READ THE EVIDENCE BEFORE SEEING THE ANSWER: HYPOTHESIS-INDEPENDENT EVIDENCE READING FOR DETECTING CONFIDENT HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence-based fact verification normally asks a leading question: the candidate answer is placed in the verifier's context and the model is asked whether the evidence supports it. We study the confident, self-consistent errors on which such verification is most needed and least trusted. Auditing 9,762 of them across eight benchmarks shows a benchmark-dependent mixture of genuine hallucinations, ambiguous questions and wrong gold labels; on the genuine part the best internal signal we tried reaches 0.69 AUROC, versus 0.90 or more once the same model reads a passage, so evidence is needed. Candidate-conditioned verification has a measurable cost: holding question, evidence and verifier fixed and varying only whether the candidate is the verifier's *own* output, a model judges its own wrong answer more leniently than another model's equally wrong answer (19,718 matched pairs over five benchmarks, $p < 10^{-5}$), increasingly so with its confidence, and not at all when neither candidate is its own. We therefore make the evidence-interpretation stage hypothesis-independent: a reader answers the question from the evidence before seeing the candidate, $R(x, e) \to \hat{a}$, and an external matcher compares $\hat{a}$ with the candidate, while the calibrated candidate-conditioned score is retained as a base term. This improves detection on every benchmark tested ($+0.016$ AUROC with benchmarks as units, p=0.015), by a similar margin whether or not the verifier produced the answer: what helps is reading the evidence first, not the removal of the ownership bias that motivated it. Routing only the questions that internal signals cannot separate matches the full-retrieval residual error rate at 39% retrieval coverage.

1 INTRODUCTION

Fact verification is commonly formulated as a candidate-conditioned decision: given a question, evidence, and a candidate answer, a verifier determines whether the evidence supports that candidate. FActScore (Min et al., 2023), SAFE (Wei et al., 2024b), RARR (Gao et al., 2023), and standard yes/no grounding prompts follow this pattern, placing the candidate in the same context in which the verifier interprets the evidence. The factored variant of Chain-of-Verification (Dhuliawala et al., 2024) partially separates these steps, but the resulting answers are still evaluated against the original response. We refer to this general property as *hypothesis dependence*: the candidate enters the same computation that interprets the evidence, written as $V(x, a, e)$, where x, a, and e denote the question, candidate answer, and evidence.

This dependence becomes particularly important when the verifier also generated the candidate. We distinguish *self-verification*, where the verifier evaluates its own answer, from *third-party verification*, where the candidate comes from elsewhere. In a controlled comparison that fixes the question, evidence, and verifier while varying only candidate ownership, we find that the verifier is more lenient toward its own wrong answers. The effect is stronger for genuine hallucinations and increases with verifier confidence.

This observation raises a structural question: *should evidence interpretation itself be conditioned on the candidate being evaluated?* We answer this question with ***BEV***, a verification framework, shown in Figure 1, that separates evidence interpretation from hypothesis evaluation. BEV first asks

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.

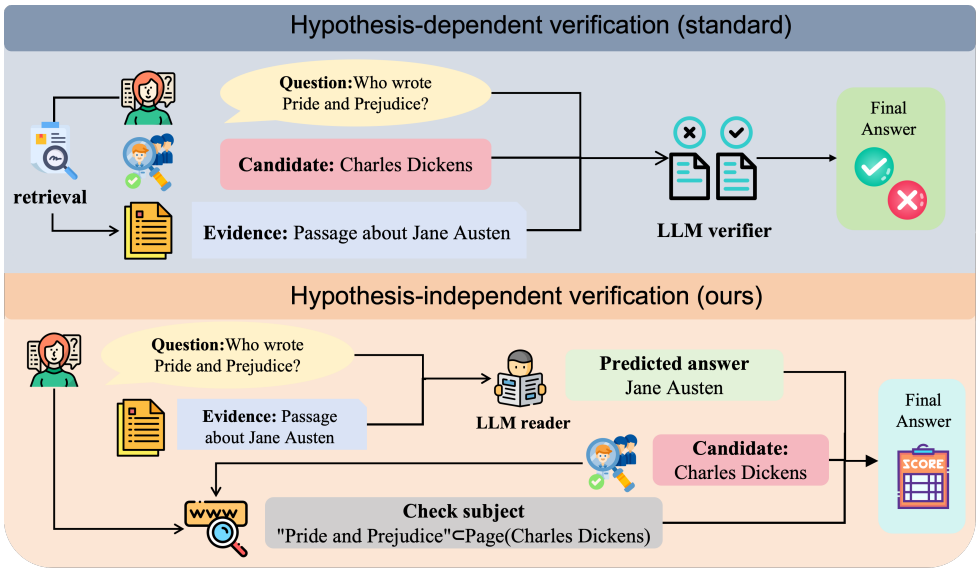


Figure 1: **Separating evidence interpretation from hypothesis evaluation.** *Top*: conventional verification provides the question, evidence, and candidate answer to the same verifier. *Bottom*: the evidence reader receives only the question and evidence and produces $\hat{a}$; the candidate is evaluated afterward by an external matcher. BEV retains conventional hypothesis-dependent verification as a complementary signal while introducing a candidate-independent evidence-reading signal.

an evidence reader to answer the question without observing the candidate, $R(x, e) \to \hat{a}$, and then compares the evidence-derived answer with the candidate outside the evidence-reading computation via $\mathrm{match}(\hat{a}, a)$.

Contributions. *(1)* We identify and characterise an ownership effect in self-verification through controlled comparisons of candidate ownership: the verifier is more lenient toward its own wrong answers, with the effect becoming stronger for genuine hallucinations and increasing with verifier confidence. Further analysis separates this effect from the broader tendency associated with candidate-conditioned verification.

(2) We propose *BEV*, which structurally separates evidence interpretation from hypothesis evaluation: an evidence reader produces an answer without observing the candidate, which is then evaluated outside the evidence-reading computation, while conventional candidate-conditioned verification is retained as a complementary signal. This design removes the candidate from the evidence-interpretation stage without discarding information captured by conventional verification.

(3) We evaluate BEV across 29 retrieval-based verification configurations. It outperforms the corresponding yes/no verifier in 24 of 29, a clustered benchmark-level gain of $+0.016$ AUROC (p=0.015); the forward evidence-reading term alone, without the optional candidate-page signal, gives $+0.017$ (t=2.75, 23 of 29). The benefit is observed in both self-verification and third-party verification.

2 Related Work

Hallucination detection. Hallucination detectors use consistency across samples (Manakul et al., 2023; Zhang et al., 2023), semantic uncertainty (Kuhn et al., 2023; Farquhar et al., 2024; Nikitin et al., 2024; Kossen et al., 2024), hidden-state representations (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Chen et al., 2024; Du et al., 2024; Orgad et al., 2025), and elicited confidence (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024); see Geng et al. (2024); Huang et al. (2025). Confident, internally consistent errors remain challenging (Chauhan, 2026; Simhi et al., 2025b;a). Our audit complements this work by showing that the residual is not homogeneous: ambiguity and label errors can be mixed with genuine hallucinations (Min et al., 2020; Cole et al., 2023; Vendrow

et al., 2025; Atasoy et al., 2026; Staliūnaitė et al., 2026), while factual errors concentrate on long-tail entities (Kandpal et al., 2023; Mallen et al., 2023; Sun et al., 2024).

Self-preference and cross-examination. LLM judges favour their own generations over others' of equal quality (Panickssery et al., 2024; Wataoka et al., 2025), and cross-examination by a second model exposes factual errors (Cohen et al., 2023). Those settings judge open-ended text and the remedy is a different judge; we measure the effect under evidence-conditioned verification with question, evidence and both candidates fixed, find it graded by confidence and absent for third-party candidates, and remove the dependence with the same model.

Evidence-based verification. Evidence-based methods check generated claims against retrieved text, including FActScore, SAFE, and RARR (Min et al., 2023; Wei et al., 2024b; Gao et al., 2023); MiniCheck instead trains a grounding classifier (Tang et al., 2024). Factored Chain-of-Verification answers verification questions without the original response in context, but subsequently evaluates those answers against that response (Dhuliawala et al., 2024). Despite differences in retrieval and decomposition, these methods generally expose the candidate to the computation that decides whether the evidence supports it. BEV, introduced in Section 5, targets this shared dependence.

Entity-centric verification and evaluation quality. Entity-aware verification aligns generated entities or relations with evidence (Liu et al., 2020; Pham et al., 2025; Yeh et al., 2025; Jia et al., 2026), supported by entity-linking and retrieval techniques (Cao et al., 2021; Wu et al., 2020; Shavarani & Sarkar, 2025). Our external matcher is deliberately simpler: it isolates the effect of separating evidence reading from candidate evaluation rather than proposing a new entity-linking system; see Appendix I. Evaluation work further shows that answer equivalence, ambiguity, and reference-label errors can substantially change factuality judgments (Si et al., 2021; Kamalloo et al., 2023; Northcutt et al., 2021; Gema et al., 2025; Chen et al., 2026). This motivates our explicit audit of whether confident residual errors are genuine hallucinations or artifacts of the question and gold label.

3 What Are We Studying? Auditing Confident Errors

3.1 An operational definition: the q-stubborn regime

We fix a panel of four signals in advance, one per information source; its members read different things about the model, not four versions of one reading: *token-level belief*, the mean token log-probability of the greedy answer a_0; *sampling stability*, the fraction of $k{=}10$ samples matching a_0, the SelfCheckGPT-style signal of Manakul et al. (2023); *prompt robustness*, the fraction of four paraphrases answered with a_0; and *representation-level evidence*, a linear probe on the last-prompt-token hidden state. Answer matching uses a bidirectional whole-word alias match.

Definition. Let $\mathcal{C}$ be the model's correct answers on the dataset, $\mathcal{E}$ its errors, and $Q_q(s \mid \mathcal{C})$ the q-th percentile of signal s over $\mathcal{C}$. The *q-stubborn* errors are

$$\mathcal{S}_q \;=\; \big\{\, x \in \mathcal{E} \;:\; s_k(x) \;\geq\; Q_q(s_k \mid \mathcal{C}) \;\text{ for all } k = 1, \ldots, K \,\big\},$$

with $K{=}4$ the panel above; $q{=}25$ is the primary setting and $q \in \{10, 50\}$ is used for sensitivity. This is an operational regime defined relative to a prespecified panel; we do not claim these errors form a natural category or are hard for every detector. It isolates the cases where cheap signals give a model no reason to doubt itself, and Section 3.3 asks how much of that is an artifact of the panel. The definition is label-free apart from the calibration set $\mathcal{C}$; a variant that drops the probe ($K{=}3$) behaves the same and is used where hidden states are unavailable. "Stubborn" below means membership in $\mathcal{S}_q$ and nothing more.

Hierarchical location relations such as birthplace and headquarters, and multi-valued relations admit several defensible answers, so we drop them throughout; all reported numbers use this strict subset.

We use eight benchmarks: PopQA (Mallen et al., 2023), EntityQuestions (Sciavolino et al., 2021), TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), AmbigQA (Min et al., 2020), SimpleQA (Wei et al., 2024a), HaluEval QA (Li et al., 2023) and TruthfulQA (Lin et al., 2022), and four model families: Qwen2.5-Instruct at 1.5B–32B, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3 and Gemma-3-4B-it, plus Qwen2.5-7B base. After the strict subset the benchmarks

Table 1: What cross-model stubborn errors contain. The core comprises strict-subset questions that are 25-stubborn for at least two models. Label columns report percentages of the core; rows are ordered by the genuine-hallucination share. Totals may differ from 100 due to rounding. Rankings are stable across judges, but absolute shares vary; see Table 10.

		Judge label (% of core)				
Benchmark	Core size	Genuine hallucination	Ambiguity/ wrong gold	Coarser correct	Under-determined	Unparsed
SimpleQA	753	86	8	1	3	2
HaluEval QA	604	50	22	6	19	4
TruthfulQA	243	44	28	2	21	5
TriviaQA	94	41	28	5	18	7
EntityQuestions	134	37	41	1	19	1
AmbigQA	629	26	48	8	17	1
PopQA	286	25	65	0	10	0
Natural Questions	438	20	49	6	24	0

hold 787–11,816 questions; accuracy ranges from 3.5% on SimpleQA to 71% on TriviaQA, and q=25-stubborn errors are 0.9–5.1% of errors on the template benchmarks and 4–40% elsewhere; sizes are in Table 12.

3.2 How much of it is genuine hallucination, and where it concentrates

For the audit, every question 25-stubborn for at least one model is exported with its gold label, aliases and each model's answer. A judge, DeepSeek-v4-pro, answers five sub-questions and a fixed rule turns them into one of four labels, H genuine hallucination, G ambiguity or wrong gold, C coarser-but-correct, and U undetermined, with a strict rule counting a different real entity of the same kind as H. 9,762 strict-subset questions carry a label, covering 12,878 model–question pairs; Appendix B gives the protocol.

On the stratified 350-item sample of Table 10, two other judge families, a rewritten protocol and AmbigQA's human ambiguity annotations agree with the main judge on the ordering; a second family then re-judged the entire audit under the same protocol, and the ranking holds at full scale, Spearman $+0.929$ and κ 0.670, while the absolute H share moves up on every benchmark. We claim the ranking, not the absolute share, which no judge we ran estimates accurately; the verifier comparison changes by at most 0.006 under the second judge's labels, and no judge is a human annotator.

Table 1 shows the share of genuine hallucinations among cross-model stubborn errors running from 86% on human-verified SimpleQA down to 20–26% on Natural Questions, AmbigQA and template-built PopQA, where ambiguous questions and wrong gold labels take up to 65%; Figure 6 in the appendix plots it. The template benchmarks instantiate "⟨relation⟩ of ⟨entity⟩" from a knowledge base without checking whether the entity name has a dominant referent, so the question points at a famous work while the gold label belongs to an obscure homonym. Corpus co-occurrence confirms this for G items, where the subject co-occurs with the model's answer 12× more often than with the gold label (p=$9 \cdot 10^{-70}$), and not for H.

The genuine component is small but concentrated; a pre-registered criterion that asked it to be large failed, as Appendix A records. Inside the top confidence decile, the answers a system would actually serve, 16–63% of the errors that remain are stubborn and 5–53% are judge-confirmed genuine, with medians of 53% and 17% over eighteen model–benchmark configurations, against 0.3–16% of all errors, median 3%, broken down in Appendix G. Abstaining on low confidence makes this worse: it removes the ordinary errors and leaves these.

3.3 Is the regime a definitional artifact?

Because the regime is carved out by four signals, their failure on it is nearly tautological; Appendix F separates the two. Thresholding does not manufacture the difficulty: a signal that took no part in

the selection is still at chance (0.45–0.51), and detectors used in no definition, EigenScore and semantic entropy, are at chance under every variant. Once ordinary errors are matched on token confidence, however, representation-based detectors recover most of their power, answer-position probes reaching 0.66–0.68 on H against 0.75–0.78 on the matched errors, so part of the difficulty is high confidence itself; the re-query detectors, $P(\mathrm{True})$, InterrogateLLM and CoVe, show no residual beyond 0.03.

4 Why Verifying These Errors Is Hard

4.1 Internal signals do not separate them, so evidence is needed

Table 11 in the appendix separates the three populations. On H the four defining signals are near chance by construction, as are EigenScore (Chen et al., 2024) at 0.37–0.47 and discrete semantic entropy at 0.44–0.51, which enter no definition. The one internal exception is a probe read at the answer position instead of the prompt's last token, at 0.66–0.69, discussed in Appendix F; signals that re-query the model, InterrogateLLM (Yehuda et al., 2024) and Chain-of-Verification (Dhuliawala et al., 2024), reach 0.43–0.65. Nothing internal approaches the 0.89–0.96 the same model reaches with a passage in front of it, so the rest of the paper works with evidence. On G every signal is at chance, and earlier reports that self-verification fails there follow from pooling the two populations.

4.2 But evidence-based verification has a problem of its own: self-anchoring

Section 3.3 established that inside this regime the model's own signals rank its errors like its correct answers, so the natural move is to bring in evidence. The standard way of doing so is hypothesis-dependent: the verifier is shown the candidate and asked whether the evidence supports it. A candidate that pulls the verdict towards itself we call *hypothesis anchoring*. We measure what that costs separately in the two deployments, because the answer differs between them. Everything in this subsection is about *self-verification*; Section 4.3 reports the third-party case, where the same measurement finds nothing.

Fix the verifier R, the question x and the passages e, let a_R and a_M be the wrong answers two different models gave to x, and measure

$$\Delta_R(x) \;=\; \frac{V_R(x, a_R, e) - V_R(x, a_M, e)}{\sigma_R}, \qquad a_R \neq a_M, \text{ both wrong},$$

with σ_R the within-verifier SD of the paired difference. Both candidates are scored by the same R with the same prompt on the same evidence, and neither is supported by it, so $\mathbb{E}[\Delta_R] > 0$ is attributable to the identity of the candidate. One asymmetry survives this control: the verifier's own answer occurs literally in the passage somewhat more often; Appendix N measures it separately. Three 7–8B models on EQ and PopQA give twelve $(R, M, \text{benchmark})$ combinations and 10,634 matched pairs; replacing a_R by a candidate the verifier did not produce turns Δ into the third-party control of Section 4.3, the bottom row of Table 2.

Table 2 is the measurement. Over all wrong answers the leniency is small but unambiguous ($+0.025$ SD, $p{=}2 \cdot 10^{-4}$); sorted by the verifier's confidence in its own answer it is monotone, from -0.106 in the lowest quintile to $+0.122$ in the highest; and on judge-confirmed genuine errors it reaches $+0.117$. One caveat: the verifier's own answer also appears literally in the passage more often, and on pairs where neither candidate does the population effect survives ($+0.047$, $p{=}8 \cdot 10^{-11}$) while the judge-confirmed figure does not ($+0.074$, $n{=}207$, n.s.), so the subset numbers combine ownership with that lexical asymmetry; Appendix N separates them. It holds even where the average runs the other way: Qwen2.5-7B on EQ is *harsher* on its own answers overall yet flips on its stubborn ones. On the items the two rows share, a verifier discriminates about as well on its own errors as another model does, 0.919 against 0.921, so the leniency coexists with unimpaired discrimination.

A verifier that rewarded any candidate placed before it would produce the same table, so we add a third condition C: a candidate drawn at random from a different question of the same relation, equally unsupported and produced by nobody here. With A, B, C the verifier's scores on its own answer, the other model's and the random one,

Table 2: Ownership bias among wrong candidates: the standardised difference between a verifier's scores on two equally wrong candidates for the same question and evidence, positive meaning that it favours the first: its own answer in the top block, a candidate it did not produce in the bottom. In the bottom block neither candidate is the verifier's own, so the sign depends on which of the two fixed models is subtracted; there the offset does not grow on the judge-confirmed subset. The last block is a different quantity, overall discrimination, on the items both settings share.

ownership	**population**	**candidate visible** while the evidence is read	**candidate hidden** until the comparison
verifier	*setting*	self-verification	BEV, self-reader
= generator	all wrong answers	**+0.025**	−0.052
	judge-confirmed (H)	**+0.117**	−0.026
verifier	*setting*	third-party verification	BEV, independent reader
≠ generator	all wrong answers	−0.102	−0.034
	judge-confirmed (H)	−0.064	0.000
overall	*correct vs.* H *AUROC on the shared items (a different quantity):*		
discrimination	verifier = generator	0.919	0.937
	verifier ≠ generator	0.921	0.941

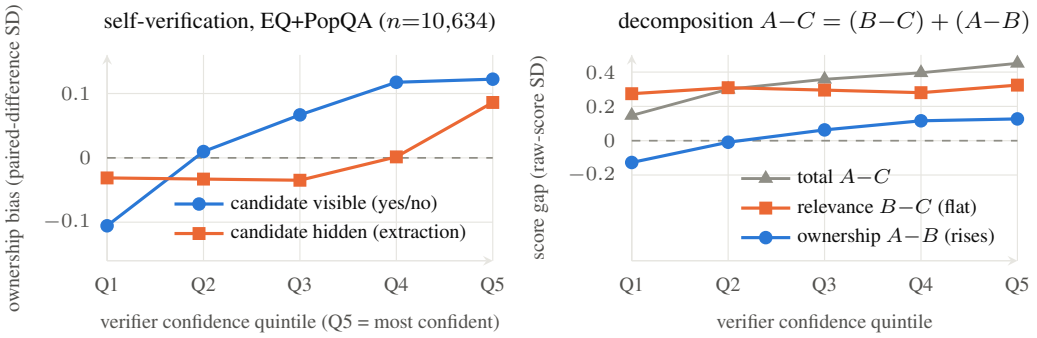


Figure 2: Ownership leniency grows with the verifier's confidence. *Left*: standardised bias towards the verifier's own wrong answer by confidence quintile, candidate visible against candidate hidden. *Right*: the three-way decomposition of Section 4.2, in raw-score SD; the relevance term is flat and the ownership term carries the whole trend.

$$\underbrace{A-C}_{+0.330} = \underbrace{(B-C)}_{\text{relevance } +0.297} + \underbrace{(A-B)}_{\text{ownership } +0.034},$$

in a common denominator, the raw-score SD instead of the paired-difference SD of Δ_R; Appendix E gives the derivation. Across confidence quintiles the relevance term is flat (0.274–0.324) while the ownership term moves from -0.127 to $+0.127$, the opposite of what confident questions being merely easier to adjudicate would predict; Figure 2 plots both terms.

Outside the template benchmarks, repeating the measurement on the multi-hop HaluEval QA, TriviaQA and Natural Questions adds 9,084 matched pairs and the effect is stronger: $+0.067$ SD overall (p=$2 \cdot 10^{-10}$), $+0.302$ on the stubborn subset and $+0.236$ on judge-confirmed genuine errors, against -0.021 on ordinary errors. Pooled, the measurement rests on 19,718 matched pairs over five benchmarks; the three cohorts are mapped in Appendix D.

The right column of Table 2 applies the same design to extraction and shows no leniency on the population as a whole (-0.052) or on judge-confirmed genuine errors (-0.026, n.s.). On the raw stubborn subset a residual of $+0.226$ remains (Table 3), and its source is measurable: there the model's own answer occurs literally in the evidence twice as often as the other's and is the extracted answer three times as often, whereas on judge-confirmed errors that gap is nil; see Appendix N.

Table 3: The factorial ablation. Candidate in the verifier's context × yes/no judgement or extraction; the fourth cell does not exist, since without a candidate there is no proposition to judge. Entries are leniency as in Table 2, not AUROC, on the good-passage corpora. Rows are nested: judge-confirmed genuine errors are a subset of the stubborn ones. All columns use the same 23,225 matched pairs.

	n	visible, yes/no		visible, extraction		hidden, extraction	
		bias	p	bias	p	bias	p
All wrong answers	23,225	$+0.025$	10^{-4}	-0.017	1.00	-0.052	1.00
stubborn	633	$+0.291$	$2 \cdot 10^{-7}$	$+0.239$	10^{-5}	$+0.226$	$5 \cdot 10^{-4}$
judge-confirmed (H)	278	$\mathbf{+0.117}$	0.047	$\mathbf{+0.066}$	0.14	$\mathbf{-0.026}$	0.64

Removing the candidate also changes the task, so we separate the factors as far as the design allows: only three cells of the 2×2 exist, since without a candidate there is no proposition to judge, so identification is partial. The third cell is an extraction prompt that *shows* the candidate while still asking the model to answer from the passage. Table 3 reports all three on the same items: on judge-confirmed genuine errors the bias is $+0.117$ (p=0.047) with the candidate visible, $+0.066$ (n.s.) when the task changes to extraction but the candidate stays, and -0.026 when it is hidden. Both moves are associated with a reduction and neither suffices alone; we stop short of attributing the effect to candidate visibility alone, and note only that a change of format with the candidate still present leaves a bias that is no longer significant but not yet absent.

4.3 Ownership or candidate conditioning? The two factors are separable

The bottom row of Table 2 runs the same design on a verifier that produced *neither* candidate: for each verifier and the two other models we score both of their wrong answers to the same question, candidate visible exactly as before. It shows no preference that grows on the judge-confirmed subset; its overall offset is a fixed quality difference between two models. Hypothesis dependence is therefore necessary for anchoring but not sufficient, and the cost we document belongs to self-verification, the arrangement most self-checking and single-model retrieval pipelines use; Appendix M rules out a family prior.

5 Making the Evidence-Interpretation Stage Hypothesis-Independent

BEV is the minimal implementation of the factorisation $\hat{a} = R(x, e)$, $s = \mathrm{match}(\hat{a}, a)$ stated in the introduction, together with a second instantiation that applies the same move to retrieval. The claim concerns the dependence structure, not the implementation, so the robustness experiments vary everything except that structure.

Forward: extract, then compare. Given passages P and question x, the reader answers x *from* P *only*, replying "not in passage" if it cannot; the candidate a_0 never appears in the prompt. Since the base term keeps the yes/no score, BEV reads the passages twice, one classification pass and one generation pass, against one for a single-pass baseline. Writing $\hat{a} = R(x, P)$ and $\varnothing$ for abstention,

$$m = \begin{cases} +1 & \mathrm{match}(\hat{a}, a_0) \\ 0 & \hat{a} = \varnothing \\ -1 & \text{otherwise,} \end{cases} \qquad s_{\mathrm{fwd}} = \mathrm{Pc}(\text{yes/no}) + w\, m.$$

Here $\mathrm{Pc}(\cdot) \in [0, 1]$ is the percentile of the yes/no log-odds among the model's own correct answers, held out five-fold for the correct items, and w=0.5 was fixed before the held-out corpus was collected; the AUROC is flat in w over $[0.25, 1]$ and every positive weight beats w=0, as Appendix L shows. The matcher is whole-word and bidirectional after normalisation, so "Spielberg" matches "Steven Spielberg"; gold aliases are used only in the audit, never at inference; Appendix Q gives the matcher in full. BEV is thus a hybrid: a candidate-free evidence correction to a hypothesis-dependent base score.

Table 4: Search-corpus verification by benchmark. Mean correct-vs-H AUROC across configurations. Gain is BEV minus the yes/no verifier; "positive" counts configurations with gain above zero. Every benchmark mean is positive; the "All" rows average over configurations, and the mean of the seven benchmark means is $+0.016$ (t=3.39, p=0.015). Seven of 29 confidence intervals lie entirely above zero and none lie below. Per-configuration results are in Table 5.

Benchmark	Configs	Yes/no AUROC	BEV AUROC	Gain Δ	Positive configs
Full score (candidate-page signal included where applicable)					
PopQA	5	0.755	0.797	$+0.041$	5/5
EntityQuestions	4	0.829	0.849	$+0.019$	4/4
TriviaQA	4	0.729	0.745	$+0.017$	4/4
AmbigQA	4	0.651	0.667	$+0.015$	3/4
SimpleQA	4	0.536	0.544	$+0.008$	3/4
HaluEval QA	4	0.631	0.639	$+0.008$	3/4
Natural Questions	4	0.639	0.642	$+0.003$	2/4
All	29	0.684	0.701	$\mathbf{+0.017}$	**24/29**
Forward term only (candidate-page signal removed from both verifiers)					
Template benchmarks (PopQA, EQ)	9	0.713	0.747	$+0.034$	8/9
Open-domain benchmarks	20	0.637	0.647	$+0.010$	15/20
All	29	0.661	0.678	$\mathbf{+0.017}$	**23/29**

Candidate-page consistency: an optional auxiliary. The same move applies to retrieval, which normally queries with the question and inherits the anchor that produced the error. Querying *by the answer*, we fetch the page of the candidate and test whether the question's subject occurs in it:

$$r \;=\; \mathbf{1}[\,\mathrm{subj}(x) \subseteq \mathrm{page}(a_0)\,] \in \{0,1\}, \qquad s \;=\; s_{\mathrm{fwd}} + r,$$

at zero model calls, and keyed on the *candidate* a_0, not on $\hat{a}$. A hit is worth 1.0, twice a full forward agreement, while r=0 subtracts nothing, because correct answers often fail this test and a veto would suppress them, not the errors. Every baseline receives the same bit on the same items, as Table 7 verifies. It is an optional auxiliary, outside the general claim: it presumes a direct subject–answer relation and its value is relation-dependent, which we measure: "author" 0.73 on correct answers against 0.10 on genuine errors, "educated at" 0.02 even on correct answers; Appendix P reports every relation.

5.1 Experiments

All verifiers receive the same model and passages: a FActScore-style claim verifier (Min et al., 2023), the yes/no verifier, a trained entailment model, extract-then-compare alone, and BEV; baselines without retrieval are $P(\mathrm{True})$, verbalised confidence and cross-model agreement. Three corpora: *gold page*, an oracle covering 68% of PopQA; *DPR*, the top-3 shipped with EntityQuestions; and *search*, intentionally naive top-1 Wikipedia search. Metrics are correct-vs-H AUROC with bootstrap CIs and H recall at 10% false-positive rate. We also record each corpus's *extractable rate*, the fraction of correct answers recoverable from the passage.

Table 4 reports the search corpus. Configurations within a benchmark are not independent, so benchmarks are the units of analysis: the mean of the seven benchmark means is $+0.016$ (t=3.39, p=0.015) and a hierarchical bootstrap gives $+0.017$, CI $[+0.008, +0.030]$, in Appendix O; the 24-of-29 count is descriptive, and Figure 4 in the appendix shows each interval. Against the best of three baselines chosen per dataset after seeing the labels we win 19 of 29 with a mean of $+0.001$, which bounds what per-benchmark selection could recover. The candidate-page bit does not produce the headline: removed from both verifiers, the forward term alone still gives $+0.017$ (t=2.75, 23 of 29 positive), $+0.034$ on the two template benchmarks and $+0.010$ on the five others, positive in both aggregates. It survives every substitution we tried, listed in Appendix C.

Across 24 model–corpus configurations the extractable rate almost perfectly determines how well any verifier can do, r=$+0.95$ against the yes/no verifier's AUROC and $+0.97$ against ours, as Figure 5 in the appendix shows, and is uninformative about the difference between them: r=-0.02,

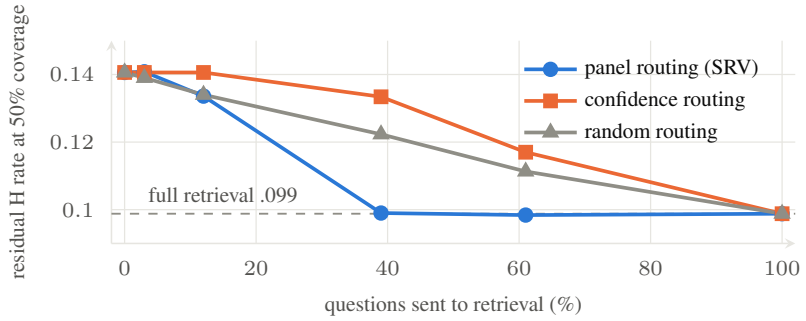


Figure 3: Retrieval budget versus residual error. Mean over six settings. At 39% retrieval, panel routing reaches the full-retrieval rate (0.099); confidence and random routing yield 0.133 and 0.122.

p=0.94, with +0.013 below an extractable rate of 0.3 against +0.018 above. The corpus sets the ceiling, not the margin: internal signals without retrieval give 0.37–0.48, the good passages of Table 9 0.92–0.97 and the search corpus 0.64–0.80, so the absolute level of any correct-vs-H AUROC here reflects the evidence available, not the class. The margin does move with evidence quality *within* a fixed benchmark, where DPR passages at increasing rank depth erode it to nothing by ranks 18–20 in Table 8; across benchmarks it is confounded with question difficulty and disappears.

The gain is not the removal of anchoring. If it were, the advantage should shrink when the verifier did not produce the answer; over eighteen reader×generator configurations the margin is +0.012 when the verifier is the generator (6/6) and +0.020 when it is not (11/12), not significant and running the other way. Together with Appendix D, where neither factor improves standalone ranking, this places the effect in the combination of the two moves, not in a mechanism we can name.

The forward term is the consistent half: at or above yes/no on 6/6 good-passage pairs, 17/18 reader×generator configurations and 9/9 prompt variants in Appendix C.

In deployment, retrieval is the expensive stage, so the question is which questions to spend it on. Routing by the panel, i.e. sending only questions whose cheap signals cannot separate them from correct answers, and composing stages by the minimum of their percentiles gives $s_{\text{SRV}} = \min_i \text{Pc}_i$, under which an added evidence stage can only lower a score: evidence vetoes and never inflates. Figure 3 is the cost–quality curve: at 39% retrieval coverage it matches the full-retrieval residual error rate, against 0.122 for random routing on the same budget. Random routing is the neutral comparison: a veto-only rule cannot promote an item, which handicaps the least-confident baseline by design and leaves it about a sixth of the benefit, since the questions that need verification are not those the model is least sure of. The minimum is conservative, not optimal; Appendix R compares operators.

6 Limitations and Conclusion

Limitations. *(i)* The q-stubborn regime is operational and relative to a finite signal panel, and, as Section 3.3 showed, part of its difficulty is high confidence itself. *(ii)* The corpus, not the method, sets the absolute level, as Section 5.1 showed, and because the decision is a string match the method verifies entity-valued answers, with failures by answer type in Appendix K; we compared against neither soft-alignment scorers (Zha et al., 2023) nor the dedicated RAG hallucination benchmarks (Niu et al., 2024; Ravi et al., 2024), so generality beyond short-answer QA is untested. *(iii)* The audit rests on language-model judges, so the benchmark ranking is the claim, not the absolute shares; and the margin against a per-benchmark oracle baseline is not significant.

Conclusion. A leading question is not universally bad for verification. A verifier that reads the evidence conditioned on a candidate it produced itself is lenient towards that candidate, graded by its own confidence; one that did not produce it is not. Separating evidence interpretation from candidate evaluation removes the dependence without forbidding the question, and it helps in both deployments.

The Use of Large Language Models

Large language models were used to aid and polish the writing of this paper. They are also the object of study: the models whose errors we audit and verify are LLMs, and the audit labels of Section 3 are produced by the LLM judges described in Appendix B.

Ethics Statement

This work uses publicly released question-answering benchmarks and openly available model weights; no human subjects were involved and no personal data were collected. All audit labels come from LLM judges rather than human annotators, and we claim only the benchmark ranking they agree on, not the absolute shares. The method is a detector of confident errors and is intended to reduce, not produce, misinformation. We are aware of no conflicts of interest.

Reproducibility Statement

Every quantity needed to reproduce the results is specified in the paper. Section 3 defines the stubborn regime and the signal panel, and Table 12 gives the resulting sample sizes per benchmark and model; Appendix B gives the judge prompt, the labelling rule and the calibration study; Appendix A records the pre-registered criteria; Appendix Q gives the matcher, abstention rule, prompts and inference settings; Appendix O describes the clustered and bootstrap analyses; Appendix C lists every ablation with its setting. Models, decoding settings and the retrieval corpora are stated where first used in Sections 3 and 5.1.

References

I. F. Atasoy, B. Mutlu, E. A. Sezer, and A. Wahdan. Do benchmarks underestimate LLM performance? evaluating hallucination detection with LLM-first human-adjudicated assessment. arXiv preprint arXiv:2605.08462, 2026.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *International Conference on Learning Representations*, 2021.

Mohit Singh Chauhan. DECK: A consistency x confidence taxonomy of LLM hallucinations. arXiv preprint arXiv:2606.02289, 2026.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Wenbo Chen, Veena Padmanabhan, Tootiya Giyahchi, Elaine Wong, and Leman Akoglu. Rethinking evaluation for LLM hallucination detection: A desiderata, a new RAG-based benchmark, new insights. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14912–14931, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. LM vs LM: Detecting factual errors via cross examination. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12621–12640, Singapore, December 2023. Association for Computational Linguistics.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 530–543, Singapore, December 2023. Association for Computational Linguistics.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102948–102972. Curran Associates, Inc., 2024.

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

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, July 2023. Association for Computational Linguistics.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188.

Runsong Jia, Zhen Fang, Mengjia Wu, Jie Lu, and Yi Zhang. Evidence-aligned entity verification for hallucination detection in retrieval-augmented generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29538–29554, San Diego, California, United States, July 2026. Association for Computational Linguistics.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221, 2022.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. arXiv preprint arXiv:2406.15927, 2024.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, Singapore, December 2023. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained fact verification with kernel graph attention network. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7342–7351, Online, July 2020. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8901–8929. Curran Associates, Inc., 2024.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Curtis Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024.

Hoang Pham, Thanh-Do Nguyen, and Khac-Hoai Nam Bui. Verify-in-the-graph: Entity disambiguation enhancement for complex claim verification with interactive graph representation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5181–5197, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. Lynx: An open source hallucination evaluation model. arXiv preprint arXiv:2407.08488, 2024.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6138–6148, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Hassan Shavarani and Anoop Sarkar. Entity retrieval for answering entity-centric questions. In Weijia Shi, Wenhao Yu, Akari Asai, Meng Jiang, Greg Durrett, Hannaneh Hajishirzi, and Luke Zettlemoyer (eds.), *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pp. 1–17, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. What's in a name? answer equivalence for open-domain question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9623–9629, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. Distinguishing ignorance from error in LLM hallucinations. arXiv preprint arXiv:2410.22071, 2025a.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, I'm wrong: LLMs hallucinate with certainty despite knowing the answer. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14665–14688, Suzhou, China, November 2025b. Association for Computational Linguistics.

Ieva Raminta Staliūnaitė, James Bishop, and Andreas Vlachos. The role of ambiguity in error prediction via uncertainty quantification. arXiv preprint arXiv:2606.02093, 2026.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 311–325, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? arXiv preprint arXiv:2502.03461, 2025.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-Judge. arXiv preprint arXiv:2410.21819, 2025.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. arXiv preprint arXiv:2411.04368, 2024a.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 80756–80827. Curran Associates, Inc., 2024b.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6397–6407, Online, November 2020. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Min-Hsuan Yeh, Max Kamachee, Seongheon Park, and Yixuan Li. HalluEntity: Benchmarking and understanding entity-level hallucination detection. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9333–9347, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15445–15458, Singapore, December 2023. Association for Computational Linguistics.

A PRE-REGISTRATION

Three criteria were pre-registered. (A) *Existence*: stubborn errors should be at least 8% of all errors at q=50. (B) *Irreducibility*: a set built from three of the four signals should not be separable by the held-out fourth. (C) *Systematicity*: they should recur across models above chance. All three were registered before the audit; Appendix A reports their outcomes, including the failure of (A).

Before running the full-scale experiments we fixed three criteria for the definition and one for the method, with magnitudes. We report each as written.

1. **(A) Existence.** At q=50, stubborn errors should be at least 8% of all errors. *Failed:* 0.87% on the full set and 0.14% on the strict subset (16.1% / 7.5% at q=10). We did not move the threshold. The conjunction "all four signals above the median of correct answers" is severe (only 21% of correct answers pass it), so the more informative quantity is residual-error attribution: for a detector that fuses the four signals (AUROC 0.917 on all errors), stubborn errors (q=10) make up 44.8% of its false negatives at 95% correct-answer retention, 61.3% at 90%, and 78.7% at 80%. The class is small among all errors and dominates what is left after the detector toolbox has been applied.
2. **(B) Irreducibility (leave-one-out).** A stubborn set defined by three of the four signals should not be separable by the held-out fourth: AUROC ≤ 0.60. *Passed* for $q \geq 25$ (0.33–0.61).
3. **(C) Cross-model systematicity.** Restricted to questions two models both answer wrongly, the overlap of their stubborn sets should exceed the hypergeometric expectation by $\geq 2\times$ with $p < 10^{-3}$. *Passed* (3.40×, p=$3.2 \cdot 10^{-39}$ at the pilot scale; 3–14× for genuine errors alone at full scale, Section 4).
4. **(D) Forward verifier, held-out.** Fixed after the good-passage experiments and before fetching the search corpus: extract-then-compare with w=0.5 must be $\geq$ the yes/no verifier (correct vs. H AUROC) on at least 4 of the 6 seven-billion-parameter model–corpus pairs (Qwen/Llama/Mistral × PopQA/EQ) on the search corpus, otherwise the forward term is demoted to "a cheaper equivalent of yes/no" and the method claim is restricted to routing and the reverse check. *Outcome:* 5/6 (Mistral on EQ is -0.005, Table 5).

The alias-matching convention (strict vs. lenient) does not change any of these outcomes (7.45% vs. 7.48% at q=10 on the strict subset).

B JUDGE PROTOCOL AND CALIBRATION

Items. Every question that is 25-stubborn for at least one model on a dataset is exported once with: the question, the gold answer and its aliases, the Wikipedia title of the intended entity where the dataset provides one (PopQA), and every model's answer with its per-model signals. Items stubborn for $\geq$ 2 models are labelled first (priority 1). Totals: 7,253 items over the six original datasets, 2,182 for HaluEval, 502 for TruthfulQA, 1,343 for Gemma-3-4B on PopQA and EQ, and 204 additional for Qwen2.5-32B on SimpleQA, 11,484 item records exported in all, of which 11,399 returned a parsable label; because the Gemma batch re-exports questions already present in the first batch, those records cover 10,980 distinct questions. Three counts should be kept apart: 11,484 records exported, 11,399 labelled, and the 9,762 strict-subset questions stubborn for at least one

model that carry a valid label, covering 12,878 model–question pairs; the last is what every reported number uses. Parse failures were re-queried; the residual, 0.3–2.8% per batch, is left *unlabelled* instead of folded into U, and appears as the n/a column of Table 1.

Prompt (v4). The judge (DeepSeek-v4-pro, thinking disabled, temperature 0) answers five sub-questions and returns a JSON object; a fixed rule derives the final label instead of asking for it directly:

- Q1 `gold_correct`: for the entity named by the Wikipedia title, is the gold answer factually correct? (yes / no / unsure)
- Q2 `question_unambiguous`: reading only the question text, does it uniquely refer to the intended entity, or would a well-informed reader take it to mean a different, better-known entity of the same name?
- Q3 `model_defensible`: is at least one model answer correct under some reasonable reading? *Strict near-neighbour rule:* a different real entity of the same type (a relative, a sequel or another work in the same series, a co-worker, a namesake with a different role) is not defensible and is labelled H.
- Q4 `coarser_correct`: is a model answer correct but coarser than the gold (*Cuba* for *Cienfuegos*?)
- Q5 `error_type`: name collision / slot error / type error / near neighbour / unrelated.

Q1, Q2 and Q4 are item-level; Q3 is asked once but the judge also returns a per-answer entry, and the rule below is applied to each answer separately, the item label being the worst case (H $>$ G $>$ C $>$ U). Per answer: C if it is coarser-correct; G if the gold is wrong, or the question is ambiguous *and this answer* is defensible; H if the gold is correct, the question unambiguous, and this answer wrong; U otherwise. Two consequences are intended. An answer that is indefensible but whose question is ambiguous falls to U instead of H: an ambiguous question does not license calling any particular answer a hallucination. And one model's defensible answer does not protect another's: the labels are per answer, so a defensible reading by one model and an unrelated entity from another give G and H respectively on the same item.

Second judge. The 94 PopQA cross-model core items were relabelled by a second LLM judge (Claude, with web search enabled) instructed to verify each item against the web before answering; we did not record the exact checkpoint for this 94-item spot check, which is a reproducibility gap; the full-scale second-judge relabel below uses GLM-5.2 and is fully specified. Agreement with the first judge: 83%. Every disagreement was the first judge accepting a near-neighbour answer (father for son, another film in the series) as defensible; none went the other way. The near-neighbour rule in v4 was introduced in response and the first judge was rerun.

Calibration against human labels. AmbigQA questions carry a human annotation of whether the question has multiple valid answers. Over all 2,043 audited AmbigQA items the judge assigns G to 48% of human-ambiguous questions and to 29% of single-answer questions (odds ratio 2.34, Fisher two-sided $p{=}3 \cdot 10^{-20}$). The two notions are related but not identical: AmbigQA marks answer multiplicity, whereas G also covers wrong gold labels, so the rate on single-answer questions is not zero.

Examples. G: "Who was the director of *The Shining*?" (gold: Mick Garris, for the 1997 miniseries; models: Stanley Kubrick); "Who is the father of George Washington?" (gold refers to a racehorse); "Who was the screenwriter for *Gravity*?" (gold: George Clooney; models: Alfonso Cuarón; the gold is wrong). H: "Where is the headquarters of Anorthosis Famagusta?" (gold: Larnaca; models: Famagusta, confidently and consistently); "Who performed the David Lee Roth discography?" (models: Van Halen). C: "Where was Yanet Bermoy born?" (gold: Cienfuegos; models: Cuba).

C Ablations and robustness checks in full

On the oracle page and the shipped DPR top-3 (Table 9, appendix) the forward verifier is above yes/no on all 10 model–corpus pairs and $\geq$ yes/no in 17 of 18 reader$\times$generator configurations,

the FActScore-style verifier being weakest in all 18; rewriting a question as a declarative claim is the strongest form of the conditioning we remove. Appendix C adds three prompt wordings (nine model–wording configurations), the DPR rank sweep, granting the auxiliary bit to every baseline, and a fourth model family added after the design was frozen.

The sieve as a gate (SRV). The principle also says *when* evidence is worth fetching, which turns the definition of Section 3 into a gate instead of adding a third contribution. Each free signal is mapped to its percentile among the model's correct answers; only questions with $\mathrm{Pc}(\mathrm{conf}) \geq q$ are sent to retrieval, and the final score is the minimum of the percentiles of the stages taken, so evidence can only veto. At q=0.25 the gate routes 31–42% of questions over eight model–corpus configurations; it lets through three quarters of the correct answers by construction, but covers 89–100% of the genuine stubborn errors and only 5–15% of the ordinary ones, and so retains 70–100% of the H recall that full retrieval achieves, against 27–57% for a random gate on the same budget. The verifier can also be much smaller: Qwen2.5-1.5B and 3B verifying the 7B/8B models' answers beat yes/no in all nine configurations, by a *larger* margin than self-verification ($+0.036$ at 3B against $+0.019$), because extraction asks a small model only to read; together, this costs roughly one small-model call per three questions. Read as a selective-prediction rule (Table 6), the same scores show that *abstaining by confidence is actively harmful for this class*, raising the residual stubborn rate from 6.9% to 9.3% and from 11.6% to 15.6%, while BEV is the best rule at almost every answer rate though the reduction is modest (11.6% to 7.0% at half coverage). Full numbers in Appendices C and G.

Neither construction detects a hallucination on its own; both convert external evidence into a verdict, so the corpus's extractable rate, the fraction of the model's correct answers recoverable from the passage, is the prerequisite, not a nuisance parameter. It is measurable without labels and sets the ceiling: it correlates $+0.95$ with the yes/no verifier's absolute AUROC and $+0.97$ with ours, but *not* with the margin between them ($r = -0.02$), which stays positive on average even below an extractable rate of 0.3 ($+0.013$ against $+0.018$ above; Appendix C). The percentile calibration needs no target-domain labels either: pseudo-calibration from confidence alone costs at most 0.02 AUROC (Appendix J).

Scale (1.5B→32B), post-training and corpus co-occurrence all fail to account for the regime; the one probe that separates H from G, reverse association, does not separate genuine errors from correct answers (Appendix C).

Systematicity. Restricted to questions two models both answer wrongly, the overlap of their stubborn sets exceeds the hypergeometric expectation by 3–14× for H alone (median over model pairs: PopQA 13.8×, EQ 5.6×, TriviaQA 11.0×, NQ 9.0×, AmbigQA 7.9×, HaluEval 4.1×, SimpleQA 2.8×; all $p < 10^{-3}$ except five small-count pairs) and by up to 35× for G. Both are properties of the question. At n=6,000 we had concluded the opposite for H from 0–4 co-occurrences; the full-scale result retracts that.

Cross-model convergence, the signature usually read as "shared pre-training error", is on template benchmarks mostly the signature of ambiguity. Evaluating a detector on the stubborn tail of PopQA or EntityQuestions penalises it for errors the model did not make; the sieve enriches such items 2–3× over a random sample of errors.

Scale, post-training, corpus. Across Qwen2.5 1.5B→32B the share of stubborn errors among all errors rises 2.7%→3.3%→5.5%→9.6%→9.5% at q=10 (the scale series is reported at q=10, where every size has a workable count; at q=25 the same ordering holds on smaller numbers), saturating at 14B: larger models remove the ordinary "don't know" errors and keep this class. The base model is not cleaner: its stubborn share is higher (17.8% vs. 9.2% at 7B), 49% of the instruct model's stubborn questions receive the same wrong answer from the base model (overlap 3.3×, p=10^{-59}), and instruction tuning only inflates confidence uniformly. Genuine stubborn errors show no distinctive co-occurrence pattern in the pre-training corpus; their mechanism remains unexplained.

Reverse association. The one internal probe that separates H from G asks the model the inverse question ("Which films did a_0 direct?") and checks whether the original subject appears: 0.13–0.28 for H versus 0.41–0.67 for G. A genuine stubborn error is a one-directional surface attractor; an

ambiguous question's "wrong" answer is a true fact about the dominant referent. This observation motivates the reverse check in Section 5.

When is the reverse check worth it? Applying it to five benchmarks gives a rule, not a uniform gain, governed by one quantity a practitioner can measure without labels, the reverse hit rate on the model's own correct answers, and by the forward verifier's strength. It is worth it when entities are prominent and forward verification is weak (PopQA, EQ: $+0.07$ to $+0.17$); unnecessary when forward verification is already strong (TriviaQA); of no purchase when entities are obscure (SimpleQA, hit rate 0.15–0.24 even for correct answers); and premise-violating on multi-hop questions (HaluEval), where the entity named in the question is usually not the one the answer stands in a relation to. Appendix C gives the four regimes in full.

Is the advantage just one more retrieval call? Table 7 (appendix) grants the reverse bit to every verifier at the same weight. It helps all of them (mean $+0.08$), so it is an orthogonal signal any verifier should use, not a private gain of ours; with it granted to all, our forward term still gives the best combination in 8 of 9 pairs. Used alone (one page fetch, one substring test, no model call) it reaches 0.61–0.79, at or above self-verification (0.48–0.77). Giving every verifier the answer page as *text* instead of as a bit makes all of them *worse* (0.72–0.87 against 0.77–0.90), so we keep the reverse check symbolic.

What this buys a deployed system. AUROC is not what a practitioner spends, so Table 6 (appendix) reads the same scores as a selective-prediction rule: abstain on the lowest-scoring fraction and measure what is left. *Abstaining by confidence is actively harmful for this class*: it strips out ordinary errors (21.0%→2.2%, 25.5%→5.0%) and raises the residual stubborn rate from 6.9% to 9.3% and from 11.6% to 15.6%. One of our two rules is best or tied at every answer rate, and the bidirectional one is best in seven of the eight columns; $P(\mathrm{True})$ also lowers both populations at once, so that is not a property unique to ours. The reduction is modest either way, 11.6% to 7.0% at half coverage. An AUROC of 0.71–0.89 on correct-vs-H is not a solved problem; answer-entity page coverage is 0.55 and 0.52 for these two configurations (0.24–0.60 across all nine), and uncovered questions get no bonus, so these numbers are conservative.

Is it the wording of one prompt? We re-ran the forward verifier with three different extraction instructions: (a) "answer using only the passage, reply *Not in passage* if it does not contain the answer" (the default), (b) "answer from the passage, answer only, reply *unknown* if you cannot", and (c) "extract the span that answers the question verbatim, reply *none* if there is none". On EQ with DPR top-3 the correct-vs-H AUROC is 0.931/0.942/0.942 (Qwen), 0.921/0.914/0.913 (Llama) and 0.922/0.916/0.915 (Mistral): all nine variants are above the corresponding yes/no baseline (0.908/0.906/0.903), and the spread within a model is at most 0.011. The extractable fraction varies slightly (0.68–0.75; the verbatim-span instruction is the strictest). The gain comes from the structure, the absence of the candidate from the prompt, and not from one wording.

Does it hold in a fourth family? Gemma-3-4B-it (Google) was added after all design choices were fixed, and run through the same pipeline on PopQA and EQ. The phenomenon replicates: stubborn errors are 4.3% and 7.2% of its errors, and its stubborn set overlaps those of the other families 3.0–9.4$\times$ above the hypergeometric expectation on questions both models answer wrongly. So does the method: on four corpora (EQ DPR top-3, PopQA gold page, and the search corpus for both) the forward verifier beats yes/no 4/4 (0.940 vs. 0.918, 0.941 vs. 0.935, 0.788 vs. 0.760, 0.684 vs. 0.645), with the two largest gains on the realistic corpus, as the quality sweep below predicts.

Is the advantage an artifact of a weak retriever? Table 8 (appendix) answers this with the retriever shipped with EntityQuestions, taken at increasing rank depth. As passages get worse the extractable fraction of correct answers falls 0.74→0.30 and every verifier degrades monotonically (yes/no 0.91→0.69, ours 0.93→0.70). The realistic search corpus sits in the middle of that range (0.42–0.46), not at the bottom, so the held-out comparison is not run at an artificially weak operating point. The advantage is largest at intermediate quality (+0.01 to +0.05, peaking at ranks 4–6 and 11–13), small at the top where yes/no is near its ceiling, and absent at ranks 18–20: with no extractable evidence there is nothing to compare. That boundary is visible in the same experiment that rules out the artifact.

Ambiguity flag. Verifying both the model's answer and the gold label against the same passages gives a benchmark-hygiene signal: *model answer supported* $\wedge$ *gold not supported* fires on 25–36% of G items and 3–8% of H items on EntityQuestions (2–3% of correct answers, 1% of ordinary errors).

The reverse term rewards ambiguous questions by design (their answer's page does mention the subject), so correct-vs-G AUROC drops from 0.60–0.69 to 0.56–0.69 on PopQA; recall on ordinary errors is 1–4 points lower on PopQA and 1–27 points higher on EQ. Verifier ranking is model-sensitive on the gold page: at 14B the yes/no prompt malfunctions (ordinary-error recall 0.31) and the FActScore-style verifier is best, 0.955 against our 0.947, the one setting in which we are not first; at 32B the ordering reverts (ours 0.945, yes/no 0.942, FActScore-style 0.899). BEV is first or second in every setting we ran and never last.

How cheap can the verifier be? The verifier need not be the model under test, and it need not be large. We let Qwen2.5-1.5B and Qwen2.5-3B verify the answers of the 7B/8B models on EQ with DPR top-3. With the same reader, extract-then-compare beats yes/no in all nine configurations, and the margin is *larger* for the small readers (mean $+0.036$ at 3B versus $+0.019$ for self-verification): the yes/no prompt asks for a judgement, a small model's weakest task, whereas extraction only asks it to read. The extractable fraction is essentially independent of reader size (0.73–0.77). A 1.5B verifier with our method reaches 0.894/0.880/0.885 on the three answer sources, against 0.908/0.906/0.903 for a 7B/8B verifier with yes/no: a five-fold smaller verifier costs about 0.02 AUROC. Combined with the router, which sends only a third of the questions to retrieval, verification cost drops by roughly an order of magnitude.

Routing, thresholds, and what did not work. On the gold-page corpus SRV at $q{=}0.10/0.25$ routes 49%/31% of PopQA questions and keeps H recall at 0.79/0.69 for Llama (full retrieval 0.86), 61%/40% at 0.77/0.77 for Mistral (0.77), and 49%/34% at 0.72/0.75 for Qwen-32B (0.82); a random gate on the same budgets keeps 0.36/0.23, 0.50/0.41 and 0.45/0.35 respectively. On EQ, where DPR passages are strong on ordinary errors too, it saves 38–67% of calls at a cost of 10–22% of H recall: an efficiency device, not an accuracy one. Because every stage is a percentile among the model's own correct answers, thresholds transfer: set at 10% false-positive rate on one model–dataset pair and applied to the others, the median realised rate is 0.10 (range [0.04,0.16] within a benchmark, [0.01,0.24] across). Negative results: fusing retrieval with $P(\mathrm{True})$ or verbalised confidence lowers H AUROC on 9 of 10 good-corpus pairs; a "does the passage state a *different* answer" prompt is model-dependent (0.58/0.63/0.91); a corpus built from the page whose title equals the subject string covers 38% of questions with mostly homonym pages and is unusable. One exploratory, non-pre-registered observation runs the other way: on the *weak* corpus, adding $P(\mathrm{True})$ helps consistently (forward 27 of 29, bidirectional 6 of 8, up to $+0.11$); self-verification is informative only when the passage does not contain the answer.

D THE FACTORIAL ABLATION IN FULL

Which cohort is which. Three matched-pair samples appear in Section 4.2 and they are nested differently. *10,634*: the three 7–8B models on EQ and PopQA with good passages, the sample behind the quintile table, the three-way decomposition and the lexical control. *9,084*: the same design on HaluEval QA, TriviaQA and Natural Questions with the search corpus; *19,718* is the union of the two, and the pooled figure quoted in the abstract. *23,225*: the good-passage factorial, which adds Qwen2.5-14B and Gemma-3-4B to the first cohort so that all three prompt conditions run on identical items. This third cohort, not the 10,634-pair one described alongside Δ_R in Section 4.2, is what Table 2 reports. The stubborn subset is 633 in both the first and the third cohort, because the two added models contribute no new *jointly* stubborn pairs on these benchmarks.

Does it hold outside template entity questions? Repeating the measurement on three differently built benchmarks, HaluEval QA (HotpotQA-derived, multi-hop), TriviaQA and Natural Questions, with the search corpus adds 9,084 matched pairs, and the effect is stronger there: $+0.067$ SD overall ($p{=}2 \cdot 10^{-10}$), $+0.302$ on the stubborn subset ($p{=}5 \cdot 10^{-44}$, $n{=}2{,}470$) and $+0.236$ on judge-confirmed genuine errors ($p{=}2 \cdot 10^{-15}$), against -0.021 on ordinary errors, again confined to the confident tail. The three agree ($+0.274$, $+0.316$, $+0.323$ on their stubborn subsets) and

the monotone dependence on confidence reappears (-0.138 to $+0.284$, $\rho = +0.131$). Pooled, the measurement rests on 19,718 matched pairs over five benchmarks.

Two factors: whether the candidate appears in the verifier's context, and whether the task is a yes/no judgement or an extraction. The fourth cell does not exist, since without a candidate there is no proposition to judge, so three cells are measurable. All three columns are computed on the same items (23,225 matched pairs for which all six sources are available), normalised by the within-verifier SD of the paired difference. This is the same cohort as Table 2, so the two tables agree cell for cell where they overlap; it is larger than the 10,634-pair sample behind the quintile and decomposition tables, so absolute levels are not comparable with those.

The three columns decrease monotonically in every row. On judge-confirmed genuine errors, changing the task while leaving the candidate in context removes about half the bias and leaves a residual that is no longer significant; removing the candidate as well takes it to zero. Both moves contribute and neither is sufficient alone, which is the reason BEV makes both.

The same ablation on ranking accuracy. Scoring the three conditions as standalone detectors (correct vs. H AUROC, raw signals without percentile calibration, twelve model–corpus configurations) gives a different and equally informative answer: hiding the candidate changes AUROC by -0.007 on average (4 of 12 positive) and switching to an extraction task by -0.008 (3 of 12 positive). Neither factor improves discrimination on its own. This is consistent with the bias table, not in tension with it: the hypothesis-independent construction buys an *unbiased* score, not a sharper one, and the gain reported in Section 5.1 comes from composing that unbiased evidence term with the calibrated verifier. s_{fwd} is *additive*, so a matching extraction raises the score by w; the veto-only property belongs to the $\min_i \mathrm{Pc}_i$ aggregator of the router (Appendix R), not to s_{fwd}.

Sensitivity to the extraction variant. The candidate-hidden column above uses the single-call extraction over all passages, which BEV also uses. Taking the per-passage maximum instead leaves a small residual on judge-confirmed errors ($+0.036$, p=0.26) and a larger one on the raw stubborn subset ($+0.299$). A maximum over passages amplifies the asymmetry that the model's own answer is the more extractable of the two, since it only has to be found in one passage, another reason to prefer the single-call form beyond its lower cost.

E Decomposing candidate sensitivity

Bias by the verifier's confidence, in full. The quintile rows dropped from Table 2 for space:

	n	candidate-conditioned (yes/no)		candidate-free reading (extract)	
		bias	p	bias	p
Q1 (least confident)	2,126	-0.106	1.00	-0.031	0.99
Q2	2,128	$+0.010$	0.31	-0.033	0.98
Q3	2,126	$+0.067$	$1 \cdot 10^{-3}$	-0.035	0.94
Q4	2,127	$+0.118$	$2 \cdot 10^{-7}$	$+0.002$	0.48
Q5 (most confident)	2,127	$+0.122$	$6 \cdot 10^{-7}$	$+0.086$	$1 \cdot 10^{-3}$

The decomposition by confidence quintile. A is the verifier's score on its own wrong answer, B on another model's wrong answer to the same question, C on a random wrong answer taken from a different question of the same relation. **Units differ from the table above**: the three-way decomposition is reported in the within-(verifier, benchmark) SD of the *raw* yes/no log-odds over wrong answers, a scale shared by all three terms, because the identity $A - C = (B - C) + (A - B)$ holds only under a common denominator. The quintile table above and Δ_R of Section 4.2 instead divide by the SD of the *paired difference*, so the two sets of magnitudes are not directly comparable.

Confidence quintile	n	ownership $A-B$	relevance $B-C$	total $A-C$
Q1 (least confident)	2,126	-0.127	$+0.274$	$+0.147$
Q2	2,128	-0.009	$+0.309$	$+0.300$
Q3	2,126	$+0.063$	$+0.295$	$+0.358$
Q4	2,127	$+0.116$	$+0.280$	$+0.396$
Q5 (most confident)	2,127	$+0.127$	$+0.324$	$+0.451$
all	10,634	$+0.034$	$+0.297$	$+0.330$

The relevance term varies over $[0.274, 0.324]$ with no trend while the ownership term moves monotonically from -0.127 to $+0.127$: what tracks the model's confidence is the ownership component, not the topical-plausibility component. (Row sums differ from $A-C$ by at most 0.001 due to independent rounding.)

Per verifier and benchmark, same units:

	all wrong answers			stubborn subset		
Verifier / benchmark	n	$A-B$	$B-C$	n	$A-B$	$B-C$
Llama-3.1-8B / EQ	2,411	$+0.113$	$+0.516$	139	$+0.382$	$+0.843$
Llama-3.1-8B / PopQA	1,235	$+0.083$	$+0.268$	103	$+0.381$	$+0.470$
Mistral-7B / EQ	2,421	$+0.148$	$+0.184$	185	$+0.294$	$+0.461$
Mistral-7B / PopQA	1,273	$+0.057$	$+0.157$	133	$+0.006$	$+0.064$
Qwen2.5-7B / EQ	2,186	-0.179	$+0.312$	56	$+0.511$	$+0.592$
Qwen2.5-7B / PopQA	1,108	-0.049	$+0.227$	17	$+0.034$	$+1.613$

The ownership term is positive in all six stubborn cells and significant in four; the two that are not are both on PopQA. Qwen2.5-7B is the informative case: it is *harsher* on its own answers overall on both benchmarks (-0.179, -0.049) and yet is the most self-preferring of the three on its stubborn EQ answers ($+0.511$). A verifier can be globally strict and still be lenient where it is committed.

F Is the target set an artifact of its own definition?

A wrong answer is called stubborn when four internal signals all fall inside the range of the model's correct answers, so the finding that those four signals cannot separate the class is close to tautological. Three checks separate the tautology from the phenomenon.

Held-out signals. We rebuild the set using only two or three of the four signals and evaluate the ones that played no part in the selection, together with signals that play no part in any definition. Median correct-vs-H AUROC over six model–benchmark pairs; the answer-probe column is one variant throughout, the probe read at the answer's last token (the other two read positions are within 0.01 of it):

Set defined by	held-out defining signals	EigenScore	sem. entropy	$P(\text{True})$	verb. conf	CoVe	answer probe
sc + pc	conf 0.485 / probe 0.463	0.434	0.441	0.660	0.598	0.590	0.686
conf + probe	sc 0.509 / pc 0.458	0.499	0.506	0.645	0.612	0.610	0.679
sc + pc + conf	probe 0.448	0.429	0.439	0.652	0.602	0.581	0.685
all four	—	0.439	0.445	0.660	0.608	0.593	0.663

Two things follow. A signal that took no part in the selection is at chance on the resulting set (0.448–0.509), so the failure is not produced by having thresholded that signal. And every external detector returns essentially the same number whether the set was carved out with two thresholds or four; adding thresholds does not manufacture difficulty. Discrete semantic entropy, computed here from the ten sampled answers and used nowhere in the paper's definitions, is at chance under all four definitions (0.439–0.506).

Confidence-matched ordinary errors. We pair every genuine stubborn error with an ordinary error of nearly identical token confidence (mean difference ≤ 0.0002) and re-run the same detectors:

Detector	correct vs. H	correct vs. confidence-matched ordinary	difference
answer-position probes	0.663–0.678	0.750–0.782	$+0.09$ to $+0.10$
EigenScore	0.439	0.545	$+0.107$
semantic entropy	0.445	0.537	$+0.092$
$P(\text{True})$	0.660	0.647	-0.013
verbalised confidence	0.608	0.590	-0.017
InterrogateLLM	0.564	0.550	-0.014
Chain-of-Verification	0.593	0.570	-0.023

The result splits in two. For detectors that read the model's internal state or its sampling distribution, the stubborn set is a genuinely harder region than ordinary errors of the same confidence: they recover 0.09–0.11 AUROC once confidence is held fixed. For detectors that ask the model a second question, there is no residual at all: their failure on this class is what token confidence alone already predicts. Stubbornness is therefore not a regime these methods fail in for reasons beyond confidence; our claim, that this is the region where verification is needed and where hypothesis-conditioned verification is most biased, does not depend on it. This matching fixes only one of the four defining dimensions, so a gap on signals correlated with the other three is partly definitional; the non-circular evidence is the held-out panel of Appendix F.

G How much of a deployed system's error is this class?

Judge-confirmed stubborn hallucinations are 0.3–16% of all errors depending on the benchmark (median 3% over eighteen model–benchmark configurations), which is small, and one of our pre-registered criteria failed on exactly this count. The relevant quantity for a deployed system is different: among the answers a system is most confident about, the ones it would actually serve, what is left after the ordinary errors have been filtered out. Errors remaining inside the top confidence deciles, as a share of the errors in that bucket (judge-confirmed genuine / all stubborn):

	all answers	top 20% confidence	top 10% confidence
Llama-3.1-8B / EntityQuestions	2.0% / 4.3%	15.2% / 35.7%	19.7% / 48.9%
Mistral-7B / EntityQuestions	2.3% / 5.1%	14.6% / 34.4%	19.6% / 50.3%
Llama-3.1-8B / PopQA	1.4% / 4.4%	10.4% / 33.2%	12.2% / 46.5%
Qwen2.5-7B / TriviaQA	2.0% / 4.1%	14.4% / 39.2%	20.6% / 52.9%
Llama-3.1-8B / Natural Questions	4.2% / 14.3%	12.8% / 56.3%	9.4% / 58.7%
Llama-3.1-8B / HaluEval	6.0% / 10.1%	29.2% / 53.6%	31.0% / 63.4%
Qwen2.5-7B / SimpleQA	10.8% / 12.7%	42.1% / 49.4%	49.9% / 58.8%

The class is rare relative to all errors and common relative to the errors a confident system still makes: in the top confidence decile, 16–63% of the remaining errors are stubborn (median 53%) and 5–53% are judge-confirmed genuine (median 17%); the table above lists seven of the eighteen configurations. Abstention measured on the whole answer population, not conditioned on H, shows the same asymmetry: on TriviaQA, abstaining on the least confident half cuts the overall error rate from 43% to 16% while *raising* the genuine stubborn rate among the answers kept from 0.86% to 1.73%. Whether that warrants the extra retrieval call is a deployment question with an explicit answer in our numbers: the router sends about a third of questions to retrieval, and a 1.5B verifier suffices, so the cost is roughly one small-model call per three questions.

H Trained entailment and grounding baselines

Two off-the-shelf discriminators scored identical passages and claims: a generic entailment model (DeBERTa-v3-base on MNLI+FEVER+ANLI, 184M) and a specialist grounding checker (MiniCheck-RoBERTa-Large, 355M; Tang et al., 2024). Over 21 model–corpus configurations across five benchmarks the generic model beats the FActScore-style prompt but beats the 7B yes/no verifier in only 1, and the specialist is below it in almost all. Both are hypothesis-conditioned by construction, since the candidate is their hypothesis, though they are also 20–100$\times$ smaller than the generative verifiers and trained on a different task format, so this comparison bounds what an off-the-shelf discriminator delivers here, without isolating hypothesis conditioning. One uncalibrated extraction pass beats the generic model in 11 of 21 (Appendix H).

Correct vs. H AUROC on identical passages and identical claims, over 21 model–corpus configurations. "extract" is the raw $\{-1, 0, +1\}$ extraction signal with no percentile calibration, included so that all columns are on the same footing.

	n_{cor}	n_H	DeBERTa-MNLI	MiniCheck	yes/no (7B)	FActScore-style	extract
Qwen2.5-7B / EQ DPR	5,149	43	0.844	0.743	**0.918**	0.726	0.868
Llama-3.1-8B / EQ DPR	6,419	98	0.858	0.768	**0.905**	0.825	0.867
Mistral-7B / EQ DPR	5,958	119	0.855	0.781	**0.909**	0.815	0.850
Qwen2.5-7B / PopQA gold	1,878	13	0.927	0.908	**0.945**	0.921	0.924
Llama-3.1-8B / PopQA gold	2,639	70	0.942	0.884	**0.970**	0.964	0.920
Mistral-7B / PopQA gold	2,576	70	0.933	0.884	**0.936**	0.905	0.888
Qwen2.5-7B / EQ search	264	42	0.666	0.661	**0.746**	0.667	0.736
Llama-3.1-8B / EQ search	394	98	0.719	0.650	**0.782**	0.709	0.746
Mistral-7B / EQ search	391	116	0.749	0.689	**0.803**	0.668	0.724
Qwen2.5-7B / PopQA search	397	24	0.590	0.557	0.642	0.618	**0.827**
Llama-3.1-8B / PopQA search	661	102	0.660	0.612	0.688	0.646	**0.708**
Mistral-7B / PopQA search	675	109	0.694	0.641	0.683	**0.702**	0.689
Qwen2.5-7B / HaluEval	360	225	0.584	0.547	**0.642**	0.610	0.563
Llama-3.1-8B / HaluEval	350	275	0.525	0.526	**0.591**	0.578	0.575
Mistral-7B / HaluEval	366	432	0.554	0.577	**0.661**	0.647	0.611
Qwen2.5-7B / TriviaQA	306	44	0.707	0.652	**0.753**	0.715	0.684
Llama-3.1-8B / TriviaQA	422	27	0.733	0.705	**0.802**	0.764	0.707
Mistral-7B / TriviaQA	401	56	0.584	0.598	**0.683**	0.643	0.685
Qwen2.5-7B / NQ	419	125	0.627	0.599	**0.665**	0.641	0.613
Llama-3.1-8B / NQ	471	169	0.598	0.577	**0.663**	0.644	0.601
Mistral-7B / NQ	470	139	0.618	0.604	0.624	0.604	**0.632**

The specialist grounding checker is the weakest column for a reason: MiniCheck is trained to decide whether a document supports a summary-style claim, and it does that well; here "supported" requires knowing the fact, and a short ⟨subject, relation, answer⟩ triple gives it little surface to align. The table shows that a trained discriminator, generic or specialist, does not remove the dependence we are measuring, since the candidate is still the hypothesis being scored, and does not by itself outperform a plain prompt on this contrast.

I Entity linking: what the reverse check assumes, and what brackets it

The reverse check is an entity-retrieval step reduced to its crudest form: take the answer string, look it up as a Wikipedia title (following redirects), and test whether the question's subject string occurs on the page. Two questions follow: how much the crude version loses against proper entity linking, and how much the substring test loses against a softer matcher. Both are measured here, not assumed.

The lookup procedure and its fallback. The answer string is first resolved as a page title. When that fails, we fall back to the Wikipedia search API and take the first hit. On PopQA the fallback fires on 9.2% of answer–subject pairs (301 of 3,281), and only 7% of those fallback lookups return a page whose title is consistent with the answer string; the fallback is poor. It is also immaterial: deleting every fallback item changes the reverse signal's correct-vs-H AUROC by less than 0.002 (Llama 0.695→0.694, Qwen2.5-14B 0.714→0.713). The signal rests on the exact-title lookups; the fallback neither helps nor hurts.

Naive and oracle entity linking bracket the corpus. Two corpora bound what entity-centric retrieval can deliver on PopQA. The lower bound is naive string-to-title linking on the *subject*: it retrieves a page for only 38% of questions, and on 82% of the covered questions the model answers correctly the page does not contain the answer, because the subject string resolves to a famous homonym. We report this as unusable and retract an earlier 0.72 AUROC we obtained from it, which was an artifact of mean-imputing the 62% of questions with no page. The upper bound is the gold entity's page, available for 68% of questions and correct by construction, an oracle since it uses the benchmark's own entity identifier. The realistic search corpus sits between them and is what all

held-out numbers use. The gap between the two bounds is the headroom a real entity linker (Cao et al., 2021; Wu et al., 2020; Shavarani & Sarkar, 2025) would compete for; we did not use one, and a reader should read our reverse-check numbers as a floor for that family of methods, not its best achievable value.

A looser matcher is worse, not better. A natural objection is that exact substring inclusion is brittle: aliases, inflections and reorderings should be absorbed. We compared five matchers on the full text of the 3,281 PopQA answer pages, averaged over six models (correct vs. H AUROC): exact substring 0.656, punctuation- and case-normalised substring 0.656, continuous word-coverage fraction 0.644, any shared bigram 0.631, surname/last-token match 0.608. Loosening the match makes the signal worse, and the hit rates say why: relaxation raises the hit rate on correct answers and on genuine hallucinations together (for Qwen2.5-14B, H from 0.20 to 0.45 while correct goes 0.63 to 0.71), so the contrast is diluted. The brittleness is the selectivity: requiring the subject's full name to appear is the evidence that the answer entity really stands in the queried relation.

J CALIBRATION WITHOUT TARGET-DOMAIN LABELS

Every stage of our score is a percentile among the model's *correct* answers, which presumes access to correct answers on the target domain. We measured what happens when that access is removed, on the search corpus, reporting the change in correct-vs-H AUROC against the oracle calibration (six model–dataset tags): calibrating on another dataset's correct answers costs -0.015 to $+0.002$; calibrating with *no labels at all*, by treating the model's top 10% or 20% most confident answers as if they were correct, costs -0.022 to $+0.031$ (it is better than the oracle for two tags); 50 labelled correct answers cost -0.003 to $+0.013$, and 200 are within $\pm.001$ of the full set. Calibration is therefore not a hidden label requirement: a practitioner with no annotations can use the confidence-based pseudo-calibration and lose at most about 0.02 AUROC, and fifty labels make the question moot.

Full-scale judge comparison. The 350-item study above perturbs the judge on a stratified sample; we also re-ran a second family (GLM-5.2) with the same protocol over every audited item. On the 9,770 items both judges labelled:

Benchmark	n	main H%	GLM H%	Δ	main G%	agreement (H vs. rest)
SimpleQA	1,599	86.4	88.4	+2.0	7.4	0.851
HaluEval QA	2,041	66.1	74.4	+8.3	16.1	0.875
TriviaQA	271	57.2	69.4	+12.2	23.2	0.871
EntityQuestions	1,095	53.3	61.0	+7.7	21.0	0.874
TruthfulQA	476	47.3	70.6	+23.3	27.7	0.733
PopQA	1,125	46.8	59.8	+13.1	41.9	0.818
AmbigQA	1,995	39.9	55.1	+15.2	37.1	0.822
Natural Questions	1,168	32.7	46.1	+13.4	40.5	0.824

Rank correlation between the two columns is $+0.929$ ($p = 0.001$); pooled agreement is 0.840 with $\kappa = 0.670$. The disagreement is directional, not random: the second judge's H share is higher on *every* benchmark, because it resolves many of the first judge's ambiguity and undetermined verdicts as genuine errors, while the reverse move is rare. TruthfulQA is the extreme case ($+23.3$ points, agreement 0.733), as one would expect of a benchmark built from adversarial misconceptions, where "genuine error" and "defensible reading of a contested claim" are hardest to separate. This table covers all audited items with a valid label from both judges, so its H shares are not comparable to Table 1, which reports the cross-model stubborn core. GLM-5.2 returned no parseable label on 3% of items after one retry (content-filter refusals and malformed JSON); those are excluded, not imputed.

K WHERE THE STRING MATCH BREAKS: ERROR ANALYSIS BY ANSWER TYPE

The "extract" column below is the *uncalibrated standalone* extraction score, not the deployed composite s_{fwd}; standalone extraction is not expected to beat yes/no on its own (Appendix D), so the

column shows where the extraction signal itself survives, not where BEV wins. We did not recompute the composite per answer type.

The final decision is made by extraction plus an alias/string match, which is brittle for answers that are not entity names. We classified gold answers into five types and recomputed, per type, the fraction of the model's *correct* answers the matcher recovers from the passage and the correct-vs-H AUROC of the two verifiers (EQ, PopQA, NQ and TriviaQA; 12 model–corpus configurations):

Answer type	n	n_H	match rate on correct answers	yes/no	extraction
Two-word person name	7,696	480	0.54	0.818	0.789
Long entity ($\geq$3 words)	4,652	185	0.52	0.834	0.751
Other short entity	9,457	169	0.47	0.787	0.734
Year or date	213	48	0.37	0.583	**0.645**
Pure number	287	64	0.22	0.602	0.566

The match rate falls monotonically from 0.54 on person names to 0.22 on pure numbers, and on numeric answers both verifiers collapse towards chance. This is a real limit of the construction and we name it as one: the method is a verifier for *entity-valued* answers, and on quantities it inherits the matcher's weakness. One detail runs the other way and is worth noting: on dates the extraction signal is *better* than the yes/no prompt (0.645 against 0.583), presumably because a date either appears in the passage or does not, whereas asking a model whether a date is "supported" invites it to reason around near-misses. The numeric and date rows rest on 287 and 213 comparison items from three configurations each, so their intervals are wide and we read them as a direction, not a point estimate. The practical rule is the same one the extractable rate gives: measure the matcher's hit rate on a handful of known-correct answers of the type you care about, and do not deploy the method where it is near zero.

Does the method comparison depend on the judge? Separately from the shares, we re-ran the verifier comparison itself under three definitions of the target: H according to the main judge, according to the second judge, and according to both (the high-precision intersection). Over twelve model–corpus configurations the forward verifier's margin over yes/no is essentially unchanged, 9/12 wins with mean $+0.023$ (main), 10/12 with $+0.017$ (second) and 10/12 with $+0.023$ (intersection), and the per-configuration differences agree almost cell by cell (for example $+0.017/+0.016/+0.015$ on Qwen2.5-7B with DPR passages, and $-0.009/-0.001/-0.009$ on Llama-3.1-8B with search passages). The judge determines how large this population is; it does not determine which form of verification works better on it.

L Sensitivity of the fusion weight, and whether it can be chosen without labels

The forward score is $s_{\mathrm{fwd}} = \mathrm{Pc(yes/no)} + w\,m$ with $m \in \{-1, 0, +1\}$ and w=0.5 fixed in advance. Correct-vs-H AUROC over the five settings of w, twelve model–corpus configurations:

	w=0	w=0.25	w=0.5	w=0.75	w=1
mean over 12 configurations	0.828	**0.848**	**0.848**	0.846	0.845
configurations above w=0	—	11/12	9/12	8/12	8/12

The curve is flat over $[0.25, 1]$ and every positive weight beats w=0: the extraction term only needs to enter at all; its exact weight barely matters. Against a per-configuration oracle choice of w, the fixed w=0.5 gives up 0.006 AUROC on average and 0.014 at worst.

We also asked whether w can be chosen without labels. The answer is no; we report this because it bears on how the method should be deployed. The rule we designed uses only the model's correct answers, the same data the percentile calibration already needs, and picks the w that maximally separates the three levels of m (a between-over-within variance ratio). It selects w=1 in every configuration and gives up 0.009 on average and 0.025 at worst against the per-configuration oracle, i.e. *worse* than simply fixing w=0.5. The practical recommendation is to fix w, not select it. w=0.25

would also have been marginally better than our pre-registered w=0.5; we do not change it, because choosing it now would be exactly the test-set tuning this analysis is meant to rule out.

M IS THE EFFECT OWNERSHIP, OR A FAMILY PRIOR?

The matched-pair design compares a verifier's score on its own wrong answer against another model's wrong answer. A verifier that simply preferred, say, Qwen-style answers would produce a similar table. The control is to ask the same verifier to compare two candidates *neither* of which is its own. For each verifier R and the two other models M_1, M_2, we score M_1's and M_2's wrong answers to the same question under R:

$$\Delta_R^{\text{3rd}}(x) \;=\; \frac{V_R(x, a_{M_1}, e) - V_R(x, a_{M_2}, e)}{\sigma_R},$$

with M_1, M_2 fixed in a canonical model order, so the sign of the offset reflects that order, not any preference. The comparison is in the main text (Table 2, middle block).

The third-party column is not zero overall, which is expected: M_1 and M_2 are two fixed models and a constant difference in their answer quality or style shifts the whole comparison. This nuisance effect is *flat* across subsets (-0.105 on ordinary errors, -0.042 on stubborn), while the ownership effect *grows* by an order of magnitude over the same subsets ($+0.017 \rightarrow +0.291$). On the subsets where ownership peaks, a verifier shows no systematic preference between two candidates that are not its own. The effect is therefore specific to having produced the answer, not to a family prior or to generic anchoring.

N IS THE RESIDUAL JUST DIFFERENTIAL EXTRACTABILITY?

Section 4.2 attributes the residual bias of the extraction verifier on the raw stubborn subset to the model's own answer being the more extractable of the two. That is a testable statement. For every matched pair we record whether each candidate occurs literally in the evidence (normalised whole-string match) and whether the reader's extracted answer equals it:

	n	literal occurrence in evidence			extracted answer matches		
		own	other	diff	own	other	diff
ordinary wrong answers	22,592	0.038	0.051	-0.013	0.009	0.012	-0.003
stubborn (all)	633	0.198	0.111	$\mathbf{+0.087}$	0.066	0.027	$\mathbf{+0.039}$
not judge-confirmed	355	0.220	0.107	$+0.113$	0.101	0.034	$+0.068$
judge-confirmed (H)	278	0.169	0.115	$+0.054$	0.022	0.018	$+0.004$

On ordinary errors the model's own answer is slightly less likely to appear in the passage. On the stubborn subset it is $1.8\times$ as likely to appear and $2.4\times$ as likely to be the extracted answer, which is the asymmetry behind the $+0.226$ residual reported there; the gap is carried almost entirely by the items the audit does *not* confirm as genuine. On judge-confirmed genuine errors the extracted-match gap is $+0.004$ (essentially nil), which is where the extraction bias is also absent (-0.026, p=0.64; Table 3). The confound is real, measurable, and removed by the audit, not by the prompt.

Does the ownership effect survive the lexical control? Partly. Literal occurrence could itself drive the yes/no bias, since a verifier may say "yes" more readily when the candidate appears verbatim in the passage, so we recompute the bias on the pairs where *neither* candidate occurs literally (89% of pairs), and also measure it on the composite score s_{fwd} instead of on raw yes/no. Both candidates are still scored by the same verifier on the same evidence.

	n	yes/no bias		bias on s_{fwd}	
		bias	p	bias	p
all wrong answers	10,634	$+0.042$	$2 \cdot 10^{-9}$	$+0.016$	$3 \cdot 10^{-4}$
neither candidate literal	9,454	$+0.047$	$8 \cdot 10^{-11}$	$+0.016$	$8 \cdot 10^{-5}$
stubborn	633	$+0.273$	$2 \cdot 10^{-5}$	$+0.279$	$7 \cdot 10^{-5}$
neither candidate literal	464	$+0.082$	0.06	$+0.038$	0.14
judge-confirmed (H)	278	$+0.110$	0.02	$+0.038$	0.11
neither candidate literal	207	$+0.074$	0.13	$+0.006$	0.39

Three readings; the second is a limitation. *(i)* On the whole wrong-answer population the effect survives the control and is slightly larger ($+0.047$ against $+0.042$), so ownership leniency is not an artifact of lexical priming at population scale. *(ii)* On the stubborn subset it is: the large figure falls from $+0.273$ to $+0.082$ once literal occurrence is removed, and on judge-confirmed errors from $+0.110$ to $+0.074$, neither significant at these sample sizes. The headline subset numbers of Table 2 should therefore be read as *ownership together with the lexical asymmetry that accompanies it*, not as ownership alone. *(iii)* Measured on the composite score the effect on judge-confirmed errors drops to $+0.038$ (n.s.) and to $+0.006$ under the lexical control, which is the sense in which the construction reduces the bias it was motivated by: a reduction, measured on the score we actually deploy, not an elimination.

O Pooled analysis with benchmark clustering

Model–dataset configurations are not independent: four configurations share a benchmark and several share a generator. A sign test over 29 such cells overstates the evidence, so we report three things instead. The distribution of paired differences against the yes/no verifier is min -0.019, Q_1 $+0.005$, median $+0.017$, Q_3 $+0.022$, max $+0.093$; 24 of 29 are positive, seven have bootstrap 95% intervals entirely above zero and *none* entirely below. Averaging within each benchmark and treating the seven benchmarks as the units of analysis, the most conservative reading, gives a mean of $+0.016$ (t=3.39, two-sided p=0.015), with every benchmark mean positive (PopQA $+0.041$, EQ $+0.019$, TriviaQA $+0.017$, AmbigQA $+0.015$, SimpleQA $+0.008$, HaluEval $+0.008$, NQ $+0.003$). A hierarchical bootstrap that resamples benchmarks and then configurations within benchmarks gives $+0.0168$ with a 95% interval of $[+0.0075, +0.0296]$. Dropping the three configurations with the widest intervals lowers the mean to $+0.0125$ (21 of 26 positive) and dropping six lowers it to $+0.0114$ (18 of 23): the small-n_H cells inflate the estimate somewhat, and the effect survives their removal.

P The reverse check by relation

The reverse bit tests whether the question's subject occurs on the answer entity's page, which is entity co-occurrence in an answer-centric source, not a test of the queried relation. Its value is therefore relation-dependent, and we measure it (hit rate on the model's *correct* answers, on judge-confirmed genuine errors, and on ambiguous items):

Relation	n_{correct}	hit | correct	hit | H	hit | G	gap
PopQA director	677	0.91	0.41	0.87	$+0.50$
PopQA author	667	0.85	0.31	0.86	$+0.54$
PopQA screenwriter	529	0.87	0.47	0.82	$+0.40$
PopQA father	397	0.66	0.22	0.75	$+0.44$
EQ P50 (author)	330	0.73	0.10	0.55	$+0.63$
PopQA producer	311	0.91	0.72	0.90	$+0.19$
EQ P175 (performer)	140	0.66	0.05	0.63	$+0.61$
EQ P69 (educated at)	113	0.02	0.00	0.00	$+0.02$

Three things follow. Ambiguous items behave like correct answers (0.55–0.90), which is the mechanism behind the cost we already report: the reverse term rewards them by design. "Educated at" fails completely: a person's name essentially never appears on a university's page, so the bit carries no signal at all (0.02 on correct answers), which is the reviewer-anticipated case where co-occurrence

and the queried relation come apart. And "producer" is the weakest discriminator among the relations that do work ($+0.19$), because producers co-occur with many works. A high hit rate on *correct* answers is *necessary* for the check to be usable, and that much is visible without audit labels, the same label-free screen as the extractable rate. It is not sufficient: *producer* matches *director* at 0.91 on correct answers yet separates far worse ($+0.19$ against $+0.50$), because the subject also sits on the page of a wrong producer 0.72 of the time. Ruling out that kind of collision needs the hit rate on errors, and that does need labels.

Q MATCHER AND INFERENCE DETAILS

The external comparison $m = \mathrm{match}(\hat{a}, a_0) \in \{+1, 0, -1\}$ decides every forward verdict, so we give it in full. Both strings are lower-cased; punctuation and leading articles ("the", "a", "an") are stripped; whitespace is collapsed. The comparison is whole-word and bidirectional: $\hat{a}$ matches a_0 if either is a contiguous whole-word subsequence of the other, so "Spielberg"↔"Steven Spielberg" matches while "Berg" does not. A reply is treated as an abstention, giving $m{=}0$, if it is empty or contains any of *not in passage*, *unknown*, *none*, *no answer*, *n/a*. A multi-span reply is split on commas and on "and"/"or" and gives $m{=}+1$ if any span matches. The benchmark's gold aliases are used only to decide correctness when constructing the audit; at inference the matcher sees the two strings and nothing else, so the method does not assume an alias table at deployment. Decoding is greedy, 24 tokens maximum, all retrieved passages in a single call; the reader may be any model, including one much smaller than the generator. The percentile map $\mathrm{Pc}(\cdot)$ returns a value in $[0, 1]$: for wrong answers it is the fraction of the model's correct answers scoring no higher, and for correct answers it is computed with five-fold cross-fitting so that an item is never ranked against a reference set containing itself. The reverse bit is keyed on the pair (question subject, candidate a_0), enters the final score with weight 1.0, and is granted identically to every baseline in the comparisons that use it. The resulting ranges are $\mathrm{Pc} \in [0, 1]$, $w\,m \in \{-0.5, 0, +0.5\}$, $r \in \{0, 1\}$, hence $s_{\mathrm{fwd}} \in [-0.5, 1.5]$ and $s \in [-0.5, 2.5]$; all four endpoints are attained in practice. With $w{=}0.5$ the three bands of m overlap by design, so that an item with $m{=}-1$ and a high percentile can outrank one with $m{=}0$ and a low percentile, whereas at $w{=}1$ they separate completely and the extraction term overrides the calibrated score, so ordinary errors whose passage does not contain the answer would never be flagged.

R AGGREGATION OPERATOR

Every stage i is mapped to a percentile $\mathrm{Pc}_i \in [0, 1]$ (written with the same Pc as in Section 5, and distinct from the passage set P) by the same rule: the rank of its raw score among the model's correct answers on the calibration split, so Pc_i is the fraction of correct answers the stage scores no higher than this item. For the evidence stage the raw score is s_{fwd} of Section 5; for the free signals it is the signal itself. Stages then compose by the minimum of their percentiles, $s_{\mathrm{SRV}} = \min_i \mathrm{Pc}_i$, a different score from the additive s of Section 5. The choice is conservative, not optimal, and we check that the conclusion does not depend on it. Over six model–corpus configurations, correct-vs-H AUROC and the residual genuine-hallucination rate among answers kept at 50% coverage:

Operator	AUROC	residual H at 50% coverage
$\min_i \mathrm{Pc}_i$ (used)	0.525	**0.099**
$\frac{1}{2}\min_i \mathrm{Pc}_i^{\text{cheap}} + \frac{1}{2}\mathrm{Pc}^{\text{evidence}}$	**0.597**	**0.097**
$\mathrm{mean}_i \mathrm{Pc}_i$	0.515	0.108
$\max_i \mathrm{Pc}_i$	0.474	0.099

The minimum and a weighted average are within 0.002 of each other on the deployment metric and both beat the mean; on AUROC the weighted average is better. We keep the minimum because it makes the safety property explicit, that an added evidence stage can only lower a score; we do not claim it is the best aggregator.

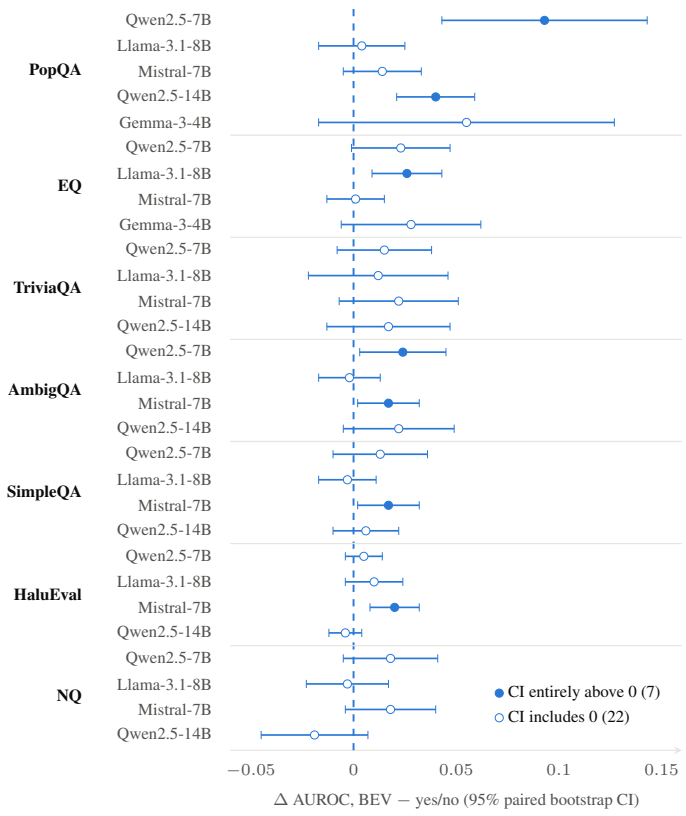


Figure 4: **Per-configuration gain on the search corpus.** Paired bootstrap difference in correct-vs-H AUROC between BEV and the yes/no verifier for all 29 configurations of Table 4, with 95% intervals; both verifiers receive the candidate-page bit where it applies. Seven intervals lie entirely above zero, none entirely below, and 24 of 29 point estimates are positive.

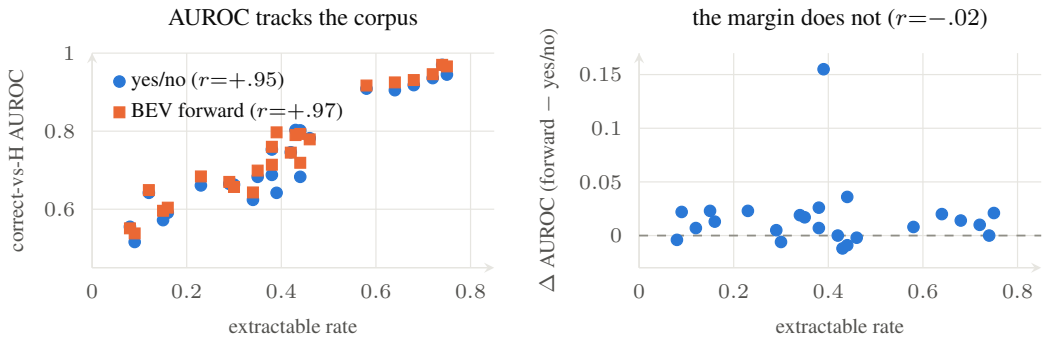


Figure 5: **The corpus sets the ceiling, not the margin.** Twenty-four model–corpus configurations. *Left*: absolute correct-vs-H AUROC of the yes/no verifier and of the forward term against the corpus's extractable rate, the fraction of correct answers recoverable from its passages. *Right*: the difference between the two against the same quantity; it is flat (r=−0.02, p=0.94) and positive on average on both sides of an extractable rate of 0.3.

S Additional results

Data and models in full. PopQA (Mallen et al., 2023) (14,267), EntityQuestions (Sciavolino et al., 2021) (22,075), TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), Am-

Table 5: Held-out evaluation on the realistic search corpus (the top-1 Wikipedia search result for the question text). Nothing was tuned on this corpus. Correct vs. H AUROC; "extr." is the fraction of correct answers recoverable from the passage; best per row in bold; 95% CI shown for the bidirectional score. The reverse term needs a subject entity and a direct subject–answer relation, so it is reported for the two template benchmarks. The "yes/no+r" column is the baseline of Table 4, the yes/no verifier granted the same candidate-page bit, so the benchmark means reported there are recoverable from this table. Recall is H recall at 10% FPR, yes/no → forward → bidirectional. "—" marks a cell where the reverse check does *not apply* by construction (no subject entity, or no direct subject–answer relation to test on a page); it is not a missing measurement, and the corresponding recall column then has two stages instead of three. The one exception is Gemma-3-4B on EQ, where the check does apply but we did not fetch the answer-entity pages at full scale; Table 7 reports it on the subset that has them.

	Model	n_H	extr.	$P(\text{True})$	yes/no	yes/no+r	forward	reverse	BEV bidir.	recall
PopQA	Qwen2.5-7B	24	0.39	0.713	0.641	0.729	0.764	0.667	**0.823** [0.75,0.89]	0.12→0.29→0.54
	Llama-3.1-8B	102	0.38	0.688	0.688	0.823	0.699	0.750	**0.827** [0.78,0.87]	0.22→0.27→0.58
	Mistral-7B	109	0.35	0.622	0.682	0.755	0.700	0.674	**0.769** [0.72,0.82]	0.22→0.30→0.46
	Qwen2.5-14B	69	0.38	0.753	0.672	0.816	0.736	0.744	**0.856** [0.81,0.90]	0.10→0.33→0.67
	Gemma-3-4B	70	0.40	0.657	0.645	0.652	0.684	0.608	**0.708** [0.61,0.82]	0.03→0.14→0.35
EQ	Qwen2.5-7B	42	0.42	0.642	0.743	0.829	0.762	0.711	**0.853** [0.79,0.90]	0.21→0.33→0.52
	Llama-3.1-8B	98	0.46	0.715	0.780	0.835	0.792	0.717	**0.861** [0.81,0.90]	0.31→0.36→0.53
	Mistral-7B	116	0.43	0.700	0.802	0.894	0.797	0.787	**0.895** [0.86,0.93]	0.39→0.40→0.64
	Gemma-3-4B	63	0.46	0.562	0.760	0.760	**0.788**	—	—	0.14→0.40
TriviaQA	Qwen2.5-7B	54	0.41	**0.779**	0.755	—	0.770	—	—	0.24→0.24
	Llama-3.1-8B	31	0.47	0.775	0.807	—	**0.818**	—	—	0.29→0.39
	Mistral-7B	66	0.47	0.710	0.696	—	**0.717**	—	—	0.24→0.24
	Qwen2.5-14B	8	0.35	**0.762**	0.657	—	0.675	—	—	0.00→0.12
AmbigQA	Qwen2.5-7B	266	0.30	0.681	0.668	—	**0.692**	—	—	0.14→0.25
	Llama-3.1-8B	274	0.31	0.633	**0.670**	—	0.669	—	—	0.17→0.18
	Mistral-7B	350	0.30	0.616	0.635	—	**0.652**	—	—	0.13→0.18
	Qwen2.5-14B	81	0.31	0.608	0.632	—	**0.654**	—	—	0.21→0.20
HaluEval	Qwen2.5-7B	225	0.12	**0.661**	0.641	—	0.646	—	—	0.15→0.19
	Llama-3.1-8B	275	0.16	0.588	0.591	—	**0.601**	—	—	0.11→0.14
	Mistral-7B	432	0.23	0.671	0.660	—	**0.680**	—	—	0.22→0.27
	Qwen2.5-14B	227	0.10	**0.691**	0.633	—	0.628	—	—	0.15→0.15
NQ	Qwen2.5-7B	125	0.29	0.634	0.665	—	**0.683**	—	—	0.18→0.25
	Llama-3.1-8B	169	0.30	0.642	**0.662**	—	0.659	—	—	0.17→0.22
	Mistral-7B	139	0.34	0.636	0.624	—	**0.642**	—	—	0.10→0.13
	Qwen2.5-14B	47	0.31	**0.607**	0.603	—	0.585	—	—	0.13→0.11
SimpleQA	Qwen2.5-7B	306	0.09	0.523	0.517	—	**0.529**	—	—	0.12→0.14
	Llama-3.1-8B	388	0.08	0.506	**0.553**	—	0.550	—	—	0.07→0.09
	Mistral-7B	526	0.15	0.544	0.574	—	**0.591**	—	—	0.10→0.14
	Qwen2.5-14B	224	0.08	0.499	0.501	—	**0.507**	—	—	0.16→0.17

bigQA (Min et al., 2020), SimpleQA (Wei et al., 2024a), HaluEval QA (Li et al., 2023), TruthfulQA (Lin et al., 2022); Qwen2.5-Instruct at 1.5B/3B/7B/14B/32B, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Gemma-3-4B-it, plus Qwen2.5-7B base with 5-shot prompting, four model families in all. Paraphrases are hand-written templates for the two template benchmarks and generated by an external model and similarity-filtered (cosine 0.80–0.995 against the original) for natural questions. Questions after the strict subset: PopQA 8,767 of 14,267; EntityQuestions 6,380 of 22,075; TriviaQA 7,644; Natural Questions 5,907; AmbigQA 11,816; SimpleQA 4,198; HaluEval QA 9,690; TruthfulQA 787 of 817 (the rest lacking usable paraphrases). Accuracy ranges over models: PopQA 7–19%, EQ 12–22%, TriviaQA 57–71%, NQ 26–35%, AmbigQA 32–42%, HaluEval 27–34%, SimpleQA 3.5–5.5%, TruthfulQA 13–20%. Stubborn errors as a share of all errors at q=25: 0.9–4.9% (PopQA), 1.7–5.1% (EQ), 4.1–4.8% (TriviaQA), 10.0–14.3% (NQ), 9.6–12.6% (AmbigQA), 9.0–15.9% (HaluEval), 12.7–18.8% (SimpleQA) and 23–40% (TruthfulQA, the highest of the eight: a benchmark built from misconceptions people repeat confidently elicits errors the model also repeats confidently). Gemma-3-4B, run with the three-signal variant, gives 4.3% on PopQA and 7.2% on EQ. The per-model table is Table 13.

Criterion C per dataset (genuine errors only). Median over model pairs of observed/expected co-occurrence, with the range: PopQA 13.8× (8.1–27.3, 15 pairs), EntityQuestions 5.6× (2.7–

Table 6: Selective prediction on the search corpus: abstaining on the lowest-scoring items, the share of *answers still given* that are genuine stubborn hallucinations (in parentheses: ordinary hallucinations). Lower is better. Abstaining by token confidence makes the residual stubborn rate *worse*, because it removes ordinary errors and concentrates this class among the answers kept. "—" marks the no-detector row, where no abstention rule is defined and the coverage columns do not apply.

	answer rate	100%	90%	80%	70%	50%
PopQA Llama-3.1-8B (n_H=102)	no detector	6.9 (21.0)	—	—	—	—
	confidence (free)		7.6 (13.3)	8.6 (7.8)	9.2 (5.1)	**9.3** (2.2)
	$P(\text{True})$		**6.4** (19.2)	6.5 (15.6)	**6.3** (11.3)	**5.3** (5.3)
	yes/no retrieval		6.8 (20.3)	6.7 (19.1)	6.2 (17.1)	6.6 (11.9)
	BEV forward		7.0 (20.1)	6.2 (20.0)	**6.0** (19.1)	6.3 (12.9)
	BEV bidirectional		**6.4** (18.8)	**6.1** (15.4)	6.2 (11.0)	**5.3** (6.3)
EQ Mistral-7B (n_H=116)	no detector	11.6 (25.5)	—	—	—	—
	confidence (free)		12.9 (18.3)	14.5 (11.6)	16.6 (7.6)	**15.6** (5.0)
	$P(\text{True})$		12.3 (21.4)	12.4 (17.5)	12.3 (14.4)	9.2 (10.4)
	yes/no retrieval		11.1 (23.6)	10.1 (21.2)	10.0 (17.9)	8.0 (12.2)
	BEV forward		10.8 (23.4)	10.1 (21.8)	9.4 (19.2)	8.6 (12.4)
	BEV bidirectional		**10.4** (22.3)	**9.7** (20.5)	**8.9** (15.4)	**7.0** (7.2)

Table 7: Fair comparison: the candidate-page bit added to *every* verifier with the same weight 1.0, on the search corpus restricted to the questions that have an answer-entity page. Correct vs. H AUROC, without $\rightarrow$ with the bit; best per row in bold. All columns are percentile-calibrated (five-fold held out on correct answers), as in Table 5. The bit is an orthogonal signal that improves every verifier on these benchmarks; with it granted to all of them our forward term still gives the best combination in 8 of 9 pairs (Mistral-7B on PopQA goes to the FActScore-style verifier, and Mistral-7B on EQ ties with yes/no).

Model	(n, n_H)	$P(\text{True})$	FActScore	yes/no	extract only	extract $\oplus$ yes/no
PopQA Qwen2.5-7B	(518, 24)	0.710→0.774	0.618→0.718	0.634→0.730	0.816→0.797	0.774→**0.826**
PopQA Llama-3.1-8B	(817, 102)	0.726→0.819	0.654→0.802	0.693→0.823	0.712→0.823	0.705→**0.826**
PopQA Mistral-7B	(931, 109)	0.648→0.726	0.725→**0.779**	0.693→0.756	0.667→0.738	0.712→0.772
PopQA Qwen2.5-14B	(729, 69)	0.772→0.852	0.722→0.818	0.668→0.816	0.787→0.849	0.732→**0.856**
PopQA Gemma-3-4B	(367, 23)	0.484→0.605	0.547→0.628	0.567→0.650	0.642→0.684	0.637→**0.714**
EQ Qwen2.5-7B	(377, 42)	0.646→0.741	0.673→0.758	0.758→0.828	0.769→0.815	0.773→**0.854**
EQ Llama-3.1-8B	(463, 98)	0.751→0.813	0.736→0.794	0.815→0.836	0.727→0.803	0.826→**0.859**
EQ Mistral-7B	(521, 116)	0.749→0.862	0.685→0.835	0.825→**0.893**	0.704→0.852	0.824→**0.893**
EQ Gemma-3-4B	(192, 16)	0.501→0.699	0.670→0.773	0.787→0.811	0.695→0.785	0.799→**0.834**
mean gain from the bit		+0.100	+0.097	+0.078	+0.070	+0.072

14.6, 6 pairs), TriviaQA 11.0× (0–32.4, three pairs with <5 counts), Natural Questions 9.0× (5.6–24.7), AmbigQA 7.9× (4.0–17.2), HaluEval 4.1× (3.1–6.4), SimpleQA 2.8× (1.9–4.5, 10 pairs). For G: PopQA 17.5×, EQ 23.4×, TriviaQA 24.4×, NQ 7.2×, AmbigQA 9.0×, HaluEval 18.2×, SimpleQA 34.9×.

Reader × generator (18 configurations). Correct vs. H AUROC, yes/no → forward (w=0.5), on the questions every reader scored and with the percentile map fitted on that subset, so the levels sit a point or two off Table 9; these are the numbers behind the +0.012 / +0.020 split quoted in Section 5.1. EQ: Qwen answers read by Qwen/Llama/Mistral 0.918→0.931, 0.927→0.942, 0.889→0.933; Llama answers 0.904→0.929, 0.905→0.925, 0.887→0.913; Mistral answers 0.905→0.920, 0.914→0.930, 0.909→0.917. PopQA gold page: Qwen answers 0.945→0.966, 0.947→0.954, 0.955→0.961; Llama answers 0.910→0.960, 0.970→0.970, 0.951→0.970; Mistral answers 0.918→0.937, 0.949→0.946, 0.936→0.946. The forward term is at or above yes/no in 17 of the 18 (Mistral answers read by Llama on PopQA is −0.003), and FActScore-style is below yes/no in all 18.

Weight w. With w=1 the three extraction levels do not overlap and ordinary-error recall at FPR 10% drops by 10–23 points on the gold page (Llama 0.96→0.73). With w=0.5: AUROC ≥ yes/no

Table 8: Retrieval quality sweep on EntityQuestions. Each row is restricted to the questions for which *that* corpus returned a passage; the percentile calibration is the same five-fold held-out rule used everywhere else, so the DPR 1–3 and search top-1 rows coincide with the corresponding entries of Tables 9 and 5. The corpora are the shipped DPR passages taken at increasing rank depth, plus the search corpus. Quality is measured as the fraction of *correct* answers that can be extracted from the passages. Correct vs. H AUROC. The realistic search corpus is not the weakest setting; the advantage of the forward term appears wherever retrieval carries signal and vanishes at ranks 18–20, where it carries almost none.

Model	Passages	extractable	FActScore	yes/no	extract	forward (ours)
Qwen2.5-7B	DPR 1–3	0.74	0.757	0.908	0.890	**0.931**
	DPR 4–6	0.58	0.698	0.808	0.812	**0.857**
	DPR 11–13	0.41	0.681	0.756	0.717	**0.788**
	DPR 18–20	0.32	0.650	0.692	0.695	**0.701**
	search top-1	0.42	0.666	0.743	0.679	**0.762**
Llama-3.1-8B	DPR 1–3	0.75	0.838	0.906	0.874	**0.921**
	DPR 4–6	0.62	0.764	0.828	0.804	**0.835**
	DPR 11–13	0.43	0.632	0.707	0.733	**0.739**
	DPR 18–20	0.33	0.650	**0.696**	0.611	0.688
	search top-1	0.46	0.710	0.780	0.769	**0.792**
Mistral-7B	DPR 1–3	0.73	0.808	0.903	0.871	**0.922**
	DPR 4–6	0.60	0.749	0.830	0.818	**0.857**
	DPR 11–13	0.41	0.688	0.734	**0.771**	0.751
	DPR 18–20	0.30	0.661	**0.732**	0.687	0.730
	search top-1	0.43	0.667	**0.802**	0.684	0.797

Table 9: Verifiers on good passages: correct vs. H AUROC, same model and same passages. EQ uses the DPR top-3 shipped with the benchmark; PopQA the gold entity's page (an oracle, available for 68% of questions). "extractable" is the fraction of *correct* answers recoverable from the passages; the last column is H recall at 10% false-positive rate, yes/no → ours. The forward verifier is above yes/no on all ten model–corpus pairs.

Corpus	Model	n_H	extractable	FActScore	yes/no	extract	BEV fwd.	H recall
EQ DPR top-3	Qwen2.5-7B	43	0.74	0.757	0.908	0.890	**0.931**	0.60→0.77
	Llama-3.1-8B	98	0.75	0.838	0.906	0.874	**0.921**	0.66→0.74
	Mistral-7B	119	0.73	0.808	0.903	0.871	**0.922**	0.76→0.77
	Gemma-3-4B	267	0.76	0.863	0.918	0.886	**0.940**	0.67→0.87
PopQA gold page	Qwen2.5-7B	13	0.73	0.897	0.909	0.905	**0.947**	0.54→0.85
	Llama-3.1-8B	70	0.74	0.964	0.961	0.931	**0.970**	0.87→0.93
	Mistral-7B	70	0.69	0.902	0.927	0.862	**0.940**	0.77→0.79
	Qwen2.5-14B	47	0.77	**0.955**	0.892	0.895	0.947	0.47→0.85
	Qwen2.5-32B	40	0.71	0.899	0.942	0.876	**0.945**	0.82→0.85
	Gemma-3-4B	107	0.71	0.927	0.935	0.812	**0.941**	0.72→0.78

on 6/6 pairs (0.917/0.926/0.938/0.944/0.970/0.937), H recall $\geq$ on 6/6, ordinary recall within $-4/+8$ points. w=0.25 is similar. w was fixed at 0.5 before the held-out corpus was fetched.

SRV frontier (gold page, questions with a page). Triples are AUROC over all errors / H recall at 10% FPR / ordinary-error recall at 10% FPR, with the confidence gate. Llama-8B: full retrieval 0.980 / 0.86 / 0.96; SRV q=0.10 (49% routed) 0.952 / 0.79 / 0.87; q=0.25 (31%) 0.924 / 0.69 / 0.77; a random gate at those two budgets reaches H AUROC 0.664 and 0.562 at H recall 0.36 and 0.23. Mistral-7B: full 0.968 / 0.77 / 0.92; q=0.10 (61%) 0.944 / 0.77 / 0.80; q=0.25 (40%) 0.910 / 0.77 / 0.67; random 0.704/0.50 and 0.644/0.41. Qwen-32B: full 0.984 / 0.82 / 0.99; q=0.10 (49%) 0.962 / 0.72 / 0.87; q=0.25 (34%) 0.936 / 0.75 / 0.79; random 0.710/0.45 and 0.609/0.35. EQ (DPR): Llama full 0.939 / 0.67 / 0.83, q=0.10 (56%) 0.910 / 0.55 / 0.70; Mistral full 0.910 / 0.79 / 0.74, q=0.10 (61%) 0.903 / 0.71 / 0.73; Qwen full 0.928 / 0.65 / 0.73, q=0.25 (42%) 0.892 / 0.56 / 0.64. Over these six configurations, a $\min(\text{conf}, \text{sc})$ gate at q=0.25 routes 27–40% of questions and covers H 89–100%, correct answers 66–71%, ordinary errors 4–7%.

Table 10: **Reliability of the audit.** Share of genuine hallucinations (H) among the 323 of 350 stratified stubborn items (50 per benchmark) that all four configurations returned a valid label for, under three judge families and one rewritten protocol (B). The *ranking* is stable (Spearman 0.96 against the main judge for all three); the absolute level is not; individual cells move by 3–20 points. Agreement is H vs. rest against the main judge, which cannot be compared against itself.

Judge (protocol)	SimpleQA	HaluEval	TriviaQA	EQ	PopQA	AmbigQA	NQ	agreement	U share
DeepSeek-v4-pro (A, main)	86%	59%	51%	46%	28%	26%	19%	—	15%
Qwen3.8-max (A)	94%	72%	63%	50%	37%	39%	23%	0.88 / κ0.76	2%
GLM-5.2 (A)	90%	67%	61%	57%	35%	46%	25%	0.89 / κ0.79	3%
DeepSeek-v4-pro (B)	92%	62%	59%	41%	24%	33%	15%	0.91 / κ0.82	2%

Table 11: Detector AUROC for *correct vs. genuine stubborn* (H), median over 4–6 models per dataset, strict subset. The first column holds the defining signals ($*$), near 0.5 by construction; sc is the sampling-consistency signal of the SelfCheckGPT family (Manakul et al., 2023); prompt robustness behaves the same (0.45) and is omitted for space. EigenScore enters no definition and is no better. The one internal exception, a probe read at the answer position, is in Appendix F. On G every signal is at chance (0.33–0.63) and on ordinary errors every signal works (0.53–0.93); both columns are in Appendix S. Retrieval rows are computed only on questions with a passage (gold page for PopQA, shipped DPR top-3 for EQ).

	sc/conf/probe*	EigenScore	Interrogate	CoVe	$P(\text{True})$	Verbalised	Agreement	Retrieval y/n
PopQA	0.38–0.45	0.40	0.56	0.65	0.68	0.64	0.70	0.89–0.96
EntityQuestions	0.37–0.48	0.43	0.55	0.52	0.66	0.61	0.65	0.90–0.92
TriviaQA	0.41–0.47	0.37	0.43	0.48	**0.76**	0.68	**0.76**	–
Natural Questions	0.43–0.45	0.45	0.49	0.47	0.65	0.59	0.65	–
AmbigQA	0.41–0.44	0.44	0.51	0.49	0.65	0.58	0.64	–
HaluEval QA	0.41–0.46	0.45	0.48	0.50	0.69	0.62	0.68	–
SimpleQA	0.37–0.42	0.47	0.49	0.48	0.53	0.51	0.54	–

Verifier at 14B and 32B (gold page). Qwen2.5-14B (n_H=47): FActScore-style 0.955, yes/no 0.892, extract 0.895, BEV forward 0.947; Qwen2.5-32B (n_H=40): 0.899, 0.942, 0.876, 0.945 (Table 9). The yes/no prompt's collapse at 14B is visible in its ordinary-error recall at 10% FPR, 0.31 against 0.99 for the FActScore-style prompt on the same passages.

Exploratory: $P(\text{True})$ on the weak corpus. Adding $\text{Pc}(P(\text{True}))$ to the forward score of Table 5 on the search corpus helps in 27 of the 29 configurations, mean $+0.031$: PopQA 0.764→0.808 (Qwen), 0.699→0.748 (Llama), 0.700→0.715 (Mistral), 0.736→0.820 (Qwen-14B), 0.684→0.732 (Gemma); EQ 0.762→0.783, 0.792→0.818, 0.797→0.811, 0.788→0.774; TriviaQA 0.770→0.840, 0.818→0.840, 0.717→0.736, 0.675→0.779; the two losses are Gemma on EQ (-0.015) and Qwen on SimpleQA (-0.005). Added to the bidirectional score it helps in 6 of 8 (mean $+0.013$): 0.823→0.853, 0.827→0.845, 0.769→0.770, 0.856→0.898, 0.708→0.695, 0.853→0.852, 0.861→0.877, 0.895→0.902. On the good corpora the same fusion *lowers* H AUROC on 9 of 10 pairs (mean -0.015).

Scale and post-training. Qwen2.5 1.5B/3B/7B/14B/32B on PopQA: stubborn share of errors at q=10 2.7/3.3/5.5/9.6/9.5%; share of the fused detector's false negatives at 80% recall 17/22/56/63/63%. Qwen2.5-7B base (5-shot) vs. instruct: stubborn share 17.8% vs. 9.2%; 49% of the instruct model's stubborn questions receive the same wrong answer from the base model; overlap 3.3× (p=10^{-59}); the median confidence of wrong answers rises from -1.05 to -0.60 after instruction tuning while the ordering is preserved. SimpleQA at 14B/32B: stubborn 551/529 questions, H share 57%/82%, $P(\text{True})$ 0.52/0.53.

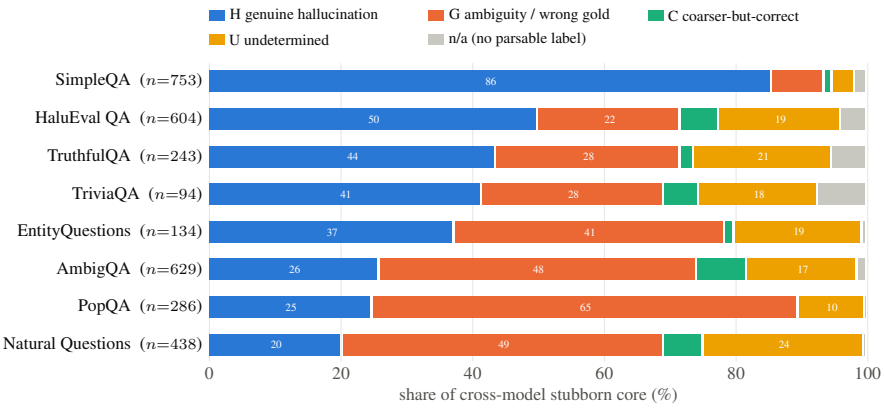


Figure 6: Composition of the cross-model stubborn core per benchmark: the same counts as Table 1, in the same order, with the unlabelled residual shown as its own segment.

Table 12: Per model and dataset: questions in the strict subset, correct answers, accuracy, 25-stubborn errors, and stubborn errors as a share of all errors. TruthfulQA (787 questions, accuracy 13–20%, stubborn share 23–40%) and Gemma-3-4B (4.3% on PopQA, 7.2% on EQ) were added after this table was built and are reported in the text instead.

model_dataset	questions	strict subset	correct	accuracy	stubborn (q=25) / share of errors
qwen_pf	14267	8767	639	0.073	73 / 0.009
llama_pf	14267	8767	1605	0.183	314 / 0.044
mistral_pf	14267	8767	1667	0.190	350 / 0.049
qwen14_pf	14267	8767	1184	0.135	214 / 0.028
qwen32_pf	14267	8767	1222	0.139	206 / 0.027
qwen_ef	22075	6380	740	0.116	96 / 0.017
llama_ef	22075	6380	1380	0.216	214 / 0.043
mistral_ef	22075	6380	1253	0.196	263 / 0.051
qwen_triviaqa	7644	7644	4358	0.570	136 / 0.041
llama_triviaqa	7644	7644	5449	0.713	101 / 0.046
mistral_triviaqa	7644	7644	5331	0.697	112 / 0.048
qwen14_triviaqa	7644	7644	5330	0.697	100 / 0.043
qwen_nq	5907	5907	1530	0.259	438 / 0.100
llama_nq	5907	5907	1895	0.321	572 / 0.143
mistral_nq	5907	5907	1913	0.324	502 / 0.126
qwen14_nq	5907	5907	2040	0.345	488 / 0.126
qwen_ambigqa	11816	11816	3754	0.318	773 / 0.096
llama_ambigqa	11816	11816	4576	0.387	825 / 0.114
mistral_ambigqa	11816	11816	4623	0.391	908 / 0.126
qwen14_ambigqa	11816	11816	4931	0.417	684 / 0.099
qwen_simpleqa	4198	4198	157	0.037	513 / 0.127
llama_simpleqa	4198	4198	229	0.055	579 / 0.146
mistral_simpleqa	4198	4198	147	0.035	760 / 0.188
qwen14_simpleqa	4198	4198	212	0.051	551 / 0.138
qwen32_simpleqa	4198	4198	224	0.053	529 / 0.133
qwen_halueval	9690	9690	2649	0.273	631 / 0.090
llama_halueval	9690	9690	2798	0.289	696 / 0.101
mistral_halueval	9690	9690	3037	0.313	1057 / 0.159
qwen14_halueval	9690	9690	3285	0.339	638 / 0.100

Table 13: Per-model detector AUROC: correct vs. genuine stubborn (H) / correct vs. ambiguous (G) / correct vs. ordinary error. Strict subset; q=25. Every column is scored with the same percentile calibration used in the main tables. The three retrieval columns are three different corpora and each is restricted to the questions that have a passage in it: *DPR* is the passage set shipped with EntityQuestions, *gold pg.* is the Wikipedia page of the intended entity, which only PopQA provides, and *search* is the top-1 Wikipedia search result, which exists for every benchmark. The first two are therefore blank wherever that corpus does not exist, not where a measurement is missing; the search column covers all rows (Qwen2.5-32B was not run on it). These three columns are the same numbers as the yes/no columns of Tables 9 and 5; the (H/G) counts in the row label are for the whole strict subset, whereas those tables report the counts inside the corpus in question, so the two sets of n_H differ. Cross-model agreement omitted for space (Table 11). "probe*" is read at the prompt's last token and enters the definition; a probe read at the *answer* position behaves differently (Appendix F). "abstain" is an indicator that the model replies *unknown* when the prompt offers that option, signed so that higher means more likely correct. Gemma-3-4B is run with the three-signal variant, so its probe and sampling-based columns are blank.

model (H/G)	sc*	conf*	probe*	EigenScore	InterrogateLLM	CoVe	$P(\text{True})$	abstain	verbalized	retr. (DPR)	retr. (gold pg.)	retr. (search)
qwen_pf (24/43)	0.323/0.247/0.929	0.367/0.272/0.938	0.372/0.338/0.827	0.330/0.277/0.925	0.674/0.365/0.738	0.650/0.365/0.781	0.697/0.357/0.762	0.517/0.436/0.820	0.648/0.403/0.757	–	0.909/0.943/0.958	0.641/0.591/0.747
llama_pf (102/172)	0.471/0.329/0.918	0.444/0.381/0.919	0.383/0.367/0.805	0.487/0.352/0.900	0.592/0.461/0.666	0.597/0.387/0.764	0.647/0.441/0.826	0.482/0.471/0.783	0.630/0.453/0.727	–	0.961/0.946/0.983	0.688/0.655/0.758
mistral_pf (112/192)	0.449/0.376/0.890	0.462/0.393/0.888	0.326/0.409/0.649	0.451/0.371/0.894	0.561/0.461/0.667	0.589/0.398/0.716	0.605/0.441/0.813	0.553/0.493/0.797	0.534/0.462/0.693	–	0.927/0.967/0.970	0.682/0.699/0.744
qwen14_pf (69/125)	0.418/0.355/0.922	0.469/0.361/0.934	0.375/0.381/0.831	0.400/0.350/0.923	0.544/0.440/0.700	0.681/0.500/0.833	0.741/0.460/0.817	0.539/0.512/0.803	0.697/0.465/0.792	–	0.892/0.889/0.873	0.672/0.619/0.667
qwen32_pf (59/128)	0.386/0.314/0.935	0.447/0.315/0.934	0.399/0.398/0.865	0.369/0.316/0.935	0.520/0.421/0.687	0.690/0.451/0.844	0.754/0.461/0.792	0.557/0.474/0.791	0.665/0.485/0.678	–	0.942/0.964/0.988	–
gemma_pf (160/121)	0.328/0.328/0.906	0.454/0.390/0.916	–	–	–	–	0.661/0.564/0.746	0.521/0.468/0.695	0.626/0.541/0.756	–	0.935/0.952/0.979	0.645/0.638/0.772
qwen_ef (43/36)	0.348/0.318/0.902	0.373/0.384/0.914	0.318/0.283/0.735	0.341/0.355/0.890	0.548/0.500/0.662	0.509/0.396/0.700	0.619/0.476/0.767	0.559/0.465/0.746	0.600/0.444/0.735	0.908/0.747/0.934	–	0.743/0.637/0.814
llama_ef (98/69)	0.491/0.377/0.919	0.489/0.421/0.899	0.378/0.383/0.753	0.510/0.397/0.879	0.546/0.462/0.640	0.625/0.534/0.714	0.695/0.546/0.823	0.468/0.468/0.766	0.615/0.493/0.742	0.906/0.760/0.949	–	0.780/0.692/0.847
mistral_ef (119/86)	0.435/0.379/0.893	0.499/0.414/0.891	0.372/0.395/0.658	0.427/0.387/0.874	0.567/0.519/0.683	0.519/0.468/0.691	0.673/0.574/0.835	0.535/0.480/0.745	0.557/0.541/0.727	0.903/0.768/0.916	–	0.802/0.748/0.843
gemma_ef (267/90)	0.412/0.364/0.894	0.465/0.331/0.884	–	–	–	–	0.637/0.526/0.629	0.505/0.506/0.572	0.656/0.535/0.632	0.918/0.794/0.939	–	0.760/0.649/0.833
qwen_ambigqa (274/307)	0.462/0.432/0.904	0.444/0.413/0.904	0.453/0.384/0.723	0.441/0.443/0.893	0.582/0.482/0.584	0.498/0.465/0.564	0.727/0.608/0.777	0.519/0.520/0.670	0.647/0.597/0.759	–	–	0.668/0.589/0.715
llama_ambigqa (281/361)	0.477/0.437/0.906	0.407/0.394/0.874	0.447/0.407/0.699	0.491/0.446/0.864	0.503/0.496/0.609	0.504/0.463/0.559	0.662/0.576/0.768	0.495/0.509/0.608	0.589/0.519/0.737	–	–	0.670/0.609/0.715
mistral_ambigqa (358/335)	0.420/0.414/0.834	0.409/0.437/0.839	0.390/0.405/0.647	0.442/0.426/0.830	0.500/0.486/0.580	0.449/0.452/0.560	0.645/0.561/0.776	0.499/0.540/0.620	0.532/0.520/0.672	–	–	0.635/0.580/0.684
qwen14_ambigqa (86/187)	0.385/0.394/0.894	0.411/0.354/0.900	0.401/0.408/0.756	0.345/0.357/0.884	0.517/0.495/0.601	0.483/0.490/0.503	0.634/0.554/0.808	0.478/0.494/0.707	0.569/0.531/0.782	–	–	0.632/0.533/0.708
qwen_simpleqa (437/47)	0.417/0.423/0.921	0.394/0.399/0.891	0.371/0.342/0.618	0.445/0.432/0.882	0.511/0.498/0.531	0.477/0.398/0.553	0.542/0.527/0.643	0.483/0.474/0.649	0.518/0.493/0.613	–	–	0.517/0.486/0.523
llama_simpleqa (496/42)	0.491/0.415/0.842	0.441/0.384/0.789	0.367/0.363/0.672	0.530/0.441/0.721	0.471/0.431/0.592	0.475/0.437/0.553	0.480/0.496/0.587	0.475/0.476/0.690	0.493/0.491/0.592	–	–	0.553/0.533/0.602
mistral_simpleqa (663/49)	0.428/0.406/0.846	0.378/0.341/0.790	0.367/0.404/0.630	0.494/0.470/0.784	0.512/0.483/0.569	0.470/0.405/0.529	0.548/0.442/0.635	0.502/0.497/0.647	0.552/0.541/0.603	–	–	0.574/0.425/0.622
qwen14_simpleqa (316/37)	0.416/0.364/0.904	0.378/0.345/0.856	0.412/0.406/0.677	0.399/0.368/0.861	0.424/0.431/0.510	0.477/0.466/0.512	0.524/0.432/0.605	0.488/0.421/0.612	0.500/0.382/0.601	–	–	0.501/0.499/0.554
qwen_triviaqa (66/39)	0.407/0.450/0.908	0.467/0.398/0.910	0.443/0.480/0.755	0.367/0.400/0.905	0.592/0.460/0.628	0.521/0.431/0.567	0.787/0.639/0.878	0.517/0.531/0.600	0.731/0.580/0.849	–	–	0.755/0.641/0.784
llama_triviaqa (43/30)	0.562/0.395/0.911	0.429/0.343/0.885	0.439/0.465/0.734	0.621/0.388/0.897	0.393/0.420/0.592	0.509/0.391/0.544	0.761/0.635/0.844	0.494/0.494/0.557	0.658/0.642/0.805	–	–	0.807/0.607/0.808
mistral_triviaqa (76/17)	0.384/0.384/0.840	0.475/0.417/0.852	0.393/0.414/0.718	0.373/0.335/0.841	0.473/0.422/0.587	0.411/0.478/0.564	0.711/0.426/0.839	0.509/0.490/0.543	0.558/0.451/0.703	–	–	0.696/0.668/0.785
qwen14_triviaqa (12/14)	0.407/0.407/0.886	0.519/0.444/0.893	0.362/0.384/0.789	0.310/0.287/0.885	0.361/0.276/0.656	0.445/0.389/0.554	0.753/0.618/0.889	0.621/0.490/0.682	0.705/0.592/0.860	–	–	0.657/0.621/0.767
qwen_nq (127/197)	0.479/0.402/0.892	0.482/0.416/0.901	0.465/0.418/0.725	0.483/0.413/0.878	0.485/0.524/0.574	0.462/0.483/0.551	0.688/0.581/0.770	0.498/0.520/0.658	0.637/0.553/0.750	–	–	0.665/0.561/0.693
llama_nq (170/239)	0.468/0.399/0.876	0.445/0.371/0.867	0.453/0.426/0.697	0.487/0.393/0.830	0.502/0.491/0.595	0.484/0.473/0.557	0.660/0.560/0.763	0.498/0.510/0.593	0.609/0.527/0.721	–	–	0.662/0.608/0.717
mistral_nq (140/207)	0.418/0.402/0.823	0.393/0.411/0.834	0.420/0.394/0.657	0.418/0.405/0.824	0.436/0.493/0.566	0.384/0.447/0.522	0.635/0.578/0.774	0.514/0.538/0.616	0.533/0.528/0.675	–	–	0.624/0.610/0.701
qwen14_nq (49/145)	0.437/0.392/0.873	0.421/0.370/0.888	0.425/0.391/0.733	0.372/0.366/0.869	0.507/0.507/0.573	0.538/0.481/0.494	0.621/0.530/0.790	0.501/0.458/0.684	0.569/0.528/0.756	–	–	0.603/0.532/0.683
qwen_halueval (366/111)	0.456/0.397/0.928	0.466/0.425/0.913	0.408/0.400/0.637	0.459/0.407/0.918	0.479/0.513/0.533	0.528/0.477/0.580	0.706/0.616/0.813	0.559/0.565/0.688	0.664/0.590/0.778	–	–	0.641/0.587/0.673
llama_halueval (412/127)	0.463/0.457/0.920	0.434/0.409/0.893	0.403/0.421/0.600	0.503/0.480/0.894	0.437/0.395/0.528	0.467/0.428/0.521	0.645/0.536/0.788	0.522/0.520/0.688	0.582/0.532/0.726	–	–	0.591/0.564/0.738
mistral_halueval (652/177)	0.462/0.470/0.846	0.437/0.420/0.841	0.408/0.402/0.645	0.449/0.443/0.811	0.478/0.447/0.580	0.474/0.446/0.524	0.671/0.532/0.812	0.527/0.511/0.614	0.558/0.512/0.698	–	–	0.660/0.592/0.686
qwen14_halueval (327/147)	0.442/0.404/0.904	0.504/0.419/0.902	0.418/0.400/0.665	0.437/0.356/0.900	0.491/0.502/0.557	0.521/0.451/0.544	0.705/0.635/0.815	0.578/0.602/0.712	0.652/0.597/0.776	–	–	0.633/0.554/0.685