held-out half of correct supported answers, and all numbers use the other half and all wrong answers.

Table 4 confirms both failure modes. For lures, retrieval heads attend more to the wrong span than to the gold span, with 0.133 against 0.065 on HotpotQA and 0.057 against 0.033 on 2WikiMultihopQA, and they point at the gold span far less often than for correct answers. For prior errors, attention on the gold span is the lowest of the three groups. On HotpotQA their attention also covers less of the context and is more diffuse. Across answers, the grounding score correlates with retrieval-head attention on the answer span, with Spearman $\rho = 0.48$ on HotpotQA and 0.40 on 2WikiMultihopQA. The retrieval heads selected independently on the two datasets share 12 of their top 20 and all lie in layers 14 to 29.

7 Limitations

Confident extractions on unanswerable questions look like correct extractions to our signals, so answerability errors call for abstention (Yona et al., 2026) and are out of scope, as shown in Appendix F. Lexical support suits extractive answers, and the retrieval account is supported by attention but is not a circuit-level proof. We study short-answer QA with a single 7B model and direct-answer prompting, and the per-stratum statistics need a batch of test queries.

8 Conclusion

This paper studies confident hallucinations in context-grounded QA, which entropy-based methods cannot detect. We show that an answer's score splits into a parametric prior and context evidence, and that confident hallucinations arise when retrieval either copies a wrong span or finds no matching span. Because benchmarks mix the two types, pooled grounding signals flip sign across datasets. We propose GPCD, which scores answers within context-support strata and corrects each type in its own direction. Experiments on five benchmarks show that GPCD detects these errors more reliably than pooled signals and gives the best average accuracy. In future work, we plan to extend this analysis to chain-of-thought reasoning and to larger models.

AI USE STATEMENT

Claude and Codex assisted with polishing the writing, preparing data, running experiments and analyzing results. The authors reviewed all outputs and take full responsibility for the content.

ETHICS STATEMENT

This work studies and mitigates hallucinations of publicly available language models on public QA benchmarks. It involves no human subjects or private data. Improving the faithfulness of context-grounded generation is intended to reduce harm from confidently wrong outputs. The analysis of context lures also clarifies how misleading context can steer models, and we recommend combining the method with source verification in deployment.

REPRODUCIBILITY STATEMENT

Section 5 and Algorithm 1 fully specify the method. All hyperparameters are listed in Section 6.1 and are identical across datasets. Data preprocessing is described in Appendix B. Anonymized code will be provided as supplementary material, including the dual-stream decoder, the stratified score, the acceptance rule, and the diagnostic scripts that produce Tables 1, 3 and 4.

REFERENCES

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in LLM reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1026–1036, 2020.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating. *arXiv preprint arXiv:2605.24024*, 2026.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024b.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. In *International Conference on Learning Representations*, 2025.

Anxin Guo and Jingwei Li. Hallucination is a consequence of space-optimality: A rate-distortion theorem for membership testing. *arXiv preprint arXiv:2602.00906*, 2026.

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, 2020.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2021–2031, 2017.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 12286–12312, 2023b.

Yee Zhing Liew, Andrew Huey Ping Tan, and Anwar P. P. Abdul Majeed. From flat facts to sharp hallucinations: Detecting stubborn errors via gradient sensitivity. *arXiv preprint arXiv:2605.00939*, 2026.

Litian Liu, Reza Pourreza, Sunny Panchal, Apratim Bhattacharyya, Yubing Jian, Yao Qin, and Roland Memisevic. Enhancing hallucination detection through noise injection. In *International Conference on Learning Representations*, 2026.

Yongchao Long, Yingying Zhang, Xianbin Wen, Xian Wu, Yuxi Zhou, and Shenda Hong. Copy-paste to mitigate large language model hallucinations. In *International Conference on Learning Representations*, 2026.

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, 2021.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Lucas H. McCabe, Rimon Melamed, Thomas Hartvigsen, and H. Howie Huang. Estimating semantic alphabet size for LLM uncertainty quantification. In *International Conference on Learning Representations*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Panjia Qiu, Mingyuan Fan, Cen Chen, and Daixin Wang. NDAD: Negative-direction aware decoding for large language models via controllable hallucination signal injection. In *International Conference on Learning Representations*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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 (Short Papers)*, pp. 784–789, 2018.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. MoreHopQA: More than multi-hop reasoning. *arXiv preprint arXiv:2406.13397*, 2024.

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 *International Conference on Machine Learning*, 2023.

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 (Short Papers)*, pp. 783–791, 2024.

Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, 2025.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *International Conference on Learning Representations*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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.

Liam van der Poel, Ryan Cotterell, and Clara Meister. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

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*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, 2025.

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 *International Conference on Learning Representations*, 2024.

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*, 2024.

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, 2018.

Samuel Yeh, Sharon Li, and Tanwi Mallick. LUMINA: Detecting hallucinations in RAG system with context–knowledge signals. In *International Conference on Learning Representations*, 2026.

Gal Yona, Mor Geva, and Yossi Matias. Hallucinations undermine trust; metacognition is a way forward. *arXiv preprint arXiv:2605.01428*, 2026.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024.

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*, 2024.

Xinyue Zeng, Junhong Lin, Liang Shi, Yujun Yan, Jun Wu, Feng Guo, and Dawei Zhou. HalluGuard: Demystifying data-driven and reasoning-driven hallucinations in LLMs. In *International Conference on Learning Representations*, 2026.

Jianxiong Zhang, Bing Guo, Yuming Jiang, Haobo Wang, Bo An, and Sean Du. Harnessing reasoning trajectories for hallucination detection via answer-agreement representation shaping. *arXiv preprint arXiv:2601.17467*, 2026.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 2024.

Yifan Zhu, Huiqiang Rong, and Haoran Luo. Token-guard: Towards token-level hallucination control via self-checking decoding. In *International Conference on Learning Representations*, 2026.

Zhengtao Zou, Ya Gao, Jiarui Guan, Bin Li, and Pekka Marttinen. Adaptive residual-update steering for low-overhead hallucination mitigation in large vision language models. *arXiv preprint arXiv:2511.10292*, 2025.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.

A Extended Related Work

Uncertainty-based detection and confident errors. Most label-free hallucination detectors assume that errors are uncertain. They measure token entropy, semantic entropy over samples (Kuhn et al., 2023; Farquhar et al., 2024; McCabe et al., 2026), self-consistency (Manakul et al., 2023; Wang et al., 2023), or perturbation-induced dispersion (Liu et al., 2026). Recent work shows that a sizeable class of errors is confident. Theory predicts that capacity-limited models must assign high confidence to some non-facts (Guo & Li, 2026; Kalai et al., 2025), and "stubborn" hallucinations defeat both entropy- and representation-based detectors (Liew et al., 2026; Yona et al., 2026). Existing responses to confident errors are detectors that need gradients (Liew et al., 2026) or trained probes (Zhang et al., 2026; Zeng et al., 2026). Entropy minimization (Wang et al., 2021; Agarwal et al., 2025) and self-rewarded test-time RL (Zuo et al., 2025) push the model toward its own confident answers, which removes uncertain errors but can reinforce confident ones. We instead explain why such errors are confident in the context-grounded setting, and we mitigate them without labels or weight updates.

Context versus parametric knowledge. Knowledge-conflict studies construct counterfactual contexts and ask whether models follow them (Longpre et al., 2021; Xie et al., 2024; Xu et al., 2024). Other studies show that irrelevant context distracts models (Shi et al., 2023; Yoran et al., 2024) and that instruction tuning can reduce context reliance (Goyal et al., 2025). On the decoding side, context-aware decoding (CAD) amplifies the pointwise mutual information between the context and the output (Shi et al., 2024; van der Poel et al., 2022). Adaptive variants set its strength from a measured conflict (Wang et al., 2025; Yuan et al., 2024), and copying-oriented training raises context faithfulness (Long et al., 2026). The same "more evidence is better" assumption drives language-prior suppression in vision-language models (Zou et al., 2025; Cheng et al., 2026). All of these methods correct in one direction, toward the context. We show that over-reliance on a wrong context span is itself a major source of confident hallucination, which global CAD makes worse. Unlike conflict studies, we analyze naturally occurring errors without labels, and we find that conflicts in both directions produce low-entropy errors.

Mechanistic detection and decoding-time mitigation. ReDeEP (Sun et al., 2025), Lookback Lens (Chuang et al., 2024a) and LUMINA (Yeh et al., 2026) detect RAG hallucinations from the balance between context use and parametric knowledge, and treat low context use as the failure mode. Lookback Lens also steers decoding, but only toward the context. Other decoding-time mitigations contrast layers (Chuang et al., 2024b), subtract a hallucination direction elicited by masking attention heads (Qiu et al., 2026), or self-check tokens during generation (Zhu et al., 2026). GPCD differs in three ways. It is bidirectional, with the direction chosen per answer. It is conservative: it changes an answer only with agreement and context support, because most flagged answers are correct. Its score takes its directions from the mechanism and needs no labels, no trained probe, and no clustering.

B Implementation Details

Prompts. Direct-answer prompt (system): "*Using the provided context, output only the final answer. Do not explain. Do not use a complete sentence.*" The user turn is

```
Context:\n{context}\n\nQuestion: {question}
\n\nAnswer:
```

so generation starts at the answer. Each paragraph is prefixed with its title. For models without a chat template we concatenate the system and user text, and the BOS token is placed outside the masked span.

Masked prompt. The full and masked prompts are tokenized as

$$[\text{prefix}]\,[\text{context}]\,[\text{suffix}] \quad \text{and} \quad [\text{prefix}]\,[\text{filler}]^{|\text{context}|}\,[\text{suffix}],$$

so their lengths and positions are identical.

Dual-stream decoding. The full and masked rows of each example are left-padded, placed in one batch, and share a KV cache. At each step the next token is chosen from the (possibly steered)

score of Eq.(8) and appended to both rows. g_t, ℓ_t and H_t are always recorded under the unsteered distributions. A sanity check confirms that batched and single-example decoding agree, and that g_t and ℓ_t match a teacher-forced recomputation to within 0.1 nats. Batches are formed under a token budget, so that $2\times$ rows $\times$ length stays bounded.

Answer span. The answer starts at the first generated token and ends at the first newline or EOS. We read g and m at this onset.

Hardware and software. All experiments run on eight NVIDIA RTX 3090 GPUs with 24 GB of memory each. The code uses PyTorch 2.3 or later and Transformers 4.45 or later.

C Prompts for the Grounding Analysis

The token-level curves of Section 3 use the following user prompts. They differ from the direct-answer prompt of GPCD in Appendix B, which asks for the bare answer span. HotpotQA, 2WikiMultihopQA and MuSiQue share one context-QA template:

```
Answer the following question briefly in one sentence using the provided context.

Context:
{context}

Question: {question}

Answer:
```

MoreHopQA, HybridQA and FanOutQA keep the prompt of their original release. MoreHopQA uses

```
Answer the following question briefly using the provided context.

Context:
{context}

Question: {question}

Answer:
```

and HybridQA uses

```
Answer the question briefly using the provided table and linked passages.

Context:
{table_and_passages}

Question: {question}

Answer:
```

SQuAD 2.0 uses the same prompt for generation and for the grounding score:

```
Answer the following question briefly using only the provided context.
If the question cannot be answered from the context, answer exactly 'unanswerable'.

Context:
{context}

Question: {question}

Answer:
```

In every case the masked prompt replaces the context tokens with the filler token and leaves the rest unchanged.

D Pilot: Correcting Every Deviation

The pilot used Llama-2-7B with the direct-answer prompt on $n = 1{,}000$ HotpotQA questions. We fitted a Gaussian mixture to the grounding features of the unlabeled test set, flagged answers far from the heaviest component, re-decoded them with type-directed CAD, and selected the candidate whose

grounding score was closest to normal. The gate flagged 24.3% of answers, all on the prior side. Repair fixed 9 answers and broke 69, and EM fell from 34.0 to 28.0. Global CAD with $\alpha = 1$ lost 13.7 EM points (34.0 to 20.3). The failure has two causes, both analyzed in Section 5. Most flagged answers were correct (base rate), and selecting by score favored steered candidates regardless of correctness.

E ALGORITHM

Algorithm 1 GPCD (label-free, no weight updates)

Require: test set $\{(c_i, q_i)\}_{i=1}^{N}$; model p_θ; ρ, b, $\alpha_{\mathrm{a}}(\cdot)$, k, T

for $i = 1, \ldots, N$ **do** ▷ Pass 1: batched dual-stream greedy decoding
  $a_i, \{g_t, \ell_t, H_t\} \leftarrow \textsc{DualDecode}(c_i, \tilde{c}_i, q_i; \alpha \equiv 0)$; read m_i^{ans}, $\bar{H}_i$; $\sigma_i \leftarrow \sigma(a_i)$ by Eq.(5)
end for
$\mathcal{P} \leftarrow \{i : \bar{H}_i \le Q_{0.5}(\bar{H})\}$; $h \leftarrow$ mean of σ_i over $\mathcal{P}$ ▷ regime
estimate mean and s.d. of m^{ans} in each stratum of $\mathcal{P}$; compute S_i by Eq.(6)
$\mathcal{F} \leftarrow$ top-ρ fraction of $\mathcal{P}$ by S ▷ type$_i$ = LURE if $\sigma_i = 1$, else PRIOR
for each $i \in \mathcal{F}$ **do**
  $\{a_i^{(j)}\}_{j=1}^{k} \leftarrow \textsc{DualDecode}(c_i, \tilde{c}_i, q_i; \alpha_{\mathrm{a}}(\mathrm{type}_i), T_j)$ with $T_1 = 0$, $T_{j>1} = T$ ▷ Eq.(8)
  $a^\star, n^\star \leftarrow$ majority normalized answer and its count
  if $n^\star \ge \lceil 3k/4 \rceil$ **and** $a^\star \ne a_i$ **and** ($h < 0.5$ **or** $\sigma(a^\star) = 1$) **then**
    $a_i \leftarrow a^\star$ ▷ accept only with agreement and support
  end if
end for
return $\{a_i\}$

F ANSWERABILITY AND AGGREGATION CASES (SQUAD 2.0, FANOUTQA)

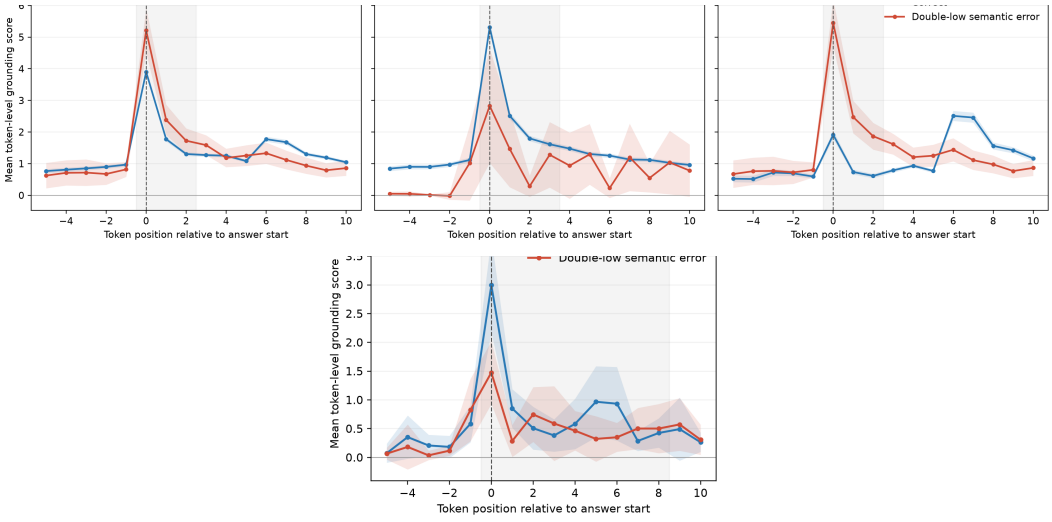


Figure 4: **Scope cases.** Top: SQuAD 2.0, all / answerable / unanswerable questions. Bottom: FanOutQA with evidence provided. Red: confident hallucinations (low conditional and context-free entropy).

On unanswerable questions, confident hallucinations peak far above correct abstentions (5.7 vs. 1.9 nats in g; masked term -5.8 vs. -2.2), and 69.7% of them are context spans, compared with 17.7% of correct answers. However, the correct group here consists of abstentions, whose first token is easy to predict without the context (66.1% argmax agreement). Once we compare against correct span extractions on answerable questions ($n = 5{,}763$; $m = -5.42$, agreement 32.0%, $g = 5.32$), the gap shrinks to -0.39 nats (95% bootstrap CI $[-0.92, 0.13]$) and the agreement is the same (32.7%). We therefore treat SQuAD 2.0 as suggestive evidence at most, and we place answerability errors outside

the scope of the method. On FanOutQA, which requires aggregating many facts, the grounding score does not separate the groups, and the confidence intervals are wide.

G COMPUTATIONAL COST

Pass 1 decodes two rows per example, about $2\times$ greedy decoding with shared batching. Repair decodes $2k$ rows for the flagged fraction of the confident pool, about 0.15 of the test set at $\rho = 0.3$. The total is about $2(1 + 0.15k) \approx 3.2$ greedy-equivalents.