ANSWERS AS PROBES: MITIGATING LLM HALLUCINATIONS WITH SOURCE-AWARE INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models that answer from a given passage still produce answers the passage does not support. We find that many of these errors are failures of reading, not of knowledge: if Llama-2-7B-chat reads each MRQA prompt in four slightly different ways and the best of its four answers is kept, macro F1 rises from 67.0 to 74.7. Decoding-time methods recover little of this gap, for two reasons. They build the alternative reading by deleting the context, which also changes how the question is encoded, and they commit to one reading token by token, before the answer that would reveal whether the reading was right exists. We propose ANSWERPROBE, a training-free decoder that addresses both. It intervenes only on what the *answer* may attend to, which leaves every state before the answer intact and lets several readings share one prefill and decode in one batch. Each reading proposes a complete answer; each answer then *probes* its source, and the unmodified model selects the one with the highest native belief plus source support: exact containment for text sources, an interventional likelihood ratio otherwise. On all 62,316 MRQA questions, ANSWERPROBE improves every domain for each of five models, gaining 3.0–4.2 F1 over greedy decoding and 2.1–2.3 F1 over the strongest published decoder at $1.7\times$ greedy cost, with no tuned hyperparameters. It also improves TruthfulQA MC1 for all five models tested. Used as an instrument, it shows that hallucinated answers do not read the passage less than correct ones, but read it through shallow layers and punctuation anchors.

1 INTRODUCTION

Large language models are increasingly asked to answer from evidence placed in the prompt, such as a retrieved document or a passage (Lewis et al., 2020). Even with the evidence in hand, they often produce answers it does not support, and they tend to follow memory when the two conflict (Ji et al., 2023; Huang et al., 2025a; Longpre et al., 2021). Decoding-time methods address this without retraining. Most follow the template of contrastive decoding (Li et al., 2023b) and compare the model with a copy of itself that lacks some evidence, obtained by deleting the context (CAD, COIECD; Shi et al., 2024; Yuan et al., 2024), by reading an earlier layer (DoLa, SLED, Activation Decoding; Chuang et al., 2024b; Zhang et al., 2024; Chen et al., 2024b), or by re-weighting context tokens by attention (DAGCD; Huang et al., 2025b).

Many of the remaining errors, however, are failures of reading rather than of knowledge (Figure 1A). If Llama-2-7B-chat reads each MRQA prompt in four slightly different ways (Section 3.3) and the best of its four answers is kept, macro F1 over all 62,316 questions rises from 67.0 to 74.7. For 14.5% of questions the greedy answer does not occur in the passage, and for almost half of these another reading produces one that does. The model can thus often produce the right answer; what fails is how it reads the passage and which reading it trusts. Existing decoders recover little of this 7.6-point gap: the strongest published decoder gains 0.7–2.1 F1, layer contrasts at most 0.4, and every baseline hurts some domain on some model (Table 1).

We trace this shortfall to two blind spots, in how current decoders construct an alternative reading and in when they choose among readings (Section 3.1). To measure what the passage contributes, CAD deletes it. In a causal model, however, the question is encoded while attending to the passage, so deletion also changes the question's representation. We show that CAD's score splits exactly into a source-information term, which ranks the better of two answers higher in 87% of pairs, and

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.

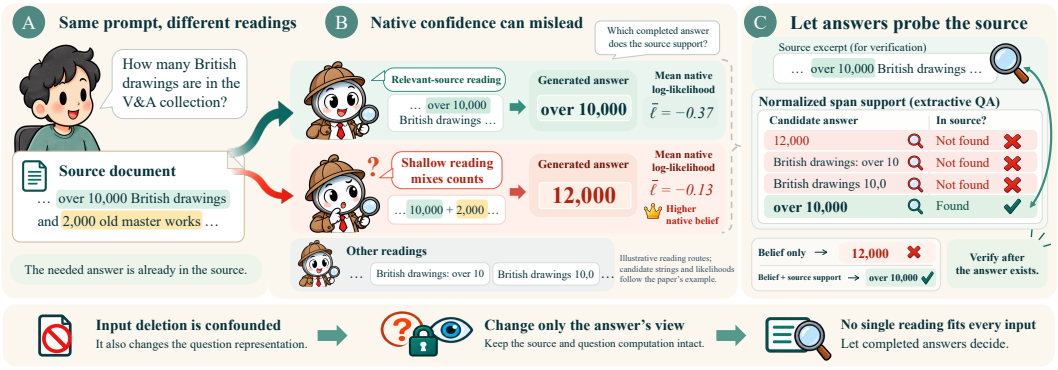


Figure 1: **Answers as probes of their source** (SQuAD example). (A) The answer is in the passage, but different readings of the same prompt give different answers: reading the relevant span gives "over 10,000", whereas a shallow reading merges two counts into "12,000". (B) The native model prefers the unsupported answer ($\bar{\ell} = -0.13$ vs. -0.37). (C) Once the answers are complete, each is checked against the source; only "over 10,000" is a passage span, and belief plus source support selects it. Bottom: the three design principles of ANSWERPROBE.

a question re-encoding term, which does so in only 20%. A token-level decoder, moreover, fixes its reading before generating, although whether a reading was right is a property of its answer, most simply whether the answer occurs in the passage: the better reading is ranked higher in only 55% of pairs at the first token, but in 75% once the answers are complete and checked against the source (Figure 1B–C).

We propose ANSWERPROBE (Figure 2), a training-free decoder built on two principles that follow from these blind spots: to change how a model reads, change what the *answer* can see rather than what the input contains, and choose among readings by their *complete answers*. Its single intervention blocks the attention of the *answer rows*, the positions that emit the answer, to one region of the prompt. Because attention is causal, every state before the answer, including the question's, is exactly that of the unmodified model, so the contrast is free of the re-encoding confound. Applied to the passage and to the punctuation-like *anchor* tokens into which models aggregate context (Wang et al., 2023; Chen et al., 2024a), and either amplified or reversed, the intervention yields a small family of readings, each of which proposes a complete answer. Each answer then *probes* its source, and the unmodified model selects the proposal with the highest posterior: its own belief plus the source's support, which is exact containment for text and an interventional likelihood ratio otherwise.

This design needs no training, auxiliary model or learned verifier, and it uses only the standard defaults of contrastive decoding. All readings share one prefill and decode as rows of one batch, at $1.68\times$ greedy cost, which is less than CAD. Because the decision is deferred to complete answers, a reading only needs to be right where the others are wrong: the two anchor readings are each worse than greedy decoding, yet each adds accuracy to the family.

Our contributions are:

- An exact decomposition that exposes a question re-encoding confound in input-deletion contrasts, and evidence that the right reading can be identified only after its answer exists (Sections 3.1 and 5).
- ANSWERPROBE: answer-row interventions, the family of readings they induce, and answer-level posterior selection for open- and closed-book QA (Section 3).
- A full 62k-question evaluation on five models from four families, on which ANSWERPROBE is the only method that improves every domain on every model (+3.0–4.2 F1 over greedy decoding), with ablations, controls and mechanism analyses of how hallucinated answers read their source (Sections 4–6).

2 Related Work

Contrastive and context-aware decoding. Contrastive decoding scores tokens by the gap between a strong and a weak distribution (Li et al., 2023b). CAD contrasts the model with and without the context (Shi et al., 2024), COIECD gates this contrast by a conflict test (Yuan et al., 2024), and DAGCD guides it with attention (Huang et al., 2025b). For factuality, DoLa, SLED and Activation Decoding contrast or refine layers (Chuang et al., 2024b; Zhang et al., 2024; Chen et al., 2024b), and ICD and ITI use a hallucination-tuned model or activation shifts (Zhang et al., 2025a; Li et al., 2023a). All fix one modified distribution in advance and apply it at every token (Appendix Table 4). AnswerProbe changes where the intervention acts, on the answer's attention rows, and when the decision is made, after answers exist.

How LLMs aggregate context. Label words and separators act as anchors that aggregate segments (Wang et al., 2023; Chen et al., 2024a; Pang et al., 2024), attention sinks collect disproportionate attention (Xiao et al., 2024; Gu et al., 2025; Binkowski et al., 2026), and over-trust in summary tokens has been linked to hallucination (Huang et al., 2024). Attention knockout (Geva et al., 2023; Meng et al., 2022; Ferrando & Voita, 2024) supplies our causal tool, which we use at decoding time.

Detection and verification. Lookback Lens (Chuang et al., 2024a) and ReDeEP (Sun et al., 2025) detect contextual hallucination. Self-consistency (Wang et al., 2022), trained verifiers (Cobbe et al., 2021) and self-evaluation (Kadavath et al., 2022; Manakul et al., 2023; Dhuliawala et al., 2024) verify sampled outputs, and RAG systems ground answers in retrieved text (Lewis et al., 2020; Asai et al., 2024; Gao et al., 2023). AnswerProbe repairs rather than detects, proposes by intervening on reading rather than by sampling, and its verifier needs no training, extra prompt or retriever.

3 Method

AnswerProbe separates *proposing* answers from *deciding* among them (Figure 2). Several readings of the prompt, each an intervention on what the answer may attend to, propose complete answers; the unmodified model decides after each answer has been checked against its source.

3.1 Setup and the two blind spots

A prompt $x = (x_1, \dots, x_n)$ contains a *source* region $\mathcal{S} \subseteq [n]$ (the passage in reading comprehension; the instruction and demonstrations in closed-book QA), a question and an answer cue. A causal model p_θ generates $y = (y_1, \dots, y_m)$, with y_t predicted at position $n + t - 1$; we call $\mathcal{T} = \{n, \dots, n + m - 1\}$ the *answer rows*. The model's belief in an answer is its mean token log-likelihood

$$\bar{\ell}(y \mid x) = \tfrac{1}{m} \textstyle\sum_{t=1}^{m} \log p_\theta(y_t \mid x, y_{<t}). \tag{1}$$

Contrastive methods pair p_θ with a *reference* q_t, the model deprived of some evidence, and decode from

$$\tilde{p}_t(v) \propto p_\theta(v \mid x, y_{<t}) \Big(\frac{p_\theta(v \mid x, y_{<t})}{q_t(v)} \Big)^{\beta}, \quad v \in \mathcal{V}_t = \{v : p_\theta(v \mid \cdot) \geq \tau \max_u p_\theta(u \mid \cdot)\}, \tag{2}$$

where $\mathcal{V}_t$ is the plausibility constraint of Li et al. (2023b). CAD deletes the source from the input, $q_t = p_\theta(\cdot \mid x_{\setminus \mathcal{S}}, y_{<t})$ (Shi et al., 2024).

Blind spot 1: input deletion confounds the contrast. The question follows the passage, so its hidden states are computed while attending to it. Deleting $\mathcal{S}$ changes both what the answer can read and how the question is represented. Let $p_\theta^{\neg \mathcal{S}}$ be the model whose answer rows cannot attend to $\mathcal{S}$ (Section 3.2), and $\bar{\ell}^{\neg \mathcal{S}}$ its analogue of Eq. (1). CAD's answer-level evidence decomposes *exactly* (Appendix D, D3):

$$\underbrace{\bar{\ell}(y \mid x) - \bar{\ell}(y \mid x_{\setminus \mathcal{S}})}_{\text{CAD evidence}} = \underbrace{\bar{\ell}(y \mid x) - \bar{\ell}^{\neg \mathcal{S}}(y \mid x)}_{\iota_{\mathcal{S}}(y)\text{: source information}} + \underbrace{\bar{\ell}^{\neg \mathcal{S}}(y \mid x) - \bar{\ell}(y \mid x_{\setminus \mathcal{S}})}_{\delta_{\mathcal{S}}^{\mathrm{enc}}(y)\text{: question re-encoding}}. \tag{3}$$

$\iota_{\mathcal{S}}$ compares two computations that share every state before the answer and differ only in whether the answer may read $\mathcal{S}$. $\delta_{\mathcal{S}}^{\mathrm{enc}}$ compares two computations in which the answer cannot read $\mathcal{S}$ but the question is encoded with or without it; nothing ties it to whether y is supported. Empirically it is

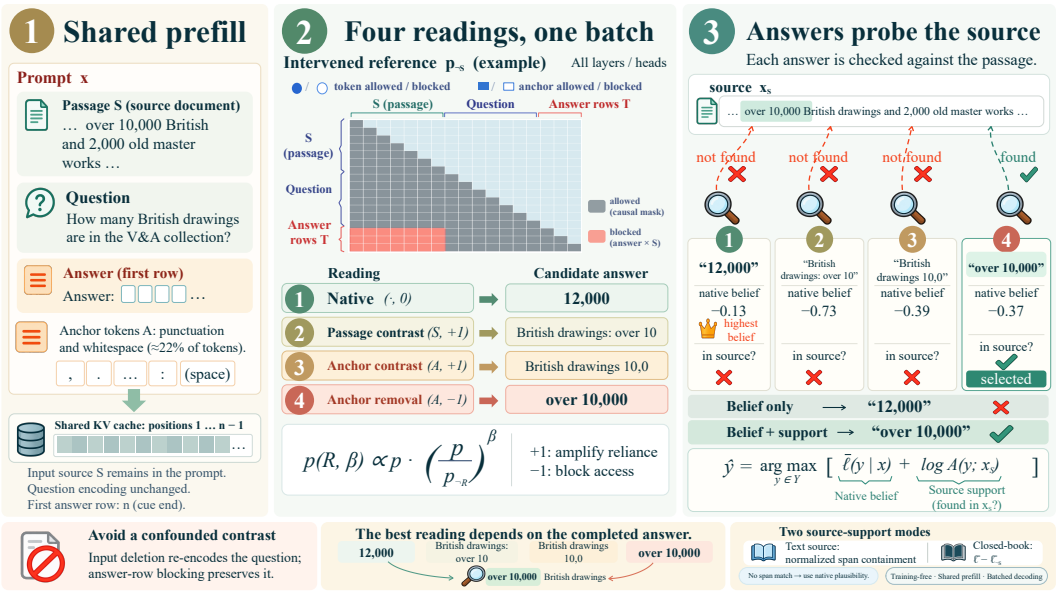


Figure 2: **Overview of ANSWERPROBE** (running SQuAD example). ① The prompt, with source passage $\mathcal{S}$, question and anchor tokens $\mathcal{A}$, is prefilled once; the key–value cache of positions $1, \ldots, n-1$ is shared by all readings, so the question's encoding never changes. ② Each reading $p_\theta^{(\mathcal{R},\beta)}$ contrasts the model with a reference whose answer rows $\mathcal{T}$ cannot attend to a region $\mathcal{R}$ (mask shown for $\mathcal{R} = \mathcal{S}$; red cells are blocked): $\beta = +1$ amplifies reliance on $\mathcal{R}$, and $\beta = -1$ removes it. The four readings decode as rows of one batch, and each proposes a complete answer. ③ Each distinct answer probes the source. Native belief alone selects the unsupported "12,000"; adding source support (Eq. (6)) selects "over 10,000", the only passage span. Bottom: design principles, and the two forms of source support: span containment for text, $\bar{\ell} - \bar{\ell}^{\neg\mathcal{S}}$ (Eq. (7)) for closed-book QA.

worse than uninformative: within a question, $\iota_{\mathcal{S}}$ ranks the better of two proposals higher in 87% of pairs, $\delta_{\mathcal{S}}^{\text{enc}}$ in 20% (Finding 1).

Blind spot 2: the reading is committed before the answer exists. Eq. (2) fixes one reference and one β for every token of every input. Whether that reading suits an input is a property of the *answer* it produces, which is unobservable while the answer is being formed: the first-token log-probability ranks the better reading higher in only 55% of pairs (Finding 2).

3.2 ANSWER-ROW INTERVENTIONS

For a region $\mathcal{R} \subseteq [n]$, the *answer-row intervention* $p_\theta^{\neg\mathcal{R}}$ is the same network with the pre-softmax attention logits

$$a_{ij}^{(l,k)} \leftarrow -\infty \quad \text{for } i \in \mathcal{T},\ j \in \mathcal{R},\ \text{in every layer } l \text{ and head } k, \tag{4}$$

which equals zeroing those attention weights and renormalising (D5). It is the smallest change that alters what the answer reads: nothing is deleted from the input, and no layer, weight or other query row is modified.

Prompt invariance. *For every $\mathcal{R}$, the hidden state of every position $i < n$ in every layer is identical under p_θ and $p_\theta^{\neg\mathcal{R}}$*, since no prompt position attends to an answer row (D4). Two consequences follow. The question is represented identically under every answer-row intervention, so $\iota_{\mathcal{S}}$ is free of the re-encoding confound by construction. And the key–value cache of positions $1, \ldots, n-1$ is shared by all interventions, which makes running several of them cheap.

3.3 A ONE-PARAMETER FAMILY OF READINGS

Setting $q_t = p_\theta^{\neg\mathcal{R}}$ in Eq. (2) gives, for each region, a one-parameter family

$$\log p_\theta^{(\mathcal{R},\beta)}(v) = (1+\beta)\log p_\theta(v) - \beta \log p_\theta^{\neg\mathcal{R}}(v) + \text{const.} \tag{5}$$

$\beta = 0$ is the native model; $\beta = +1$ doubles the evidence that flows through $\mathcal{R}$, *requiring* the answer to rely on it; $\beta = -1$ is exactly $p_\theta^{\neg\mathcal{R}}$, *forbidding* it (D2).

We use two regions. The passage $\mathcal{S}$ is the obvious one. The second is the set of *anchors* $\mathcal{A}$, the punctuation and whitespace tokens (22% of MRQA prompt tokens), into which models aggregate the information of a segment (Wang et al., 2023; Chen et al., 2024a). The passage thus reaches the answer by two routes: directly from content tokens, and through the summaries stored at anchors. The compressed route suffices for some inputs but, for others, merges details the question needs kept apart, so either direction of the anchor contrast is a plausible way to read an input. The family is $\mathcal{F} = \{(\cdot, 0), (\mathcal{S}, +1), (\mathcal{A}, +1), (\mathcal{A}, -1)\}$: native, *passage contrast*, which counteracts parametric override; *anchor contrast*, which trusts the aggregated summary; and *anchor removal*, which reads content tokens directly. $(\mathcal{S}, -1)$ cannot produce passage-supported answers and is excluded. Contrast readings decode greedily within $\mathcal{V}_t$, the others greedily, all with the contrastive-decoding defaults $\beta = 1$, $\tau = 0.1$. No member is right for every input; the family is useful because its members fail on *different* inputs.

3.4 ANSWERS AS PROBES: SELECTION AS POSTERIOR INFERENCE

Let $\mathcal{Y}$ be the distinct proposals and $y^\star$ the correct answer. We select the proposal with the highest posterior under a two-factor model, the native belief as prior and the source's support Λ as likelihood (D6):

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \underbrace{\bar{\ell}(y \mid x)}_{\text{native belief}} + \underbrace{\log \Lambda(y; x_{\mathcal{S}})}_{\text{source support}}. \tag{6}$$

The readings act only as proposal distributions; the posterior is always evaluated by the *unmodified* model. This matters because the readings are distorted distributions whose likelihoods are not comparable, and because a reading that is a poor decoder on average can still contribute the one correct proposal for a question. Selection uses no gold answers, training data or threshold.

Text source: exact support. In extractive QA, gold answers are passage spans, so $\Lambda = 1$ if the normalised y is a contiguous word subsequence of the normalised passage, and $\epsilon \to 0$ otherwise. Eq. (6) then selects the most plausible contained proposal, or the most plausible overall if none is contained. The two factors fail on different inputs: containment rejects fluent but unsupported answers, and $\bar{\ell}$ ranks several contained spans, which a string test cannot.

Non-text source: interventional support. When the source cannot be searched, we measure how much of the answer's likelihood the source supplies, with the same intervention:

$$\log \Lambda(y; x_{\mathcal{S}}) = \lambda\, \iota_{\mathcal{S}}(y) = \lambda\big(\bar{\ell}(y \mid x) - \bar{\ell}^{\neg\mathcal{S}}(y \mid x)\big), \qquad \lambda = 1. \tag{7}$$

Eq. (6) becomes $2\bar{\ell} - \bar{\ell}^{\neg\mathcal{S}}$, the answer-level counterpart of the $(\mathcal{S}, +1)$ reading, free of CAD's re-encoding term. It is a graded version of exact support: contained proposals have median $\iota_{\mathcal{S}} = +2.8$ nats per token, uncontained ones -2.0 (Finding 4).

Among decoding-time interventions, ANSWERPROBE is the only one that changes neither the input nor the layers, and the only one that decides per answer rather than per token (Appendix Table 4). Greedy decoding, a de-confounded CAD (the single reading $(\mathcal{S}, +1)$) and likelihood reranking ($\Lambda \equiv 1$) are special cases, which Section 6 evaluates. By prompt invariance, we prefill positions $1, \ldots, n-1$ once and decode all readings as rows of one batch, each intervened row receiving the additive bias of Eq. (4). A native row follows each reading's tokens and returns $\bar{\ell}$ at no extra cost. At batch size one, the four readings cost $1.68\times$ greedy wall time, against $1.75\times$ for CAD, which must prefill a second prompt (Algorithm 1).

Table 1: **ANSWERPROBE is the only method that improves every domain on every model.** F1 per domain and average F1/EM on the full DAGCD evaluation (62,316 questions per model). Families (Table 4): ID input deletion, AG attention guidance, LC layer contrast, AP answer probing (ours). †Reported by Huang et al. (2025b) under the identical protocol, whose greedy results ours reproduce; LC methods are rerun by us with their released defaults. **Bold**: best, underline: second best per model; arrows: change over greedy decoding.

Method	Fam.	HotpotQA	TriviaQA	SearchQA	SQuAD	NewsQA	NQ	NQ-Swap	**Avg. F1**	Avg. EM
Llama-2-7B-chat										
Greedy		67.40	76.85	58.12	78.73	53.90	65.52	68.77	67.04	57.83
CAD†	ID	67.16	76.83	58.11	77.91	53.26	65.00	68.85	66.73 ↓0.31	57.14 ↓0.69
COIECD†	ID	67.03	77.33	58.98	79.56	54.55	66.84	70.13	67.77 ↑0.73	58.66 ↑0.83
DAGCD†	AG	68.61	75.95	58.13	79.76	54.81	66.92	70.30	67.78 ↑0.74	58.43 ↑0.60
DoLa	LC	66.74	76.99	58.20	77.33	53.02	65.00	68.59	66.55 ↓0.49	57.26 ↓0.57
Activation Dec.	LC	67.63	76.30	57.92	79.54	54.39	66.08	69.68	67.36 ↑0.32	58.24 ↑0.41
SLED	LC	67.12	76.58	58.02	77.78	53.18	65.12	68.28	66.58 ↓0.46	57.37 ↓0.46
ANSWERPROBE (ours)	AP	**69.58**	**78.79**	**61.35**	**82.13**	**57.03**	**68.53**	**73.04**	**70.06** ↑3.02	**61.25** ↑3.42
Mistral-7B-Instruct										
Greedy		72.15	75.63	49.64	79.41	54.71	66.95	67.82	66.61	57.01
CAD†	ID	64.81	75.95	50.96	72.92	51.90	58.49	53.75	61.25 ↓5.36	50.65 ↓6.36
COIECD†	ID	72.59	72.15	42.66	81.03	55.96	68.72	73.81	66.70 ↑0.09	56.77 ↓0.24
DAGCD†	AG	73.49	75.96	52.65	80.62	56.23	69.72	72.12	68.68 ↑2.07	58.96 ↑1.95
DoLa	LC	72.28	75.70	49.82	79.45	54.62	67.18	67.87	66.70 ↑0.09	57.08 ↑0.07
Activation Dec.	LC	72.22	75.63	49.73	79.41	54.74	66.94	67.80	66.64 ↑0.02	57.03 ↑0.02
SLED	LC	72.21	75.55	49.60	79.24	54.50	66.88	67.71	66.53 ↓0.09	56.87 ↓0.14
ANSWERPROBE (ours)	AP	**74.58**	**77.33**	**55.98**	**83.14**	**58.53**	**71.02**	**74.77**	**70.76** ↑4.15	**61.54** ↑4.53
Llama-2-13B-chat										
Greedy		69.93	79.36	71.95	79.48	55.08	69.16	61.79	69.54	59.89
CAD†	ID	69.66	79.37	71.95	78.70	54.44	68.63	61.92	69.24 ↓0.30	59.42 ↓0.47
COIECD†	ID	70.43	78.96	71.93	80.39	56.16	69.81	63.21	70.13 ↑0.59	60.39 ↑0.50
DAGCD†	AG	71.69	78.77	72.73	80.76	56.24	71.31	65.20	70.96 ↑1.42	61.26 ↑1.37
DoLa	LC	64.02	76.49	69.07	74.41	49.93	65.02	57.57	65.21 ↓4.32	53.96 ↓5.93
Activation Dec.	LC	70.86	79.36	71.62	80.12	55.81	69.65	62.07	69.93 ↑0.39	60.35 ↑0.45
SLED	LC	55.69	72.73	65.96	73.63	48.22	61.72	53.44	61.63 ↓7.91	47.78 ↓12.12
ANSWERPROBE (ours)	AP	**72.91**	**80.78**	**75.16**	**83.29**	**58.71**	**72.00**	**68.29**	**73.02** ↑3.48	**64.03** ↑4.14

4 EXPERIMENTS

4.1 SETUP

For open-book QA we follow the public DAGCD protocol (Huang et al., 2025b): the six MRQA validation sets (Fisch et al., 2019) and NQ-Swap (Longpre et al., 2021), 62,316 questions in total, with its prompt, a generation budget equal to the token length of the longest gold alias, and F1 on the raw output, macro-averaged over the seven domains. Our greedy decoding reproduces the published numbers (67.04 vs. 67.03). For closed-book QA we use TruthfulQA MC (Lin et al., 2022) with the standard instruction and six-shot primer (Chuang et al., 2024b). We develop on Llama-2-7B-chat (Touvron et al., 2023) and confirm without re-tuning on Mistral-7B-Instruct (Jiang et al., 2023), Llama-2-13B-chat, Llama-3-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct, adding Qwen3-1.7B (Yang et al., 2025) for TruthfulQA. The baselines cover all three families: input deletion (CAD, Shi et al., 2024; COIECD, Yuan et al., 2024) and attention guidance (DAGCD), as reported under this protocol, and layer contrast (DoLa, Activation Decoding, SLED), which we rerun with the defaults of their releases. ANSWERPROBE uses the contrastive-decoding defaults $\beta = 1$, $\tau = 0.1$ and $\lambda = 1$, and has no tuned hyperparameter. All design choices were fixed on a 1,050-question development panel, and every confirmation run followed a protocol frozen beforehand (Appendix E).

4.2 MAIN RESULTS

ANSWERPROBE improves every domain on every model, whereas every baseline trades domains. Each baseline applies one adjustment to every question and cannot undo it where it is wrong, so each helps some domains and hurts others (Table 1; per-domain view in Appendix Figure 5). ANSWERPROBE improves all seven domains on all three models, by +3.0, +4.2 and +3.5 F1 (+3.4, +4.5 and +4.1 EM), and exceeds the strongest baseline, DAGCD, by 2.1–2.3 F1 on each. The gain transfers

Table 2: **The gain transfers to newer model families.** Full DAGCD evaluation, gold-length budget in each model's own tokens (Qwen2.5 in bf16); no baselines are reported for these models under this protocol.

Model	Method	HotpotQA	TriviaQA	SearchQA	SQuAD	NewsQA	NQ	NQ-Swap	**Avg. F1**	Avg. EM
Llama-3-8B-Instruct	Greedy	65.81	78.19	72.10	73.60	54.81	63.05	48.97	65.22	53.66
	ANSWERPROBE (ours)	**68.85**	**79.64**	**75.12**	**77.50**	**58.68**	**66.92**	**54.81**	**68.79** ↑3.57	**58.12** ↑4.46
Qwen2.5-7B-Instruct	Greedy	64.96	71.45	65.70	71.75	53.60	60.79	56.88	63.59	50.17
	ANSWERPROBE (ours)	**67.59**	**73.46**	**68.60**	**75.64**	**57.40**	**65.91**	**60.73**	**67.05** ↑3.46	**55.11** ↑4.94

to newer families without re-tuning (Table 2): +3.6 on Llama-3-8B and +3.5 on Qwen2.5-7B, again on every domain. Across five models the gain stays within +3.0 to +4.2 although greedy accuracy spans six points.

The shortfall of each baseline family follows from its design (Appendix A.2). *Input deletion* helps under knowledge conflict but loses up to 5.4 F1 elsewhere, because it re-encodes the question (Finding 1); the single reading $(\mathcal{S}, +1)$, which removes only this confound, already improves on greedy for every model. *Attention guidance* avoids the confound but still decides per token and trades domains (-0.9 on TriviaQA for Llama-2-7B). *Layer contrasts* amplify stored knowledge, the wrong source when the evidence is in the passage.

The same rule carries over to closed-book QA (Appendix A.3, Table 5). With no passage to search, the source is the instruction and demonstrations, and selection uses the interventional support of Eq. (7). It improves TruthfulQA MC1 over the native model for all five models under both summed and length-normalised scores (+2.8 and +3.5 on average), and it is the best method on average under the standard summed protocol. The input-deletion analogue instead *lowers* MC1 (36.7 $\rightarrow$ 35.6 on Llama-2-7B), the closed-book face of blind spot 1.

5 WHAT THE PROBES REVEAL

Every quantity in ANSWERPROBE is an intervention on the model's own reading, so the method doubles as a measuring instrument. We use it to test the two blind spots and to ask how the model reads when it hallucinates. Analyses use fresh questions disjoint from development and are shown for Llama-2-7B-chat; Findings 1 and 2 replicate on Mistral-7B and Llama-2-13B (Appendix H).

Finding 1. CAD's contrast is dominated by question re-encoding, which favours the wrong answer.

Splitting CAD's evidence with Eq. (3) over 2,100 questions (Figure 3a), $\iota_{\mathcal{S}}$ ranks the better of two proposals higher in 87% of pairs, CAD's evidence in 74%, and $\delta^{\text{enc}}_{\mathcal{S}}$ in only 20%. The bias is largest where CAD fails: on TriviaQA and SearchQA, CAD ranks at chance while $\iota_{\mathcal{S}}$ still ranks correctly in 83% and 88% of pairs. Deleting the passage moves the question's hidden states by $1 - \cos \approx 0.33$ (Figure 3b). A question read in isolation evokes the model's most typical answer, and contrasting against it penalises every answer the model also knows, often the correct one.

Finding 2. Which reading is right is decidable only after the answer exists.

Over all 62,316 questions, for proposals whose F1 differ by at least 0.5, the better and worse proposals share their first token in only 17% of pairs, yet the first-token log-probability ranks the better one higher in only 55%. The completed answer's $\bar{\ell}$ does so in 65%, containment in 66%, and the posterior of Eq. (6) in 75%. Readings diverge at the first step, and that step carries almost no information about which divergence is right.

Finding 3. Hallucinated answers do not read the passage less; they read it through compressed channels.

Blocking the passage in all layers lowers $\bar{\ell}$ about equally for wrong and correct greedy answers (3.42 vs. 3.76 nats per token), but wrong answers depend more on the passage in the first four layers (0.87 vs. 0.31) and on anchors (0.34 vs. 0.10; Figure 3c). Anchors absorb 21–35% of the answer's attention from layer 5 on, yet blocking them barely moves a correct answer: they are a high-traffic, low-information channel. Closing it forces direct reading, which is worse on average but right where

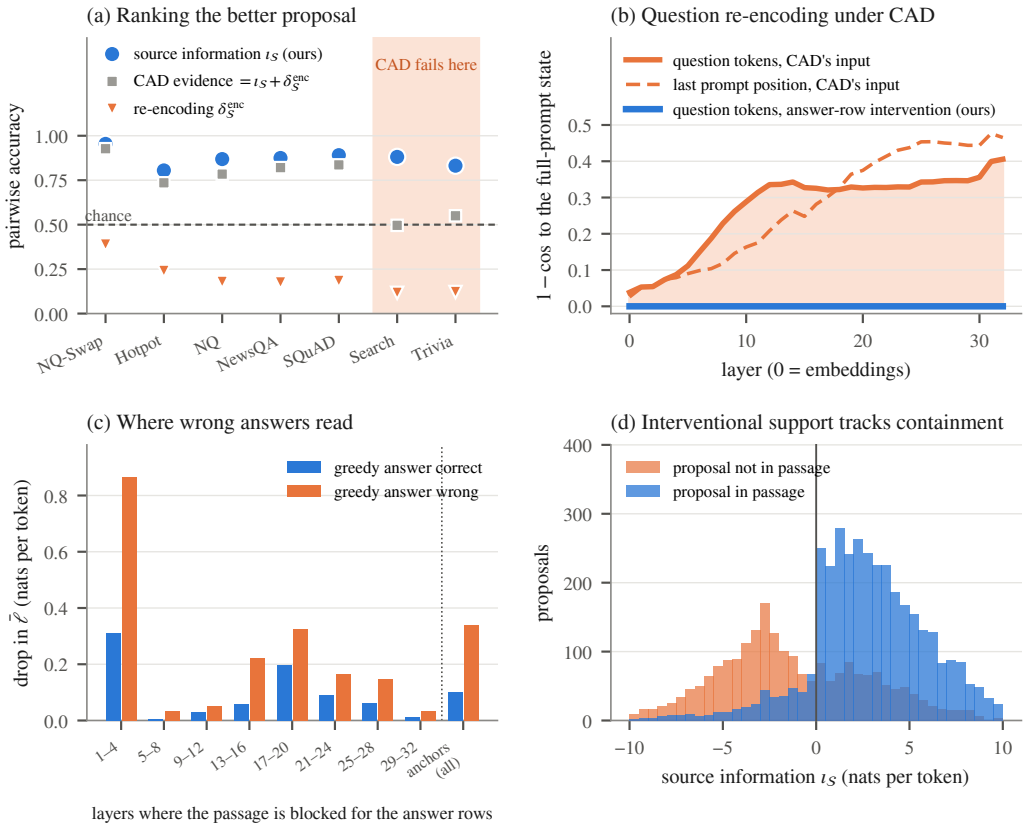


Figure 3: **What the probes reveal** (Llama-2-7B-chat, fresh MRQA questions). (a) Within-question accuracy of ranking the better of two proposals (F1 differing by ≥ 0.5) by each term of Eq. (3) (2,100 questions). (b) Drift of question-token and last-prompt-position states when the passage is deleted from the input; answer-row interventions leave both unchanged. (c) Drop in the greedy answer's $\bar{\ell}$ when its answer rows cannot read the passage in one band of four layers, or cannot read the anchors (700 questions). (d) Source information ι_S of all proposals, split by whether the passage contains them.

the compressed channel misled the model; this is the mechanism behind the anchor-removal reading (a worked example is in Appendix Figure 6).

Finding 4. Interventional support is a graded version of exact support.

Contained proposals have median $\iota_S = +2.8$ nats per token and uncontained ones -2.0 (Figure 3d); on NQ-Swap, greedy answers that follow memory have median 1.8, against 5.2 for faithful ones, which justifies Eq. (7). The change in an answer's state when the passage is blocked also organises answers by source use rather than topic, and correctness probes trained on it transfer better to knowledge conflict (8 of 8 settings; Appendix A.4).

6 Ablations and Controls

Both factors of the posterior are needed (Table 3). The native belief alone recovers a third of the gain (68.01): among plausible readings, the model still prefers the answer it would have produced anyway, often the unsupported one. Support alone, taking contained proposals in a fixed reading order, recovers three quarters (69.38). The two fail on different questions: support rejects plausible but unsupported answers, and belief ranks several contained spans, which a string test cannot.

Each intervened reading contributes what the others cannot: removing passage contrast, anchor contrast or anchor removal costs 0.81, 0.45 and 0.37 F1, although both anchor readings are worse than

Table 3: **Every reading and both factors of the posterior contribute.** Component ablation on the full MRQA set (Llama-2-7B-chat); Δ is relative to the full method. Readings as in Section 3.3, selection factors as in Eq. (6).

	Readings (proposals)				Selection			
Variant	native	$(\mathcal{S}, +1)$	$(\mathcal{A}, +1)$	$(\mathcal{A}, -1)$	$\bar{\ell}$	Λ	Macro F1	Δ
Greedy decoding	✓	✗	✗	✗	✗	✗	67.04	↓3.02
Selection: which factor of the posterior is used								
Belief only	✓	✓	✓	✓	✓	✗	68.01	↓2.05
Support only	✓	✓	✓	✓	✗	✓	69.38	↓0.68
Readings: which proposal is removed								
w/o passage contrast	✓	✗	✓	✓	✓	✓	69.25	↓0.81
w/o anchor contrast	✓	✓	✗	✓	✓	✓	69.62	↓0.45
w/o anchor removal	✓	✓	✓	✗	✓	✓	69.69	↓0.37
w/o native	✗	✓	✓	✓	✓	✓	70.04	↓0.02
ANSWERPROBE (full)	✓	✓	✓	✓	✓	✓	**70.06**	–

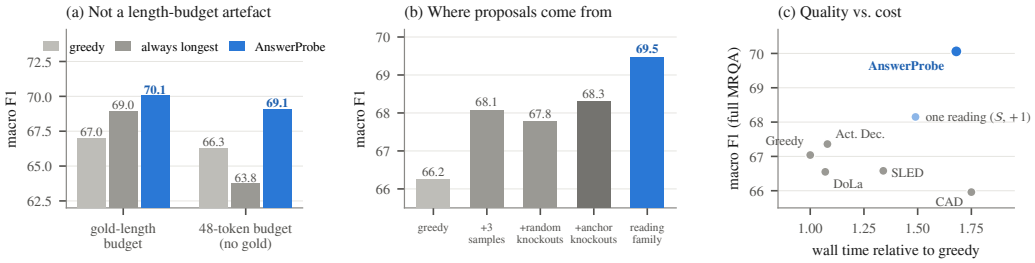


Figure 4: **The gain is not a length artefact, comes from structured readings, and is cheap** (Llama-2-7B-chat). (a) Gold-length budget (full MRQA) and fixed 48-token budget (fresh 7,000-question panel). (b) Equal-size candidate pools, same selection rule. (c) Wall time at batch size one.

greedy as decoders. They help because they are right on *different* questions (Appendix Table 14), a complementarity that a commit-in-advance decoder cannot exploit. Alone, the de-confounded reading $(\mathcal{S}, +1)$ scores 68.15, 68.21 and 70.10 on the three models, above both greedy and CAD (66.73, 61.25, 69.24): de-confounding alone removes CAD's losses, and the family adds the rest.

Controls rule out simpler explanations of the gain (Figure 4). Always taking the longest proposal gains only under the gold-length budget, whereas ANSWERPROBE gains under both budgets (+3.0 and +2.8), so its gain is not a length effect. At equal pool size, samples and random knockouts gain less than structured readings. The gain varies by 0.4 F1 over $\beta \in [0.5, 2]$, and MMLU and ARC-Challenge stay within ± 0.3 points. Answer-level selection also composes with other decoders: pairing DoLa, SLED or Activation Decoding with greedy decoding and applying Eq. (6) lifts each by 0.7–0.9 F1 (Appendix A.5).

7 CONCLUSION

Whether a reading was right becomes observable only once its answer can be checked against the source. ANSWERPROBE builds on this observation. It reads the prompt in several ways by changing only what the answer attends to, which keeps the question's representation intact, and lets the complete answers decide. Without training or tuning, it improves every MRQA domain on every model tested at $1.7\times$ greedy cost. Used as an instrument, it also shows that the contrast of input-deletion decoders is dominated by question re-encoding, and that hallucinated answers read the passage through shallow layers and anchor tokens rather than reading less of it. More broadly, our results suggest treating hallucination mitigation as a choice among readings made after answers exist, rather than as a fixed adjustment to every token.

LIMITATIONS

Exact support assumes extractive answers; abstractive tasks get only the weaker interventional support. The closed-book source region was chosen on Llama-2-7B before being frozen for the other models, and on TruthfulQA SLED is stronger under length-normalised scores. We show that hallucinated answers rely more on anchors, not that anchors cause hallucination. Answer-row interventions are weak in LLaVA-style VLMs, which absorb the image into the question (Appendix J), and require access to attention logits.

AI USE STATEMENT

We used LLM-based assistants as auxiliary tools to support implementation and experimentation, literature search, and writing. All reported numbers were computed by scripts from logged model outputs. The authors reviewed and verified all code, citations and claims, and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work aims to reduce unsupported content in LLM answers. All datasets are public research benchmarks used under their licenses, and no human subjects were involved. The method does not make models more capable of harmful behaviour. It inherits the biases of the underlying models and passages: an answer supported by a biased or false passage will be preferred.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix B specify the protocol, prompts, anchor and source definitions, normalisation and every hyperparameter (all are standard defaults). Our greedy decoding reproduces the published DAGCD greedy numbers for Llama-2-7B-chat and Mistral-7B-Instruct. Batched and sequential implementations agree on 99.4–99.7% of proposals over all 62,316 questions. We will release code, per-question outputs for every method, and the frozen analysis protocols as supplementary material.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. *arXiv preprint arXiv:2604.10697*, 2026.

Guoxuan Chen, Han Shi, Jiawei Li, Yihang Gao, Xiaozhe Ren, Yimeng Chen, Xin Jiang, Zhenguo Li, Weiyang Liu, and Chao Huang. Sepllm: Accelerate large language models by compressing one segment into one separator. *arXiv preprint arXiv:2412.12094*, 2024a.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*, 2024b.

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*, pp. 1419–1436, 2024a.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024b.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17432–17445, 2024.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd workshop on machine reading for question answering*, pp. 1–13, 2019.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

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.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems*, 43(2):1–55, 2025a.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Yanwen Huang, Yong Zhang, Ning Cheng, Zhitao Li, Shaojun Wang, and Jing Xiao. Dynamic attention-guided context decoding for mitigating context faithfulness hallucinations in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5174–5193, 2025b.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B 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 (volume 1: Long papers)*, pp. 12286–12312, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023c.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

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 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*, pp. 9004–9017, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Jianhui Pang, Fanghua Ye, Derek F Wong, Xin He, Wanshun Chen, and Longyue Wang. Anchor-based large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4958–4976, 2024.

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, 2024.

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*, volume 2025, pp. 50250–50279, 2025.

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.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9840–9855, 2023.

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. *arXiv preprint arXiv:2203.11171*, 2022.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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, 2024.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. Sled: Self logits evolution decoding for improving factuality in large language models. *Advances in Neural Information Processing Systems*, 37:5188–5209, 2024.

Yue Zhang, Leyang Cui, Shuming Shi, et al. Alleviating hallucinations of large language models through induced hallucinations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 8233–8247, 2025a.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025b.

A Extended Results and Analyses

This appendix collects the analyses that the page limit keeps out of the main text. It is referenced from the corresponding sections.

A.1 Unified view of decoding-time interventions

Table 4 places AnswerProbe among existing decoding-time interventions and makes the two blind spots concrete. The input-deletion family (CAD, COIECD, and VCD for images) re-encodes the question, so it is subject to the δ_S^{enc} term of Eq. (3). The layer-contrast family (DoLa, SLED, Activation Decoding) sharpens the model toward knowledge stored in its own layers. When the answer must come from a passage, this targets the wrong evidence. DAGCD targets the context without re-encoding the question. Like every other method in the table, however, it commits to its adjustment token by token, before the answer exists. AnswerProbe is the only method that changes neither the input nor the layers, and the only one that decides after complete answers are available.

The framework also contains simpler decoders as special cases, which makes the role of each component measurable. With the native reading alone, AnswerProbe is greedy decoding. With the single reading $(\mathcal{S}, +1)$, it is a de-confounded CAD, in which the passage is withheld through the answer rows instead of being deleted from the input. With $\Lambda \equiv 1$, it becomes a likelihood reranker that ignores the source. With the readings replaced by temperature samples, it becomes sampling with source-aware reranking. Section 6 evaluates each of these reductions, and each loses a substantial part of the gain.

A.2 Per-domain comparison with the baseline families

Input deletion helps under conflict and hurts elsewhere. CAD and COIECD delete the passage from the reference input. This helps on NQ-Swap, where the passage contradicts memory, but CAD loses up to 5.4 F1 on average (Mistral) and COIECD's conflict detector keeps both the gain and the loss below 0.8. Finding 1 (Section 5) shows why: deleting the passage also re-encodes the question, and the resulting term δ_S^{enc} ranks answers inversely to correctness, most strongly on TriviaQA and SearchQA. Removing this confound is enough to turn the loss into a gain. AnswerProbe restricted to the single reading $(\mathcal{S}, +1)$, which differs from CAD only in withholding the passage through the answer rows, improves on greedy decoding for every model (+1.1, +1.6 and +0.6; Section 6).

Attention guidance avoids the confound but commits too early. DAGCD up-weights context tokens selected by attention and does not re-encode the question. It is the strongest baseline, but it gains only 0.7–2.1 points, and it still trades domains (for instance -0.9 on TriviaQA for Llama-2-

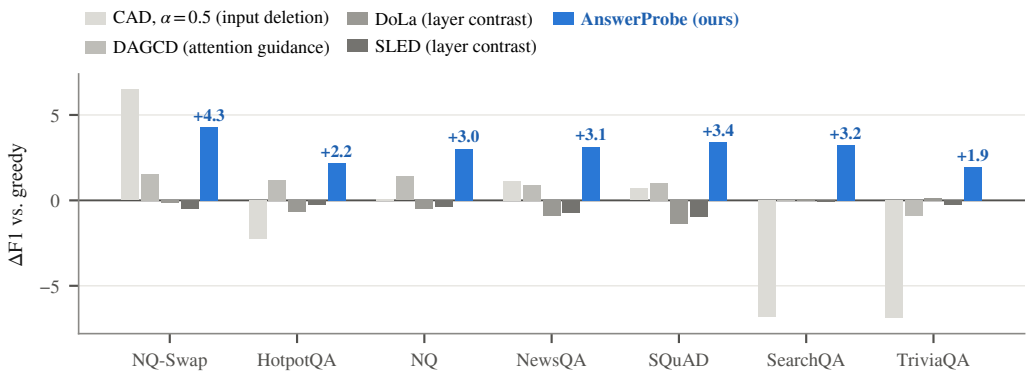


Figure 5: **Every existing family trades domains against each other.** F1 change over greedy decoding per MRQA domain (Llama-2-7B-chat, all 62,316 questions); one representative per family of Table 4 (DAGCD as reported, the others rerun by us). AnswerProbe lets the completed answers decide per question and improves all seven domains.

Table 4: **Existing decoders and ANSWERPROBE in one frame.** Each method modifies the next-token distribution relative to a reading of the prompt: CAD (Shi et al., 2024), COIECD (Yuan et al., 2024), VCD (Leng et al., 2024), DAGCD (Huang et al., 2025b), DoLa (Chuang et al., 2024b), SLED (Zhang et al., 2024), Activation Decoding (Chen et al., 2024b). Columns: what is changed, which evidence it targets, when the decision is taken, and whether the question's representation is altered.

Method	What is changed	Evidence targeted	Decision	Re-encodes question
Input deletion				
CAD	input without the context	context	per token	yes
COIECD	CAD, gated by an entropy test for conflict	context	per token	yes
VCD	input image perturbed by noise	image	per token	yes
Attention guidance				
DAGCD	context tokens up-weighted by attention and uncertainty	context	per token	no
Layer contrast				
DoLa	contrast with a premature layer	parametric	per token	no
SLED	logits evolved toward early-layer estimates	parametric	per token	no
Activation Dec.	sharpness of in-context activations	parametric	per token	no
ANSWERPROBE	answer rows blocked from a region $\mathcal{R}$	context	**per answer**	no

Table 5: **Interventional support improves TruthfulQA MC1 for every model.** All methods score the same candidates with the same prompt. ANSWERPROBE uses Eq. (7) with $\mathcal{S}$ = instruction and demonstrations, fixed on Llama-2-7B and frozen for the other models. **Bold**: best, underline: second best per protocol; DoLa, Activation Decoding and MC2 are in Appendix G.

MC1	Llama-2-7B	Llama-2-13B	Mistral-7B	Llama-3-8B	Qwen3-1.7B	Avg.
summed token scores (standard protocol)						
Native	33.8	35.4	39.7	40.8	31.3	36.2
SLED	39.2	**37.8**	38.6	42.1	**35.5**	38.6
ANSWERPROBE (ours)	**39.4**	37.3	**40.0**	**44.3**	33.8	**39.0**
length-normalised token scores						
Native	36.7	39.3	42.5	43.3	31.7	38.7
SLED	**44.8**	**44.7**	**45.5**	**54.6**	**42.7**	**46.5**
ANSWERPROBE (ours)	41.9	42.1	42.7	48.1	36.4	42.2

7B). Like the other baselines, it adjusts each token before the answer exists; Finding 2 shows that at the first token, the better and worse readings are nearly indistinguishable.

Layer contrasts target the wrong source. DoLa, SLED and Activation Decoding sharpen the model toward knowledge stored in its own layers. With the defaults of their releases, they change the average by at most $+0.4$, and on Llama-2-13B DoLa and SLED lose 4.3 and 7.9 points. In context-grounded QA the missing evidence is in the passage, and amplifying stored knowledge can conflict with it: on NQ-Swap they never gain more than 0.9 points, and they lose up to 7.9. ANSWERPROBE intervenes on the channel that carries the passage to the answer, and gains most exactly where that channel matters.

A.3 CLOSED-BOOK QA

Closed-book QA tests the interventional support of Eq. (7). There is no passage to search, and the "source" is the truthfulness-inducing context: the instruction and the six demonstrations. Table 5 shows that it improves MC1 over the native model for every model under both scoring protocols, by +3.5 on average under length-normalised scores. Under the standard summed protocol it is the best method on average. MC2 rises by 5–16 points (Appendix G).

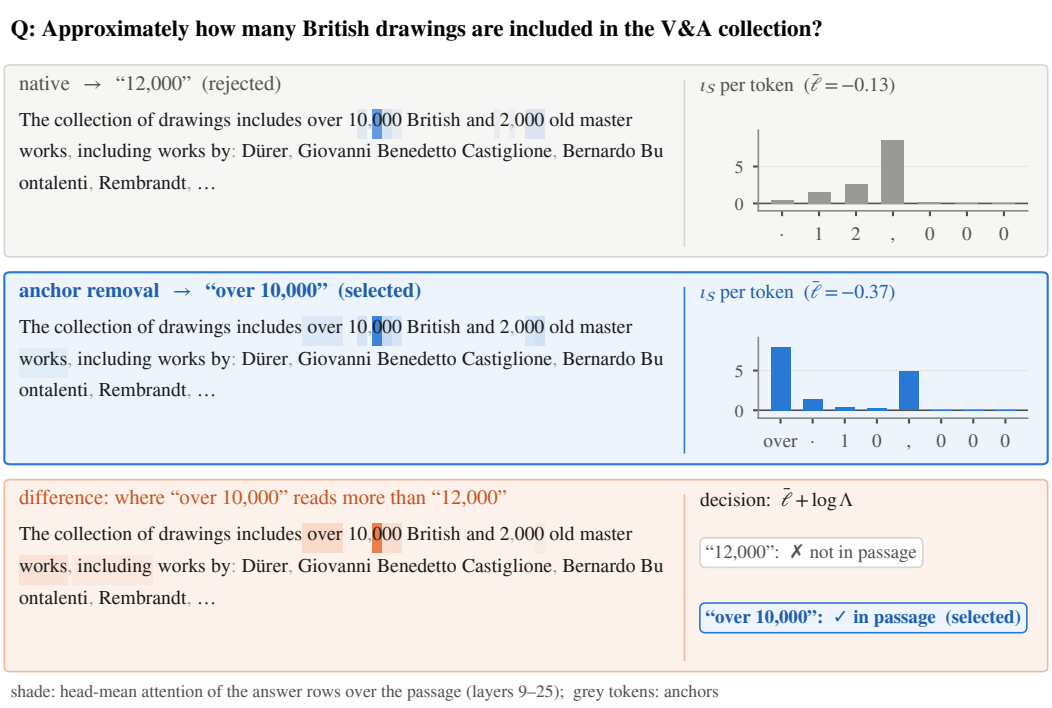


Figure 6: **Anatomy of one decision** (SQuAD; the passage, shown truncated, lists "over 10,000 British and 2,000 old master works"). *Left:* where the answer rows of each distinct proposal attend. *Right:* per-token source information ι_S with the native likelihood $\bar{\ell}$, and the decision. The native answer "12,000" reads both numbers and merges them; the anchor-removal reading proposes "over 10,000", which reads "over 10,000" itself. The native model prefers "12,000" ($\bar{\ell} = -0.13$ vs. -0.37); containment reverses this preference.

Why source information separates truthful answers. TruthfulQA's false candidates are imitative falsehoods: fluent, popular completions of the question. Their likelihood hardly depends on whether the answer can read the instruction to avoid myths, or the demonstrations of cautious answering; they are likely in any case. Truthful candidates are often the less typical, hedged or corrective ones, and their likelihood depends on that context. ι_S measures exactly this dependence, with the question's representation held fixed. The input-deletion analogue, which contrasts against the question read without instruction and demonstrations, *lowers* MC1 on Llama-2-7B (36.7 → 35.6). This is the closed-book face of blind spot 1: deleting the context also changes how the question is read. Blocking only the answer's view of the demonstrations gives 40.9 and of the instruction 38.3; blocking the question's own anchors has no effect (Appendix G). The gain therefore comes from the answer's reliance on the source, not from perturbing the question.

Relation to layer contrasts. SLED, which refines the model's belief using its intermediate layers, is stronger under length normalisation. The two methods act on different factors of Eq. (6): SLED sharpens the prior $\bar{\ell}$, while ANSWERPROBE adds the support Λ. The closed-book setting is thus not the main arena of our method. It shows that the answer-row intervention is a meaningful measurement of source reliance even when there is nothing to search.

A.4 MECHANISM FINDINGS 3–5 IN FULL

Finding 3. *Hallucinated answers do not read the passage less; they read it through compressed channels.* We block the answer rows' access to the passage in one band of four layers at a time, or to all anchors, and measure the change in the greedy answer's mean token log-likelihood $\bar{\ell}$ on 700 questions (Figure 3c). Blocking the passage in any single band costs at most 0.9 nats per token, but blocking it in all layers costs 3.4–3.8. The answer reads its source redundantly across depth, which is why the intervention acts on all layers. Comparing wrong with correct greedy answers reveals a qualitative difference. Overall, wrong answers depend on the passage about as much as

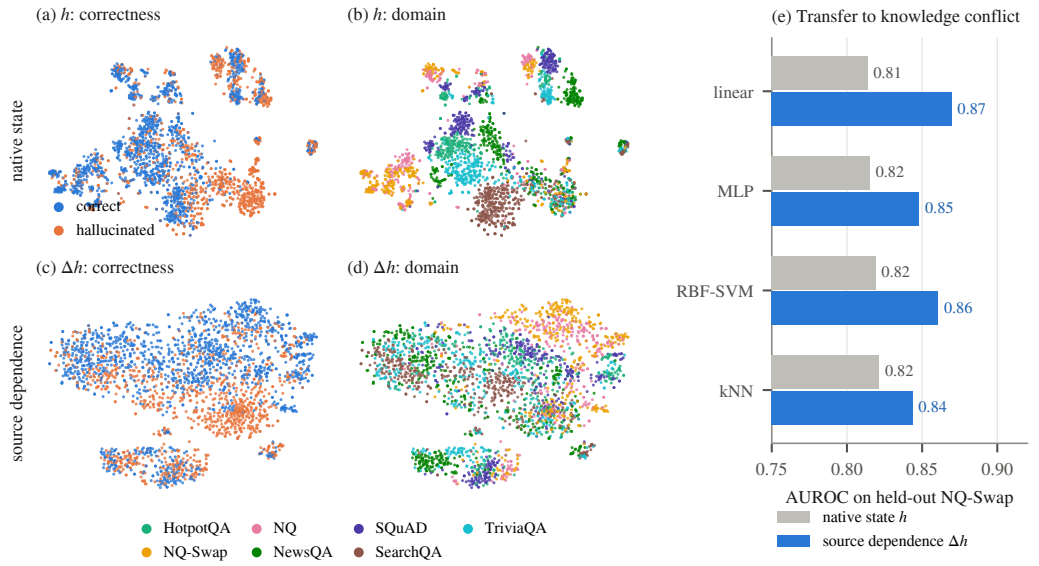


Figure 7: **Source dependence organises answers by how they use the passage, and transfers to knowledge conflict.** (a)–(d) t-SNE of the final-layer state of the last answer token for 2,500 proposals with F1 = 1 or 0 (Llama-2-7B-chat, fresh questions): the native state $h(y)$ (top) and its source dependence $\Delta h(y) = h(y) - h^{\neg \mathcal{S}}(y)$ under the answer-row intervention (bottom). (e) Correctness probes trained on the six other domains and tested on NQ-Swap, for four classifiers.

correct ones (3.42 vs. 3.76 nats), but they depend *more* on the passage in the first four layers (0.87 vs. 0.31) and on the anchor tokens (0.34 vs. 0.10). From layer 5 upward, anchors absorb 21–35% of the answer rows' attention, more than the passage's content tokens (9–16%; Appendix L). Yet blocking them barely moves a correct answer. Anchors are a high-traffic, low-information channel. Correct answers retrieve their evidence from content tokens in deeper layers; hallucinated answers lean on shallow reading and on the aggregated summary. Figure 6 shows the pattern on one question. Asked for the number of British drawings, the native answer attends to both numbers in the sentence and returns their sum, 12,000: the evidence is read, but in a merged, summarised form. With the anchors blocked, the answer reads the content tokens directly and returns "over 10,000". *This is the mechanism behind the anchor-removal reading. Closing the compressed channel forces direct reading, which is worse on average but correct precisely where the compressed channel misled the model.* It also supports, in a qualified form, the hypothesis that aggregation into anchors contributes to hallucination (Huang et al., 2024).

Finding 4. *Interventional support is a graded version of exact support.* Figure 3d shows $\iota_{\mathcal{S}}$ for all proposals. Proposals contained in the passage concentrate at positive values (median $+2.8$ nats per token), uncontained ones at negative values (median -2.0). Correct proposals have median $+3.1$ and wrong ones -1.1. On NQ-Swap, greedy answers that follow memory instead of the counterfactual passage have median $\iota_{\mathcal{S}} = 1.8$, against 5.2 for faithful ones. *This justifies Eq.* (7)*: when there is no text to search, the answer-row intervention recovers the same signal that containment provides when there is.*

Finding 5. *Source dependence encodes whether the passage supports an answer, not what the answer is about, and this transfers to knowledge conflict.* We take the final-layer state $h(y)$ of each proposal's last token and its change $\Delta h(y) = h(y) - h^{\neg \mathcal{S}}(y)$ when the answer rows cannot read the passage (Figure 7). The native states cluster by domain, and hallucinations are scattered inside every cluster. In the source-dependence space the domain structure dissolves, and hallucinations from all domains gather in one region. Among the 10 nearest neighbours of each proposal (from other questions), the share from the same domain falls from 0.70 to 0.52 (chance 0.15), while the share with the same correctness stays at 0.73 versus 0.69. Both spaces separate correct from hallucinated answers about equally well within the training distribution, with linear or non-linear probes alike (AUROC 0.85–0.92; Appendix I). The difference appears under knowledge conflict. Probes trained

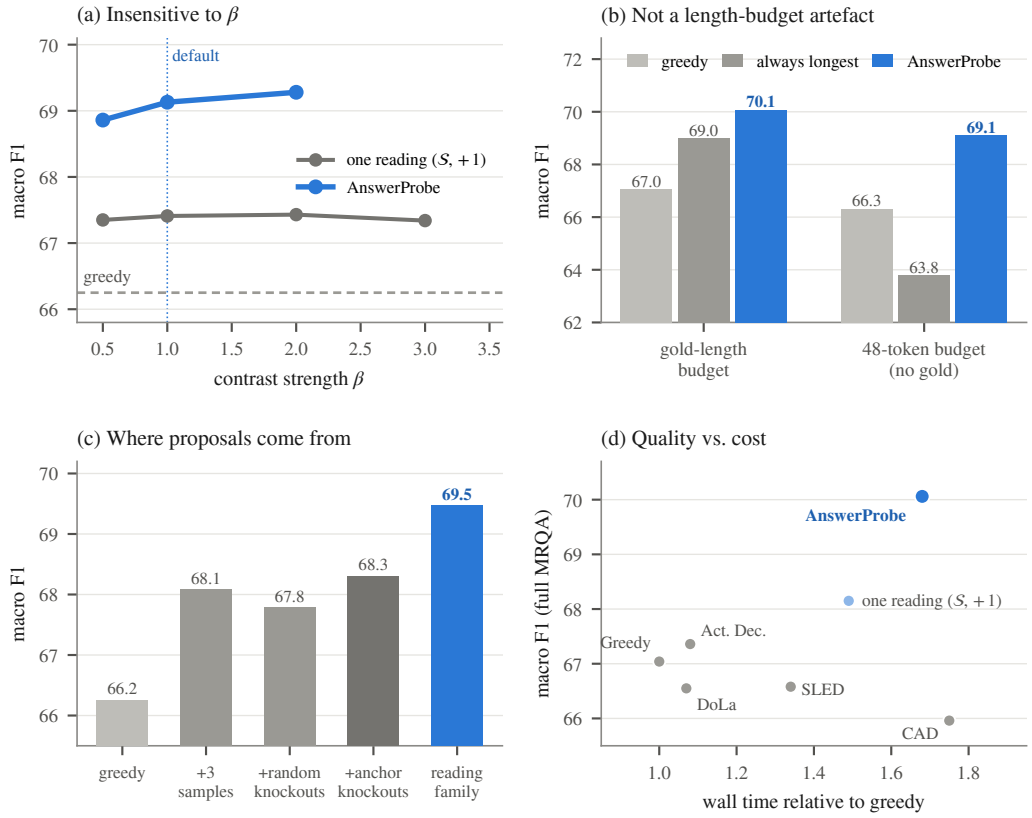


Figure 8: **The gain is robust, comes from structured readings, and is cheap.** (a)–(c): fresh 7,000-question panel (Llama-2-7B-chat), disjoint from development. (a) Contrast strength β of both contrastive readings. (b) With and without the gold-length generation budget. (c) Candidate pools of equal size under the same selection rule: greedy plus three temperature samples, plus random-token or anchor knockouts, or structured readings (native, anchor contrast, input deletion, anchor removal). (d) Macro F1 on the full MRQA set against wall time relative to greedy decoding (same GPU and questions, batch of one question).

on the six other domains and tested on NQ-Swap, whose passages contradict memory, transfer better from Δh for every classifier and layer (8 of 8 settings, by 0.02–0.07 AUROC; Figure 7e). The native state partly encodes how familiar an answer is, and in a counterfactual passage the familiar answer is the wrong one; source dependence encodes whether the passage supports the answer, which does not change with the topic or with conflict. *This is the representational counterpart of $\iota_{\mathcal{S}}$, and it explains why the answer-row intervention is most useful on NQ-Swap (Table 1).*

A.5 Robustness, cost and composability

Robustness (Figure 8). (a) The result does not hinge on the contrast strength. The single reading $(\mathcal{S}, +1)$ varies by less than 0.1 F1 over $\beta \in [0.5, 3]$, and AnswerProbe by 0.4 over $\beta \in [0.5, 2]$, so the contrastive-decoding default $\beta = 1$ is not a tuned optimum. (b) The gold-length budget of the DAGCD protocol rewards long outputs: always taking the longest proposal gains 1.9 F1 under it but loses 2.5 without it. AnswerProbe selects answers that are slightly *shorter* than greedy's in both settings and gains in both (+3.0 and +2.8). Its gain is therefore not a length effect. (c) At equal pool size, temperature samples and random-token knockouts add diversity without addressing any failure mode. Structured readings change *how the answer reads the evidence*, and they give the largest gain. Anchor knockouts beat random-token knockouts of the same size, which confirms that anchors are a meaningful region (on TruthfulQA the same comparison favours anchors on all five models).

Cost, composability and general abilities (Figure 8d; Appendix F). The four readings share one prefill and decode as rows of one batch, so ANSWERPROBE costs $1.68\times$ greedy wall time, less than CAD ($1.75\times$), which must prefill a second prompt. Answer-level verification also wraps other decoders: pairing DoLa, SLED or Activation Decoding with greedy decoding and applying Eq. (6) lifts each by 0.7–0.9 F1, but adding their outputs to the reading family changes almost nothing, so the readings are the right proposals. Finally, the method leaves MMLU and ARC-Challenge within ± 0.3 points, and on GSM8K the passage-contrast reading, applied to the problem statement, helps, with derivations serving as the source for computed answers (+3.2).

B IMPLEMENTATION DETAILS

Answer-row intervention. We replace the eager attention function of each layer. For a blocked region $\mathcal{R}$ we add a bias $b_j = -\infty$ for keys $j \in \mathcal{R}$ (0 otherwise) to the pre-softmax attention logits of query positions $i \geq n$ (the answer rows $\mathcal{T}$, which start at the last prompt position), in all layers and heads. This is equivalent to zeroing those attention weights and renormalising. Native rows receive a zero bias and are bit-identical to the unmodified model.

Regions. *Anchors* are prompt tokens before the final answer cue whose decoded text is empty after stripping whitespace, or consists only of punctuation characters (`.,;:!?"'()[]{}<>/\|-` and typographic quotes and dashes). BOS is excluded. On MRQA, anchors are 21.6% of prompt tokens (135 of 625 on average). The *passage* region is every token whose characters lie inside the passage slot of the prompt template (93% of prompt tokens). On TruthfulQA, the source is every token of the instruction sentence and the six demonstrations; the final "Q: ⟨question⟩" line and the answer cue are excluded.

Readings. The passage- and anchor-contrast readings use Eq. (2) with $\beta = 1$ and $\tau = 0.1$. The anchor-removal reading is greedy on $p_\theta^{\neg\mathcal{A}}$. All readings stop at EOS or at the budget.

Selection. Proposals are deduplicated by token sequence. $\bar{\ell}$ is the mean log-probability under the native model of the tokens actually generated. Containment uses the SQuAD normalisation (lowercasing, removing punctuation and the articles *a/an/the*, collapsing whitespace) and requires a contiguous match on word boundaries. Contained proposals are ranked above uncontained ones, and $\bar{\ell}$ ranks proposals within each group (Eq. (6) with $\epsilon \to 0$).

Baselines. DoLa uses the default premature layers $\{16, 18, \ldots, 30\}$ and mature layer 32 (for 32-layer models) with relative-top filtering 0.1. Activation Decoding uses $\alpha = 0.5$ and information layer 26. SLED uses evolution rate 2.5 and scale 75 for generation, and 2.0 and 50 for multiple choice. These are the defaults of the respective releases. All run with a KV cache under exactly our prompts; through the same code path, greedy reproduces the saved DAGCD greedy outputs on every checked item.

Algorithm 1 ANSWERPROBE (open-book)

Require: prompt x with passage positions $\mathcal{S}$, anchor positions $\mathcal{A}$; budget B

1: Prefill $x_{1:n-1}$ once; repeat the cache over rows for: native; passage contrast (native, $\neg\mathcal{S}$); anchor contrast (native, $\neg\mathcal{A}$); anchor removal (native, $\neg\mathcal{A}$)
2: **for** $t = 1, \ldots, B$ **do**
3: For each active reading, choose its token by its rule (greedy on p_θ, Eq. (2) with $\mathcal{R}{=}\mathcal{S}$ or $\mathcal{R}{=}\mathcal{A}$, or greedy on $p_\theta^{\neg\mathcal{A}}$); record $\log p_\theta$ of the chosen token from the native row
4: Advance all rows in one forward pass (answer rows blocked from their region by a $-\infty$ bias)
5: **end for**
6: **return** $\arg\max_{y \in \mathcal{Y}} \big(\mathbf{1}[\mathrm{norm}(y) \sqsubseteq \mathrm{norm}(x_\mathcal{S})],\ \bar{\ell}(y \mid x)\big)$ (lexicographic)

Compute. All experiments ran on NVIDIA 48GB GPUs. One full MRQA pass of batched ANSWERPROBE on Llama-2-7B takes about 45 minutes on 8 GPUs.

C NOTATION

Inputs and positions	
$x = (x_1, \ldots, x_n)$, $[n]$	prompt and its positions; x_n ends the answer cue and is the first answer row
$\mathcal{S} \subseteq [n]$	source region: passage (open-book) or instruction and demonstrations (closed-book)
$\mathcal{A} \subseteq [n]$	anchors: prompt tokens that are punctuation or whitespace (BOS excluded)
$\mathcal{R}$	a generic region, $\mathcal{R} \in \{\mathcal{S}, \mathcal{A}\}$
$x_{\mathcal{S}}$	the source text (prompt tokens at positions $\mathcal{S}$)
$y = (y_1, \ldots, y_m)$, $y^\star$	an answer (y_t predicted at position $n + t - 1$); the correct answer
$\mathcal{T} = \{n, \ldots, n + m - 1\}$	answer rows: positions that emit answer tokens
$x_{\setminus \mathcal{S}}$	prompt with the source deleted (CAD's reference input)
B	generation budget (token length of the longest gold alias)
Models and readings	
p_θ	unmodified model
l, k	layer ($1, \ldots, L$) and attention head
$a_{ij}^{(l,k)}$	pre-softmax attention logit of query i on key j (layer l, head k)
$h_i^{(l)}$	hidden state of position i after layer l ($h_i^{(0)}$: embedding)
$p_\theta^{\neg\mathcal{R}}$	answer-row intervention: $a_{ij}^{(l,k)} \leftarrow -\infty$ for $i \in \mathcal{T}$, $j \in \mathcal{R}$, all l, k (Eq. (4))
$p_\theta^{(\mathcal{R},\beta)}$	reading: $\log p_\theta^{(\mathcal{R},\beta)} = (1+\beta)\log p_\theta - \beta \log p_\theta^{\neg\mathcal{R}} + \text{const}$ (Eq. (5)); $\beta = 0$ native, $+1$ contrast, -1 removal
$\mathcal{F}$	reading family $\{(\cdot, 0), (\mathcal{S}, +1), (\mathcal{A}, +1), (\mathcal{A}, -1)\}$
q_t, β, τ, $\mathcal{V}_t$	reference next-token distribution, contrast strength (1), plausibility threshold (0.1), plausible set
Scores	
$\bar{\ell}(y \mid x)$, $\bar{\ell}^{\neg\mathcal{R}}(y \mid x)$	mean token log-likelihood of y under p_θ and under $p_\theta^{\neg\mathcal{R}}$ (Eq. (1))
$\iota_{\mathcal{S}}(y)$	source information $\bar{\ell}(y \mid x) - \bar{\ell}^{\neg\mathcal{S}}(y \mid x)$
$\delta_{\mathcal{S}}^{\text{enc}}(y)$	question re-encoding $\bar{\ell}^{\neg\mathcal{S}}(y \mid x) - \bar{\ell}(y \mid x_{\setminus\mathcal{S}})$
$\mathcal{Y}$, $\hat{y}$	set of distinct proposals from the reading family; the selected proposal
$\pi(y \mid x)$	prior belief, $\log \pi = \bar{\ell} + \text{const}$ (Eq. (6))
$\Lambda(y; x_{\mathcal{S}})$	source support: 1 if $\text{norm}(y) \sqsubseteq \text{norm}(x_{\mathcal{S}})$, else $\epsilon \to 0$ (text); $\exp(\lambda\, \iota_{\mathcal{S}}(y))$ (non-text)
λ	weight of interventional support (1; Eq. (7))
$h(y)$, $\Delta h(y)$	final-layer state of the last answer token; its source dependence $h(y) - h^{\neg\mathcal{S}}(y)$
norm, $\sqsubseteq$	SQuAD normalisation; contiguous word-subsequence relation

D DERIVATIONS

Throughout, $p_\theta(v)$ abbreviates $p_\theta(v \mid x, y_{<t})$, and analogously for the other distributions.

D1. Contrastive decoding as a product of experts. For every v,

$$(1+\beta)\log p_\theta(v) - \beta \log q(v) = \log p_\theta(v) + \beta \log \frac{p_\theta(v)}{q(v)} = \log \Big[p_\theta(v) \Big(\frac{p_\theta(v)}{q(v)} \Big)^{\beta} \Big].$$

Greedy contrastive decoding (Li et al., 2023b) therefore returns the mode of Eq. (2) on $\mathcal{V}_t$; the normaliser does not affect the argmax.

D2. The three members of the reading family. Setting $q = p_\theta^{\neg\mathcal{R}}$ in D1 gives Eq. (5), whose log form is $(1+\beta)\log p_\theta - \beta \log p_\theta^{\neg\mathcal{R}}$ up to the normaliser. For $\beta = 0$ it is $\log p_\theta$. For $\beta = -1$ it is $\log p_\theta^{\neg\mathcal{R}}$, so $p_\theta^{(\mathcal{R},-1)} = p_\theta^{\neg\mathcal{R}}$: removing $\mathcal{R}$ is the contrast run backwards, with no further approximation. For $\beta = +1$ it is $2\log p_\theta - \log p_\theta^{\neg\mathcal{R}} = \log p_\theta + \log(p_\theta / p_\theta^{\neg\mathcal{R}})$, which adds the log evidence ratio to the native log-probability once more. Removal and contrast are thus symmetric points of one family rather than two separate heuristics.

D3. Exact decomposition of the CAD evidence (Eq. (3)). For every token, $\log p_\theta(v \mid x) - \log p_\theta(v \mid x_{\setminus\mathcal{S}}) = [\log p_\theta(v \mid x) - \log p_\theta^{\neg\mathcal{S}}(v \mid x)] + [\log p_\theta^{\neg\mathcal{S}}(v \mid x) - \log p_\theta(v \mid x_{\setminus\mathcal{S}})]$. Averaging over the tokens of y, each conditioned on the same $y_{<t}$ in all three distributions, gives Eq. (3). The

identity is exact for every y. Its content lies in the meaning of the two terms, and in the empirical finding that $\delta^{\text{enc}}_{\mathcal{S}}$ ranks answers inversely to correctness (Section 5).

D4. Prompt states are invariant under answer-row interventions. Let $h_i^{(l)}$ be the state of position i after layer l, with $h_i^{(0)}$ the embedding. With causal attention, $h_i^{(l)}$ depends only on $\{h_j^{(l-1)}\}_{j\le i}$. Eq. (4) modifies logits only for queries $i \ge n$. Hence, by induction over l, $h_i^{(l)}$ is unchanged for every $i < n$ and every l. Therefore (i) the question's representation is identical under p_θ and every $p_\theta^{\neg\mathcal{R}}$, so the analogue of $\delta^{\text{enc}}_{\mathcal{S}}$ between any two answer-row interventions is identically zero; and (ii) the key–value cache of positions $1, \dots, n-1$ is shared by all readings.

D5. The additive bias equals zero-and-renormalise. For a query with logits a_j and blocked set $\mathcal{R}$, let $\tilde{a}_j = -\infty$ for $j \in \mathcal{R}$ and $\tilde{a}_j = a_j$ otherwise. Then $\mathrm{softmax}(\tilde{a})_j = \mathbf{1}[j \notin \mathcal{R}]\, e^{a_j} / \sum_{j' \notin \mathcal{R}} e^{a_{j'}}$. This equals the native attention weights with the $\mathcal{R}$ entries set to zero and renormalised. A zero bias leaves native rows bitwise unchanged.

D6. Selection as posterior inference. Write the posterior over the proposals as $\Pr(y^\star = y \mid x) \propto \pi(y \mid x)\, \Lambda(y; x_{\mathcal{S}})$. The prior is the native model's length-normalised belief, $\log \pi(y \mid x) = \bar{\ell}(y \mid x)$ up to a constant. The likelihood Λ is the probability of the observed source text $x_{\mathcal{S}}$ if y were correct. Under 0–1 loss the Bayes decision is the posterior mode on $\mathcal{Y}$, which is Eq. (6). *Open-book QA.* Extractive datasets such as MRQA and NQ-Swap generate gold answers as passage spans, so Λ is 1 for contained proposals and a vanishing floor $\epsilon \to 0$ otherwise; the decision is the most plausible contained proposal. If no proposal is contained, all likelihoods equal ϵ and the prior alone decides. *Closed-book QA.* There is no searchable source. We approximate $\log \Lambda$ by the interventional log likelihood ratio $\lambda\, \iota_{\mathcal{S}}(y)$: the answer is more likely to be supported by the source to the extent that its probability depends on reading the source. The two modelling choices, the length-normalised prior and the plug-in support, are assumptions, not theorems. We test them empirically. Within questions, $\bar{\ell}$ ranks the better of two proposals higher in 65% of pairs, and the full posterior does so in 75%; $\iota_{\mathcal{S}}$ separates contained from uncontained proposals (Figure 3d).

E Protocol and Statistics

All design choices (the reading set, β, τ, the selection rule, and containment on SQuAD-normalised words) were fixed on a 1,050-question development panel (150 per domain). Protocols with pass criteria were written before each full-set, cross-model and closed-book confirmation run. The full-set gain excluding development questions is identical to within 0.03 F1.

F Additional Results

Cost (Table 6). AnswerProbe runs four readings yet is cheaper than CAD. CAD must prefill a second prompt, the question without the passage, whereas all answer-row interventions share one prefill (Appendix D, D4) and decode as rows of one batch, which costs little at batch size one because decoding is memory-bound. The interventional view therefore buys a de-confounded contrast and a

Table 6: **AnswerProbe costs no more than CAD.** Wall time per question on the same GPU and the same 300 questions (Llama-2-7B-chat, batch of one question, identical eager-attention implementation for all methods), and macro F1 on the full MRQA set. CAD's F1 is our rerun with $\alpha = 0.5$.

Method	ms / question	× greedy	Macro F1
Greedy	180	1.00	67.04
DoLa	193	1.07	66.55
Activation Decoding	195	1.08	67.36
SLED	242	1.34	66.58
CAD	316	1.75	65.96
single reading $(\mathcal{S}, +1)$	268	1.49	<u>68.15</u>
AnswerProbe (ours)	303	1.68	**70.06**

Table 7: **Answer-level verification wraps any decoder, but the readings supply the best proposals.** Full MRQA set (Llama-2-7B-chat). Each decoder's output is added to a candidate pool and the pool is decided by Eq. (6).

Decoder	alone	+ greedy, verified	+ reading family, verified
DoLa	66.55	67.47 ↑0.92	70.04
Activation Decoding	67.36	68.15 ↑0.79	**70.21**
SLED	66.58	67.27 ↑0.69	70.06
ANSWERPROBE (reading family only)	–	–	70.06

Table 8: **General abilities are preserved, and reading the problem helps reasoning.** Llama-2-7B-chat. MMLU (5-shot, 14,042 questions) and ARC-Challenge (5-shot, 1,172): candidates are option letters, scored as $\bar{\ell} + \iota_{\mathcal{S}}$ with $\mathcal{S}$ = the demonstrations. GSM8K (8-shot chain of thought, 1,319): the four readings generate chains with $\mathcal{S}$ = the problem statement; for computed answers the source is the derivations, and the support is their product of experts (see text).

Method	MMLU	ARC-C	GSM8K
Native (greedy)	**47.19**	59.47	21.00
single reading $(\mathcal{S}, +1)$	–	–	22.29
ANSWERPROBE	46.89	**59.73**	**24.18**

whole family of readings for less than the price of the input-deletion contrast. Decoding the readings one after another would cost $4.8\times$.

Verification is a general wrapper, and the readings are the right proposals (Table 7). Adding just greedy decoding and applying Eq. (6) lifts every layer-contrast decoder by 0.7–0.9 F1: verification repairs some of their unsupported answers. Adding their outputs to the reading family, however, changes almost nothing (70.04–70.21 vs. 70.06). Their proposals are largely redundant with the readings, while the readings' proposals are not redundant with theirs. This is the empirical counterpart of the argument in Section 3.1: to fix context-grounded errors, one should perturb how the answer reads the passage, not how the model uses its layers.

General abilities (Table 8). A decoding method for faithfulness should not damage what the model can already do (Zhang et al., 2025a). On knowledge and science multiple choice, ANSWERPROBE changes accuracy by -0.3 (MMLU) and $+0.3$ (ARC-Challenge): scoring an option by its reliance on the demonstrations neither helps nor hurts when the demonstrations only fix the format. On GSM8K, the results clarify what the method's "source" is. Requiring the chain to rely on the problem statement (the passage-contrast reading) improves accuracy by 1.3 points: some arithmetic errors are misreadings of the problem, the same failure the method targets. Selecting among chains by $\bar{\ell} + \iota_{\mathcal{S}}$, however, falls below greedy decoding. Probing the bare final answer does too: for a computed answer, neither the problem statement nor the model's one-step belief can verify it. What can support a computed answer is a derivation. Taking the readings' derivations as the source, with support $\Lambda(y; c_1, \ldots, c_4) = \prod_{r=1}^{4} p_\theta(y \mid x, c_r)$, where x is the prompt with the problem and c_r is the derivation of reading r, cut before its conclusion, gives 24.18 (+3.2). This form reduces to majority voting when derivations are conclusive, which explains self-consistency (Wang et al., 2022) as a special case of source-supported selection. Answer-level verification therefore applies whenever the answer's source can be identified: a passage, a demonstrating context, or a set of derivations.

G CLOSED-BOOK DETAILS

SLED's multiple-choice scores are evolved logits, bounded at -100 outside the top-k tokens, not log-probabilities. Adding our support term to them changes MC1 by -0.4 (Llama-2-7B), $+2.7$ (Mistral) and -3.2 (Qwen3) under mean scoring. We therefore report the two methods separately. Anchors versus an equal number of random content tokens in the source region, MC1 difference: Llama-2-7B $+1.35$, Llama-2-13B $+0.12$, Mistral $+0.37$, Llama-3-8B $+1.22$, Qwen3 $+0.98$; pooled over the five models, $+0.81$.

Per-region analysis. On TruthfulQA we blocked the answer's view of each prompt region separately (Llama-2-7B, MC1, $\lambda = 1$). Blocking the question's anchors has no effect (36.7), and block-

Table 9: TruthfulQA MC1 (MC2) for the remaining baselines, same candidates and prompt as Table 5.

Scoring	Method	Llama-2-7B	Llama-2-13B	Mistral-7B	Llama-3-8B	Qwen3-1.7B
sum	DoLa	32.2 (60.3)	31.1 (63.1)	35.7 (64.5)	35.9 (65.0)	27.2 (59.8)
sum	Activation Dec.	34.0 (51.5)	34.3 (53.6)	30.4 (55.4)	34.9 (58.4)	31.3 (48.9)
mean	DoLa	35.6 (45.9)	28.2 (46.8)	42.4 (48.6)	43.2 (50.7)	17.6 (50.6)
mean	Activation Dec.	37.1 (47.1)	37.8 (48.0)	42.5 (48.8)	44.2 (51.6)	32.0 (45.1)

ing the whole question is harmful (33.1). Blocking the instruction (38.3), the demonstrations (40.9), or both (41.9) helps. Blocking only the anchors of instruction and demonstrations, a quarter of their tokens, already gives 39.3, about half of the effect. The benefit plateaus for $\lambda \in [0.5, 3]$ (41.5–41.9). Answers that remain likely when the answer cannot read the truthfulness-inducing context are the ones this score demotes.

H Replication of the Mechanism Findings Across Models

Table 10: **Findings 1 and 2 replicate across models.** Within-question pairwise accuracy of ranking the better of two proposals (F1 differing by ≥ 0.5), macro-averaged over domains. Finding 1: the same 2,100 fresh questions for each model, with each term of Eq. (3). Finding 2: all 62,316 questions.

	Finding 1: terms of the CAD evidence			Finding 2: prefix vs. answer		
Model	ι_S	CAD evidence	δ_S^{enc}	first token	$\bar{\ell}$	posterior
Llama-2-7B-chat	**0.87**	0.74	0.20	0.55	0.65	**0.75**
Mistral-7B-Instruct	**0.79**	0.72	0.37	0.53	0.64	**0.77**
Llama-2-13B-chat	**0.82**	0.74	0.28	0.56	0.66	**0.80**

On all three models, the re-encoding term inside CAD's evidence ranks answers below chance, and the source information isolated by the answer-row intervention ranks them better than CAD's evidence does. CAD is again at or below chance on SearchQA (0.55 and 0.47) and TriviaQA (0.46 and 0.53) for Mistral and Llama-2-13B, the domains where input deletion loses most in Table 1. On all three models the first token barely separates the better from the worse proposal, and only 17–18% of such pairs share their first token, whereas the completed answers are separated well by the posterior.

I Representation View of Source Dependence: Details

Table 11: **Probes on answer representations** (Llama-2-7B-chat, 2,100 fresh questions; proposals with F1 = 1 or 0). "All": AUROC with 5-fold cross-validation grouped by question. "NQ-Swap": AUROC on NQ-Swap for a probe trained on the other six domains. Features are standardised and reduced to 64 principal components before each classifier. Layers are numbered from 1; layer 32 is the last.

	layer 16					layer 32				
	All			NQ-Swap		All			NQ-Swap	
Probe	h	Δh	$[h; \Delta h]$	h	Δh	h	Δh	$[h; \Delta h]$	h	Δh
linear	0.902	0.899	0.907	0.786	**0.851**	0.868	0.860	0.880	0.814	**0.870**
MLP	0.907	0.898	0.906	0.830	**0.887**	0.868	0.862	0.880	0.815	**0.848**
RBF-SVM	0.919	0.909	0.919	0.837	**0.879**	0.877	0.871	0.884	0.819	**0.861**
kNN	0.897	0.888	0.899	0.848	**0.887**	0.851	0.845	0.855	0.821	**0.844**

Table 11 gives the full comparison behind Finding 5. On the pooled data, the native state and its source dependence are about equally informative, for linear and non-linear probes alike, and concatenating them adds at most 0.012 AUROC: Δh is not a strictly stronger signal of correctness. Its advantage is robustness. On the held-out knowledge-conflict domain it transfers better for every probe and layer. One further observation frames what ANSWERPROBE does and does not exploit. The native state already encodes correctness well: a supervised probe on question-centred h ranks the better of two proposals of the same question in 85% of pairs, above the native likelihood (64%). ANSWERPROBE reaches 75% without any training, using only the model's likelihood and the source; supervised probes (Kadavath et al., 2022) could be combined with the reading family at the cost of labelled data.

J VISION-LANGUAGE MODELS: WHERE THE SOURCE IS READ

We apply the same intervention to LLaVA-1.5-7B on POPE (Li et al., 2023c), with the source $\mathcal{S}$ being the 576 image tokens. Each of the 9,000 questions asks whether an object is present, so the candidates are *yes* and *no*; the passage-contrast reading scores them by $2\bar{\ell} - \bar{\ell}^{\neg\mathcal{S}}$.

Table 12: POPE with LLaVA-1.5-7B (greedy decoding, 3,000 questions per split). VCD contrasts against the image perturbed by diffusion noise (Leng et al., 2024); deletion contrasts against the prompt without the image.

	Random		Popular		Adversarial		All	
Method	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Native	87.27	85.81	86.17	84.77	**83.87**	82.65	85.77	84.39
VCD	86.87	85.66	84.77	83.74	82.07	81.38	84.57	83.56
Image deletion	87.10	85.60	86.07	84.62	**84.17**	**82.92**	85.78	84.36
ANSWERPROBE: $(\mathcal{S}, +1)$, answer rows block the image	**88.17**	**87.06**	**86.20**	**85.22**	83.37	82.70	**85.91**	**84.96**

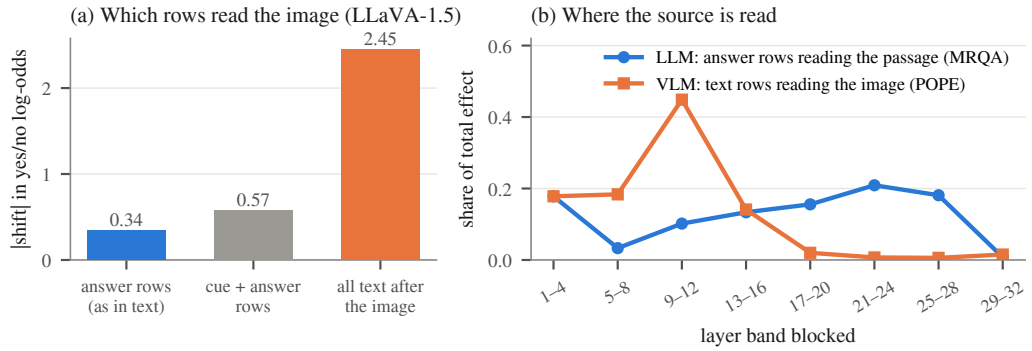


Figure 9: **In LLaVA, the image is read by the question, not by the answer.** (a) Mean change in the yes/no log-odds on 900 adversarial POPE questions when different rows cannot attend to the image. (b) Share of the effect of blocking the source in each band of four layers: for the text model, answer rows blocked from the passage (MRQA, Figure 3c); for the VLM, all text rows after the image blocked from it.

The answer-row intervention is the only rule that improves both accuracy and F1 overall, but the effect is small (Table 12), and VCD, whose reported gains are relative to sampling, falls below greedy decoding. Figure 9 explains why the answer-row intervention is weak here. Blocking the image for the answer rows changes the yes/no log-odds by only 0.34, whereas blocking it for all text after the image changes them by 2.45. The image is read by the question and cue tokens, mostly in layers 9–12, and hardly at all after layer 16. In the text model, by contrast, the answer rows read the passage throughout depth. This agrees with analyses of cross-modal information flow, which find that visual information is transferred to the question tokens in shallow and middle layers (Zhang et al., 2025b). The implication for our method is a precise one. Answer-row interventions probe what the *answer* reads, and so they are effective when the answer reads its source directly,

as in context-grounded text QA. When the source has already been absorbed into the question's representation, the only way to withhold it is to re-encode the question, which is the confound of blind spot 1. Separating visual evidence from question understanding in VLMs therefore requires interventions at the layers where the transfer happens, which we leave to future work.

K CASE STUDIES

Table 13: **How ANSWERPROBE decides.** Each block lists the proposal of every reading, whether the passage contains it, and its native mean log-likelihood $\bar{\ell}$; ▶ marks the selected answer (Llama-2-7B-chat). Successes: the native answer is unsupported (two merged numbers, a crew size completed from memory, a truncated span from the wrong clause, or a reformatted value) and another reading proposes the supported span, even though its $\bar{\ell}$ is lower. Failures: two supported spans where the more plausible one overlaps the gold answer less; and a normalisation mismatch (the passage writes "10–20" with an en dash), which makes the correct native answer look unsupported.

Reading	Proposal	In passage	$\bar{\ell}$
[success] *SQuAD*: Approximately how many British drawings are included in the V&A collection?		gold: *over 10,000*	
native	12,000	✗	−0.13
passage contr.	British drawings: over 10	✗	−0.73
anchor contr.	British drawings 10,0	✗	−0.39
▶ anchor removal	over 10,000	✓	−0.37
[success] *SQuAD*: How many people were on the test flight of the AS-206?		gold: *unmanned*	
native	Two.	✗	−0.89
▶ passage contr.	Unmanned	✓	−0.68
anchor contr.	Two.	✗	−0.89
anchor removal	Two Ex	✗	−1.16
[success] *SearchQA*: Robert Redford played hunter Denys Finch Hatton in this 1985 biopic		gold: *Out of Africa*	
native	Denys Fin	✗	−0.25
▶ passage contr.	Out of Africa	✓	−0.70
anchor contr.	Denys Fin	✗	−0.25
anchor removal	Out of Africa	✓	−0.70
[success] *NQ*: who played sonny in lemony snicket's a series of unfortunate events		gold: *Presley Smith*	
native	Tara Str	✗	−0.11
passage contr.	Tara Str	✗	−0.11
anchor contr.	Tara Str	✗	−0.11
▶ anchor removal	Presley Smith	✓	−0.44
[success] *SQuAD*: How many granal thylakoids can be in each granum?		gold: *anywhere from two to a hundred*	
native	2-100	✗	−0.20
▶ passage contr.	anywhere from two to a hundred	✓	−0.42
anchor contr.	Two to a hundred.	✓	−0.49
anchor removal	2-100	✗	−0.20
[failure] *SQuAD*: How many inhabitants does the Los Angeles area contain?		gold: *12 million*	
native	12 million	✓	−0.28
passage contr.	million How many metropol	✗	−0.86
▶ anchor contr.	over 12 million	✓	−0.24
anchor removal	12 million	✓	−0.28
[failure] *SQuAD*: How many granal thylakoids are usually in each granum?		gold: *10–20*	
native	10-20	✗	−0.19
▶ passage contr.	Most common.	✓	−0.61
anchor contr.	2-100	✗	−0.26
anchor removal	Most common.	✓	−0.61

The successes in Table 13 show the four failure types that completed answers expose: two facts merged into one (12,000 for over 10,000), an answer completed from memory against the passage (Two instead of unmanned), a surface form that differs from the passage (2-100), and an answer from the wrong clause that is also truncated by the budget (Tara Str, Denys Fin). In all but the second, the selected answer has a *lower* native likelihood than the rejected one, so plausibility alone would have failed; the source support is decisive. The failures mark the limits of exact support: it cannot rank two contained spans by overlap with the gold answer, and it inherits the brittleness of string normalisation.

This example complements Figure 6. There, the native reading merged two numbers of the passage and only the anchor-removal reading read the right one. Here, it attends to the wrong clause of a

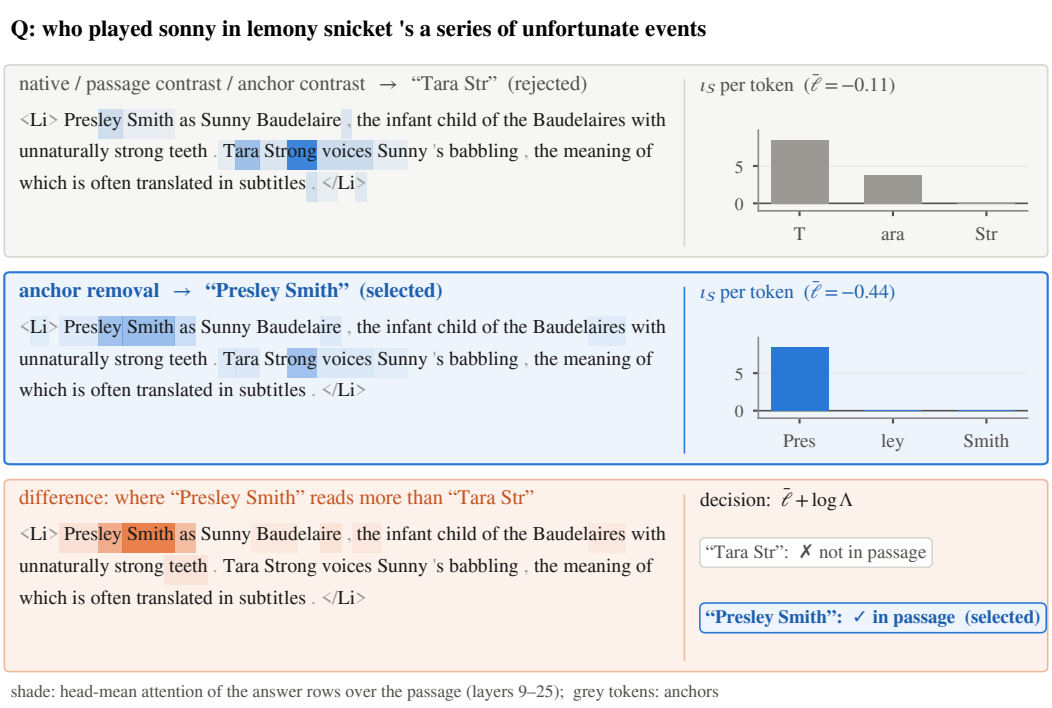


Figure 10: **A rescue by the anchor-removal reading** (NaturalQuestions; "who played Sunny in *A Series of Unfortunate Events*"). The passage names both the actress and the voice actress. Under the native reading, and under the passage- and anchor-contrast readings, the answer rows attend to the voice-actor clause ("Tara Str[ong] voices Sunny's babbling"); the answer is also truncated by the budget, so it is not contained. With anchors blocked, the answer reads the clause that answers the question ("Presley Smith as Sunny"), and containment selects it despite its lower native likelihood.

Table 14: How ANSWERPROBE uses its readings on the full MRQA set (Llama-2-7B-chat). "Changed": the selected answer differs from greedy's. "Unique rescue": greedy has F1 0 and *only* this reading proposes an answer with F1 1 (% of questions).

Domain	Changed (%)	Unique rescue: $(\mathcal{S}, +1)$	$(\mathcal{A}, +1)$	$(\mathcal{A}, -1)$
HotpotQA	10.5	1.0	0.8	1.0
TriviaQA	12.5	0.9	0.5	1.3
SearchQA	13.4	1.9	**1.0**	**2.5**
SQuAD	18.6	0.9	0.9	0.8
NewsQA	**29.1**	0.9	0.7	1.2
NQ	14.7	1.0	0.5	0.6
NQ-Swap	9.8	**2.2**	0.8	1.1

passage that contains both candidates, and only closing the anchor channel redirects it to the clause that answers the question.

Every reading uniquely rescues questions in every domain, and the division of labour follows the failure modes. The passage-contrast reading rescues most on NQ-Swap, where memory contradicts the passage. The anchor-removal reading rescues most on SearchQA and NewsQA, whose long, noisy passages are where compressed reading is least reliable.

L ANSWER-ROW ATTENTION PROFILE

Table 15 lists, layer by layer, how the answer rows of Llama-2-7B-chat distribute their attention over the regions of the prompt. It is the measurement behind Finding 3: anchors carry much of the answer's attention but little of the information it needs.

Table 15: Head-averaged attention of answer rows (gold answers of 700 fresh questions; greedy answers give the same profile to within 0.02) to each prompt region, by layer. Anchors receive more attention than passage content from layer 5 upward, yet blocking them changes a correct answer's $\bar{\ell}$ by only 0.10 nats per token (Section 5).

Layer	BOS	passage content	passage anchors	question	template
1	0.00	0.58	0.16	0.09	0.14
5	0.42	0.05	**0.35**	0.02	0.14
9	0.33	0.15	**0.25**	0.08	0.15
13	0.30	0.13	**0.21**	0.14	0.16
17	0.34	0.12	**0.22**	0.10	0.14
21	0.37	0.16	**0.26**	0.05	0.12
25	0.42	0.15	**0.25**	0.03	0.11
29	0.44	0.09	**0.24**	0.03	0.15