READ: COMBINING READER EVIDENCE FOR NON-INTRUSIVE BLACK-BOX HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination detection supports reliable use of large language models (LLMs), but many existing methods require internal access or repeated interaction with the target model. We study *non-intrusive black-box hallucination detection* (NI-BHD), which assesses previously generated responses without either requirement. An accessible model, termed a *reader*, can provide detection signals through teacher forcing on the question and observed answer tokens. Our analysis reveals a *reader suitability mismatch*: model size, answering ability, and reader–target family match do not reliably guide reader selection. Reader rankings also vary across targets and benchmarks; readers provide complementary evidence. Accordingly, we derive an information-theoretic characterization of a reader's contribution beyond the selected evidence via optimal population log-loss reduction. We therefore propose *reader evidence aggregation for detection* (READ), a supervised framework selecting and combining multiple readers. READ selects readers by cross-validated log-loss reduction and integrates their hidden-state probe scores and token-likelihood features using a fixed additive logistic model. Experiments on 6 target LLMs and 7 benchmarks achieve 0.911 cell-macro AUROC, improving over the strongest published baseline under the same access constraints by 0.116. Despite strictly less access, READ also exceeds the per-environment best of 7 published white-box detectors in 19 of 28 matched environments.

1 INTRODUCTION

Large language models (LLMs) can generate responses that sound convincing but contain factual errors or unsupported claims. These hallucinations undermine the reliability of LLM-based applications, with potentially serious consequences in education, healthcare, and finance (Harvey et al., 2025; Roustan et al., 2025; Kang & Liu, 2023). Existing hallucination detection methods can be broadly divided into white-box and black-box approaches. White-box methods use internal signals, such as hidden states and token probabilities, to detect hallucinations through learned classifiers or uncertainty estimates (Azaria & Mitchell, 2023; Kossen et al., 2024; Orgad et al., 2025). Black-box methods assess observable model responses, using mechanisms such as consistency checks across multiple answers and follow-up questioning (Manakul et al., 2023; Yehuda et al., 2024). Both categories represent important research directions in hallucination detection(Huang et al., 2025).

However, these approaches face different access constraints in deployment. White-box methods can substantially improve detection accuracy (Orgad et al., 2025), but require internal information that commercial APIs typically do not expose (Shorinwa et al., 2024). Black-box methods avoid internal access, yet many still require additional generations or follow-up queries to the target model (Huang et al., 2025). Such interaction introduces additional inference costs and may be impossible when evaluating previously collected responses whose generating model is no longer accessible. Such constraints are common in practical deployment settings (Sriramanan et al., 2024). In these settings, the detector must assess a fixed question–answer pair without further assistance from the target. This raises a central question: *how can we reliably detect hallucinations in a response that has already been generated without inspecting the target model or requesting additional generations?*

To address these access limitations, we study a more challenging setting, *non-intrusive black-box hallucination detection* (NI-BHD), where a detector learns from labeled question–answer pairs to assess previously generated responses without accessing target internals or requesting additional

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.

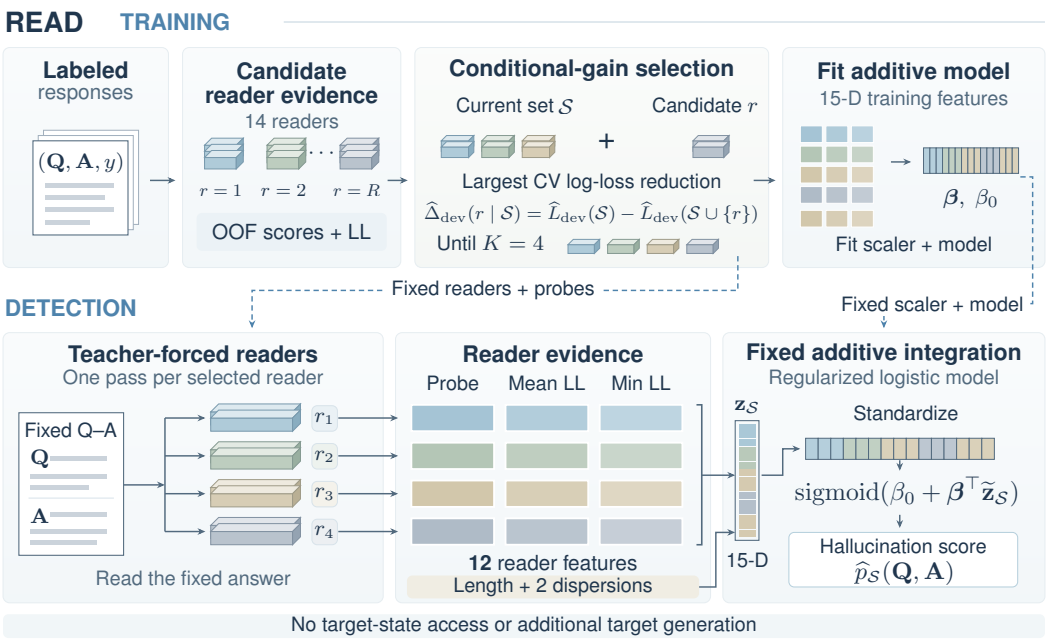


Figure 1: *Overview of* READ . Training greedily selects four readers by conditional cross-validated log-loss reduction using out-of-fold (OOF) probe scores and likelihood summaries. The population decomposition (individual signal minus redundancy) motivates this criterion. At detection, selected readers teacher-force the fixed question–answer pair once each, yielding probe scores and mean/minimum answer-token log likelihoods (LL). These 12 features, answer length, and two across-reader dispersions form a 15-dimensional vector. Training-fitted probes, standardization, and additive logistic coefficients remain fixed, without target internals or additional target generations.

generations (Definition 1). NI-BHD presents two key challenges. First, target hidden states and token probabilities, useful white-box hallucination signals, are unavailable. Second, hallucinated responses can be fluent, relevant, and internally consistent, making surface plausibility an unreliable indicator of factual correctness. Some black-box methods sample additional responses or ask follow-up questions to detect inconsistencies across responses. However, NI-BHD prohibits such interaction, requiring detection from the fixed question–answer pair alone. These challenges require informative representations of the fixed response that reveal errors beyond surface plausibility.

A natural approach is teacher forcing, where an accessible model, termed a *reader*, processes the fixed answer using the question and preceding answer tokens as context. The reader's hidden representations and token likelihoods then serve as detection signals (Sriramanan et al., 2024). When choosing a reader, larger, more knowledgeable models and those from the target's family would generally be expected to provide more reliable detection signals (Zhang et al., 2023). However, we find a counterintuitive pattern: readers that appear better suited based on these attributes are not necessarily better detectors (Section 3). Neither model size, answer-recovery accuracy, nor reader–target family match reliably guides selection. We call this a *reader suitability mismatch*. Reader rankings also vary across both targets and benchmarks. Combining readers achieves higher average AUROC than even an oracle that selects the best individual reader for each target–benchmark pair.

These observations motivate a theoretical analysis of reader value within a set. Classical information-theoretic feature selection suggests evaluating a candidate by the information it adds beyond selected features (Brown et al., 2012). Likewise, a strong reader may add little if its evidence is redundant, whereas a weaker reader may supply information that the set lacks. Accordingly, we define a reader's conditional value as the reduction in optimal population log loss upon addition. We prove that this value equals the conditional mutual information between the candidate's evidence and the hallucination label, given the available evidence (Section 4.1). We further derive a decomposition into individual signal minus signed redundancy, accounting for both overlap and synergy. This characterization provides a principled basis for selecting readers by their contribution to the set.

Based on these observations and analysis, we propose *reader evidence aggregation for detection*

(READ), a supervised framework for constructing and integrating reader evidence for NI-BHD. Each reader contributes a compact set of detection features derived from a hidden-state probe and answer-token likelihoods. During training, READ starts from an empty set and adds the candidate yielding the largest cross-validated log-loss reduction until a predefined reader budget is reached. The selected evidence is combined with a regularized additive logistic model. The reader set and learned model coefficients remain fixed at inference. Together, these components exploit complementary detection signals while requiring neither internal access to the target nor additional target generations.

We evaluate READ on 6 target LLMs and 7 benchmarks under NI-BHD. Across target–benchmark environments, READ achieves a macro-average AUROC of 0.911, improving over the strongest evaluated baseline under the same access constraints by 0.116. Comparisons on matched open-weight environments further show competitive performance with published white-box detectors that use the target model's internal information, while READ relies only on accessible readers to assess the fixed response. Ablations examine the contributions of conditional-gain reader selection and fixed additive integration (Section 5). *Our main contributions are summarized below:*

• We formulate *non-intrusive black-box hallucination detection* (NI-BHD), a detection setting in which detectors learn from a set of labeled responses and assess existing answers without accessing the target model's internal information or requesting additional generations.

• We systematically analyze teacher-forced detection and identify a *reader suitability mismatch*. Size, answering ability, and reader–target family match are unreliable selection guides. Performance is environment-dependent, while complementary errors motivate pooling readers' evidence.

• We propose READ, which selects readers on the basis of their conditional detection gain and integrates their evidence through a fixed additive model. Experiments conducted across target models and benchmarks demonstrate its effectiveness over representative baselines.

2 LEARNING SETUPS

LLMs and Token Sequences. We model a target LLM as a conditional distribution $\mathbb{P}_{\boldsymbol{\theta}_m}(\cdot)$ over token sequences, with parameters $\boldsymbol{\theta}_m$. Given a question $\mathbf{Q} = [x_1, \ldots, x_k]$, optionally including task-specific context, it generates answer $\mathbf{A} = [x_{k+1}, \ldots, x_{k+l}]$ autoregressively: $\mathbb{P}_{\boldsymbol{\theta}_m}(\mathbf{A} \mid \mathbf{Q}) = \prod_{j=k+1}^{k+l} \mathbb{P}_{\boldsymbol{\theta}_m}(x_j \mid x_1, \ldots, x_{j-1})$. Let $\mathcal{Q}$ and $\mathcal{A}$ denote question and answer spaces, respectively.

Internal Representations and Readers. For target LLM $\mathbb{P}_{\boldsymbol{\theta}_m}$, let L_m denote its number of layers and d_m its hidden dimension. Processing a question–answer pair $(\mathbf{Q}, \mathbf{A})$ with $k + l$ tokens using the target model produces the layer-ℓ representations $\mathbf{H}^{\ell}_{\boldsymbol{\theta}_m}(\mathbf{Q}, \mathbf{A}) \in \mathbb{R}^{(k+l)\times d_m}$. The j-th row, $\mathbf{h}^{\ell}_{\boldsymbol{\theta}_m, j}(\mathbf{Q}, \mathbf{A}) \in \mathbb{R}^{d_m}$, is the hidden state at token position j in the concatenated question and answer.

We also consider R auxiliary LLMs $\mathbb{P}_{\boldsymbol{\psi}_1}, \ldots, \mathbb{P}_{\boldsymbol{\psi}_R}$, termed *readers*, processing the same question–answer text. Reader r has L_r layers and hidden dimension d_r. Its tokenizer encodes input context and answer as $\mathbf{Q}_r$ and $\mathbf{A}_r = [a_{r,1}, \ldots, a_{r,l_r}]$, with lengths k_r and l_r, respectively. Processing it yields $\mathbf{H}^{\ell}_{\boldsymbol{\psi}_r}(\mathbf{Q}, \mathbf{A}) \in \mathbb{R}^{(k_r+l_r)\times d_r}$, with j-th row $\mathbf{h}^{\ell}_{\boldsymbol{\psi}_r, j}(\mathbf{Q}, \mathbf{A}) \in \mathbb{R}^{d_r}$. The target and readers may differ in architecture, depth, hidden dimension, and tokenization. Their representation matrices may therefore differ in shape, with model-specific layer indices, token positions, and hidden coordinates.

Labeled Data. For each question $\mathbf{Q}$, let $\mathbf{E}_{\mathbf{Q}}$ denote the reference answer or task-specific evidence used to assess correctness. An answer $\mathbf{A}_m \sim \mathbb{P}_{\boldsymbol{\theta}_m}(\cdot \mid \mathbf{Q})$ is labeled as hallucinated ($y_m = 1$) if it contains a material factual or task-specific error according to $\mathbf{E}_{\mathbf{Q}}$, and non-hallucinated ($y_m = 0$) otherwise. The resulting labeled dataset for target m is

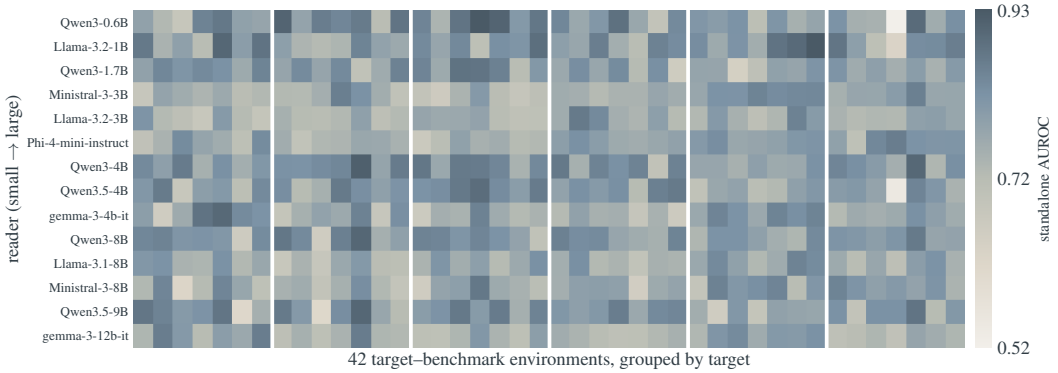
$$\mathcal{D}_m = \left\{(\mathbf{Q}^i, \mathbf{A}^i_m, y^i_m)\right\}_{i=1}^{N_m}, \tag{1}$$

where N_m is the number of labeled responses.

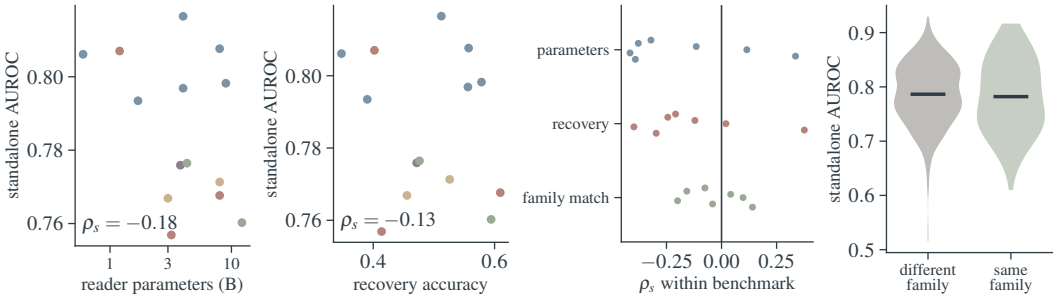
Definition 1 (Non-Intrusive Black-Box Hallucination Detection (NI-BHD))**.** *Given labeled question–answer pairs, NI-BHD aims to learn a detector*

$$D_{\boldsymbol{\phi}} : \mathcal{Q} \times \mathcal{A} \to [0, 1], \qquad s = D_{\boldsymbol{\phi}}(\mathbf{Q}, \mathbf{A}), \tag{2}$$

where $\boldsymbol{\phi}$ denotes the detector parameters and a larger score s indicates greater evidence of hallucination. At evaluation time, the detector assesses a held-out, previously generated pair $(\mathbf{Q}, \mathbf{A})$ from



(a) Standalone detection performance.



(b) Parameter count. (c) Answer recovery. (d) Correlations. (e) Family match.

Figure 2: *Model attributes serve as unreliable reader-selection guides.* (a) Standalone detection AUROC varies across readers and target–benchmark pairs. The rows are ordered by parameter count; white lines separate different target models. (b)–(c) Larger readers and those with higher answer-recovery accuracy do not consistently achieve higher cell-macro AUROC (with equal weight for each environment). (d) Within-benchmark correlations with size, recovery accuracy, and reader–target family match are not consistently positive. (e) Same-family and different-family readers' AUROC distributions substantially overlap. ρ_s denotes Spearman's rank correlation coefficient.

a target LLM, with the pair's label unknown to it. It may use auxiliary readers and access their parameters, hidden states, and token-level predictive distributions. The corresponding information of the target model is unavailable, and no additional queries to the target are permitted.

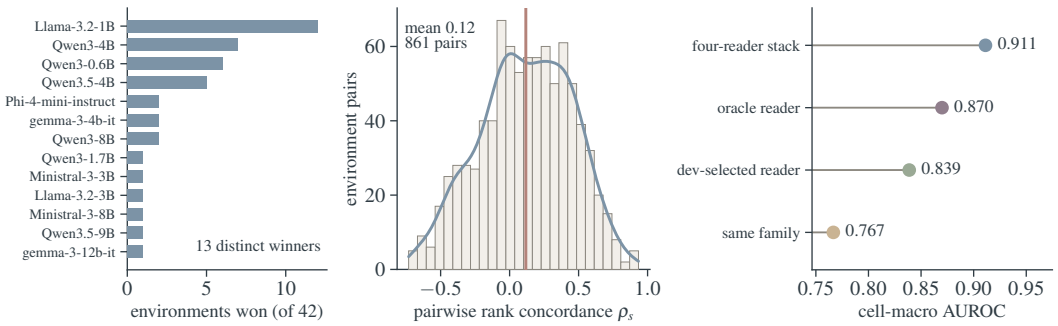
3 Analysis of Reader Evidence: What Makes a Reader Useful?

To detect hallucinations without target internals or further target queries, we need discriminative evidence from the fixed question–answer pair. Teacher forcing provides a natural way to obtain representations and token likelihoods from accessible readers. However, these signals depend on the reader that produces them, and their ability to distinguish correct from hallucinated answers remains to be assessed. Prior work on uncertainty-based detection using teacher forcing finds that larger models within a family and models with stronger general capabilities tend to perform better (Zhang et al., 2023). This motivates us to examine *whether these model attributes also indicate more discriminative representations and likelihood features across targets and benchmarks.*

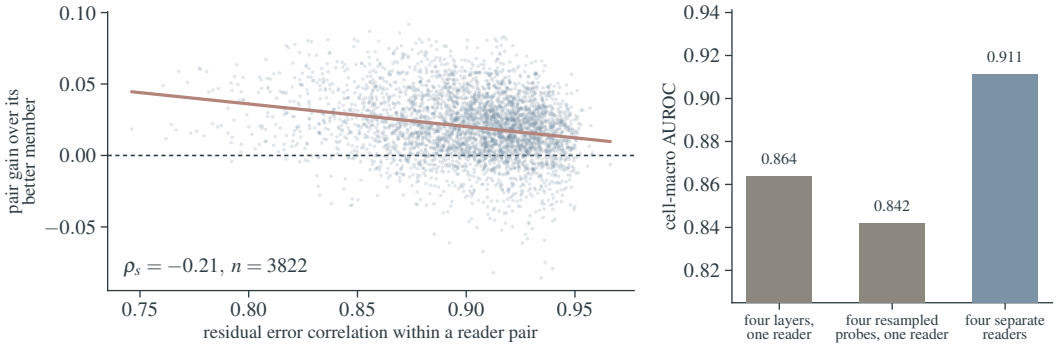
Teacher Forcing. For a fixed pair $(\mathbf{Q}, \mathbf{A})$, reader r uses its own tokenization $(\mathbf{Q}_r, \mathbf{A}_r)$. Under teacher forcing, the input context for predicting $a_{r,t}$ consists of $\mathbf{Q}_r$ and the observed answer prefix $\mathbf{A}_{r,<t} = [a_{r,1}, \ldots, a_{r,t-1}]$. The layer-$\ell$ hidden state at this prediction position is $\mathbf{h}^{\ell}_{\psi_r, k_r+t-1}(\mathbf{Q}, \mathbf{A})$. We also record the conditional probability of the observed token,

$$p_{r,t} = \mathbb{P}_{\psi_r}(a_{r,t} \mid \mathbf{Q}_r, \mathbf{A}_{r,<t}), \qquad t = 1, \ldots, l_r, \tag{3}$$

with corresponding token log likelihood $\log p_{r,t}$. Both the hidden states and the token likelihoods are obtained directly from the fixed question–answer pair without generating a new answer.



(a) Who wins where. (b) Ordering agreement. (c) One-reader rules.



(d) Errors that do not coincide. (e) Distinct models, not repeated views.

Figure 3: *Reader variability and complementary errors motivate combining readers.* (a)–(b) Win counts and rank agreement across 861 environment pairs show varying reader rankings. (c) Four-reader stacking outperforms corresponding single-reader references (cell-macro AUROC, 42 environments), including an oracle selecting each environment's best reader using evaluation labels. Same-family results cover 28 open-weight environments. (d) Lower residual correlations accompany larger pairwise gains over the better reader. (e) Four distinct readers outperform one reader's four layers or four resampled probes on cell-macro AUROC. Please see Appendix D.5.

Observation Design. We evaluate all 14 readers on the same retained responses from 6 targets and 7 benchmarks. We train each reader's logistic detector using a linear-probe score from hidden states, mean and minimum answer-token log likelihoods, and answer length, and evaluate AUROC within each target–benchmark pair. All readers use the same probe class, fitting procedure, and data splits grouped by source question. Let $\mathcal{E}$ be the set of environments, each a target–benchmark pair. Let $A_e(\mathcal{S})$ denote the held-out AUROC using reader set $\mathcal{S}$. Individual readers' results form the matrix

$$\mathbf{M} = [M_{r,e}]_{r,e}, \qquad M_{r,e} = A_e(\{r\}), \qquad \bar{M}_r = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} M_{r,e}. \tag{4}$$

Each column compares readers on the same responses, and each row shows one reader's performance across environments. Unless otherwise stated, AUROCs are averaged equally across environments (cell-macro AUROC). Further details are in Appendices D.4 and D.5.

3.1 Model Attributes Do Not Reliably Guide Reader Selection

We first ask whether reader size, answering ability, and family match predict detection performance in our experiments. Figure 2 relates these three attributes to the performance matrix $\mathbf{M}$.

Scale and Answer Capability. To test whether size and answering ability predict detection performance, we correlate $\bar{M}_r$ with parameter count and answer-recovery accuracy. We measure recovery by asking each reader to identify the reference answer among three distractors from the same benchmark, then average accuracy across environments. The Spearman correlations are -0.177 and -0.125, respectively (Figures 2b and 2c). For example, Llama-3.2-1B reaches 0.807 AUROC, compared with 0.760 for gemma-3-12b. These correlations describe the evaluated reader pool.

Table 1: READ *achieves the highest average AUROC on all six targets.* AUROC↑: per target–benchmark pair. **Avg.**: mean AUROC across benchmarks. **Bold**/underlining mark column best/second-best across method groups. "-" indicates unavailable API target internals. SE: semantic entropy. LLM-Check uses reader-side features. Gemini/Qwen3.5 and Prometheus-2 rows use prompted critics; these critics also assess fixed question–answer pairs without accessing the target.

Method	Llama-3.3-70B								Qwen3-32B								Mistral-Small-24B							
	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	**Avg.**	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	**Avg.**	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	**Avg.**
White box																								
Target likelihood	0.763	0.806	0.700	0.672	0.776	0.769	0.716	0.743	0.757	0.703	0.721	0.701	0.854	0.710	0.695	0.734	0.777	0.761	0.830	0.757	0.795	0.779	0.811	0.787
Prob.-weighted SE	0.790	0.799	0.852	0.800	0.874	0.721	0.805	0.806	0.816	0.800	0.801	0.861	0.898	0.765	0.830	0.824	0.833	0.822	0.918	0.861	0.861	0.730	0.890	0.845
EigenScore	0.844	**0.956**	0.857	0.812	0.890	0.822	0.778	0.851	0.696	0.867	0.792	0.833	0.910	0.836	0.775	0.816	0.784	0.846	**0.948**	0.866	0.865	0.722	0.712	0.820
Attn.-Laplacian	0.935	0.906	0.828	0.773	0.944	0.827	0.867	0.869	0.840	0.676	0.682	0.750	0.907	0.757	0.801	0.773	0.859	0.826	0.844	0.834	0.899	0.737	0.751	0.821
SAPLMA probe	0.934	0.927	0.862	0.775	0.887	0.893	0.921	0.885	0.892	0.813	0.826	0.853	0.909	0.855	0.880	0.861	0.895	**0.891**	0.902	0.913	0.920	0.874	0.931	0.904
Mass-mean probe	0.822	0.915	0.734	0.854	0.930	0.891	**0.949**	0.871	0.865	0.864	0.784	0.858	0.926	**0.870**	0.908	0.868	0.770	0.820	0.875	**0.933**	0.858	0.846	0.906	0.858
SE probe	0.857	0.808	0.792	0.754	0.885	0.765	0.852	0.816	0.824	0.620	0.668	0.658	0.808	0.738	0.742	0.723	0.820	0.704	0.837	0.730	0.822	0.748	0.866	0.790
Black box																								
Discrete SE	0.669	0.815	0.707	0.742	0.753	0.863	0.792	0.763	0.580	0.760	0.693	0.637	0.777	0.839	0.684	0.710	0.687	0.745	0.813	0.739	0.667	0.866	0.782	0.757
Lexical similarity	0.721	0.844	0.619	0.705	0.732	0.754	0.618	0.713	0.692	0.639	0.608	0.595	0.686	0.561	0.580	0.623	0.649	0.630	0.639	0.681	0.584	0.574	0.692	0.635
SelfCheckGPT	0.725	0.846	0.696	0.698	0.810	0.801	0.798	0.768	0.729	0.770	0.669	0.719	0.859	0.766	0.811	0.761	0.684	0.749	0.662	0.732	0.804	0.680	0.745	0.722
InterrogateLLM	0.709	0.859	0.725	0.729	0.867	0.626	0.635	0.736	0.758	0.778	0.765	0.701	0.839	0.695	0.596	0.733	0.708	0.552	0.730	0.727	0.742	0.548	0.624	0.661
Verbalized conf.	0.648	0.776	0.632	0.597	0.764	0.596	0.721	0.676	0.745	0.724	0.630	0.688	0.756	0.714	0.676	0.705	0.720	0.764	0.647	0.711	0.861	0.641	0.802	0.735
Gemini	0.664	0.756	0.681	0.770	0.754	0.700	0.559	0.698	0.550	0.645	0.552	0.712	0.666	0.563	0.615	0.615	0.621	0.728	0.590	0.710	0.531	0.594	0.676	0.636
Qwen3.5-27B	0.705	0.744	0.579	0.708	0.618	0.595	0.567	0.645	0.575	0.658	0.559	0.630	0.628	0.615	0.630	0.613	0.611	0.725	0.726	0.763	0.673	0.636	0.593	0.675
Qwen3.5-9B	0.688	0.697	0.667	0.657	0.740	0.633	0.617	0.671	0.637	0.644	0.570	0.617	0.664	0.575	0.582	0.613	0.654	0.632	0.606	0.731	0.705	0.574	0.653	0.651
Prometheus-2 (7B)	0.640	0.513	0.599	0.600	0.507	0.601	0.589	0.578	0.474	0.585	0.568	0.616	0.595	0.633	0.561	0.576	0.533	0.608	0.563	0.628	0.568	0.604	0.540	0.578
Non-intrusive black box																								
Answer length	0.497	0.553	0.460	0.506	0.494	0.477	0.431	0.488	0.535	0.532	0.480	0.514	0.458	0.516	0.439	0.496	0.610	0.463	0.402	0.553	0.437	0.452	0.514	0.490
TF–IDF logistic	0.662	0.653	0.880	0.601	0.644	0.858	0.786	0.726	0.525	0.634	0.735	0.764	0.604	0.858	0.675	0.685	0.640	0.593	0.800	0.714	0.671	0.727	0.666	0.687
DeBERTa-v3-large	0.650	0.900	0.830	0.642	0.900	0.764	0.810	0.785	0.712	0.823	0.890	0.712	0.893	0.770	0.736	0.791	0.645	0.749	0.732	0.704	0.877	0.554	0.568	0.690
ModernBERT-large	0.852	0.637	0.739	0.620	**0.967**	0.825	0.939	0.797	0.750	0.664	0.739	0.654	0.856	0.767	0.657	0.727	0.856	0.674	0.777	0.860	0.904	0.676	0.919	0.809
ELECTRA-large	0.883	0.882	**0.918**	**0.876**	0.894	0.824	0.824	0.872	0.739	0.820	0.780	0.836	0.690	0.864	0.738	0.781	0.806	0.718	0.781	0.890	0.830	0.752	0.697	0.782
LLM-Check	0.757	0.821	0.728	0.824	0.833	0.770	0.844	0.797	0.764	0.745	0.673	0.650	0.745	0.803	0.795	0.739	0.643	0.768	0.738	0.696	0.605	0.733	0.763	0.706
READ	**0.936**	0.875	0.848	0.874	0.966	**0.910**	0.925	**0.905**	**0.919**	**0.876**	**0.922**	**0.907**	**0.957**	0.853	**0.915**	**0.907**	**0.929**	0.869	0.936	0.919	**0.922**	**0.877**	**0.953**	**0.915**

Method	Gemma-3-27B								GPT-4.1								Claude-Sonnet-4							
	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	**Avg.**	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	**Avg.**	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	**Avg.**
White box																								
Target likelihood	0.732	0.784	0.799	0.744	0.777	0.830	0.790	0.779	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Prob.-weighted SE	0.629	0.757	0.820	0.796	0.745	0.796	0.810	0.765	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EigenScore	0.724	**0.890**	0.865	0.855	0.840	**0.912**	0.771	0.837	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Attn.-Laplacian	0.796	0.758	0.829	0.731	0.794	0.761	0.716	0.769	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SAPLMA probe	0.776	0.862	0.827	0.795	0.822	0.891	0.867	0.834	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mass-mean probe	0.769	0.845	0.790	0.781	0.831	0.875	0.895	0.827	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SE probe	0.732	0.674	0.773	0.763	0.824	0.809	0.801	0.768	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Black box																								
Discrete SE	0.705	0.829	0.791	0.758	0.825	0.884	0.802	0.799	0.764	0.912	0.901	0.821	0.878	**0.906**	0.829	0.859	0.741	0.806	0.825	0.781	0.838	**0.899**	0.863	0.822
Lexical similarity	0.687	0.732	0.595	0.700	0.644	0.680	0.681	0.674	0.609	0.627	0.730	0.636	0.716	0.653	0.634	0.658	0.670	0.727	0.656	0.628	0.690	0.713	0.682	0.681
SelfCheckGPT	0.655	0.745	0.772	0.723	0.712	0.766	0.745	0.731	0.633	0.773	0.672	0.723	0.790	0.779	0.821	0.742	0.673	0.714	0.631	0.702	0.768	0.776	0.801	0.723
InterrogateLLM	0.656	0.795	0.685	0.650	0.731	0.653	0.683	0.693	0.665	0.789	0.789	0.645	0.807	0.761	0.588	0.720	0.751	0.855	0.753	0.794	0.856	0.743	0.676	0.775
Verbalized conf.	0.675	0.781	0.531	0.678	0.756	0.607	0.734	0.680	0.541	0.733	0.680	0.600	0.742	0.618	0.759	0.667	0.660	0.754	0.640	0.658	0.703	0.537	0.734	0.670
Gemini	0.650	0.831	0.627	0.678	0.622	0.686	0.667	0.680	0.690	0.801	0.674	0.705	0.694	0.601	0.697	0.695	0.667	0.695	0.530	0.709	0.607	0.623	0.631	0.638
Qwen3.5-27B	0.644	0.763	0.602	0.536	0.638	0.641	0.514	0.620	0.629	0.753	0.671	0.726	0.599	0.679	0.723	0.683	0.660	0.708	0.581	0.648	0.617	0.623	0.597	0.634
Qwen3.5-9B	0.526	0.606	0.560	0.606	0.527	0.628	0.577	0.576	0.583	0.688	0.594	0.686	0.545	0.632	0.623	0.622	0.638	0.647	0.615	0.626	0.610	0.532	0.584	0.608
Prometheus-2 (7B)	0.567	0.605	0.598	0.591	0.634	0.680	0.521	0.599	0.562	0.667	0.681	0.662	0.658	0.710	0.634	0.653	0.619	0.670	0.632	0.816	0.584	0.684	0.640	0.664
Non-intrusive black box																								
Answer length	0.452	0.569	0.588	0.497	0.421	0.489	0.528	0.506	0.484	0.431	0.500	0.501	0.421	0.505	0.433	0.468	0.554	0.440	0.469	0.487	0.453	0.517	0.571	0.499
TF–IDF logistic	0.551	0.622	0.747	0.541	0.721	0.808	0.626	0.660	0.572	0.666	0.845	0.735	0.704	0.848	0.747	0.731	0.622	0.644	0.711	0.689	0.716	0.865	0.531	0.683
DeBERTa-v3-large	0.641	0.842	0.761	0.645	0.865	0.729	0.735	0.745	0.848	**0.928**	**0.953**	0.814	0.912	0.818	0.916	0.884	0.757	**0.944**	0.859	0.678	0.937	0.759	0.732	0.810
ModernBERT-large	0.713	0.677	0.822	0.622	0.874	0.752	0.848	0.758	0.850	0.849	0.825	0.687	0.925	0.783	0.844	0.823	0.868	0.716	0.716	0.661	**0.941**	0.857	0.811	0.796
ELECTRA-large	0.737	0.819	0.770	0.891	0.870	0.875	0.836	0.828	0.752	0.762	0.841	0.787	0.780	0.711	0.681	0.759	0.783	0.761	0.694	0.815	0.850	0.787	0.541	0.747
LLM-Check	0.681	0.736	0.695	0.703	0.706	0.749	0.657	0.704	0.770	0.865	0.756	0.804	0.754	0.876	0.852	0.811	0.822	0.834	0.700	0.752	0.794	0.870	0.851	0.803
READ	**0.913**	0.836	**0.951**	**0.893**	**0.908**	0.879	**0.926**	**0.901**	**0.908**	0.926	0.928	**0.905**	**0.939**	0.903	**0.959**	**0.924**	**0.955**	0.930	**0.875**	**0.914**	0.936	0.876	**0.926**	**0.916**

To check whether averaging across benchmarks hides differences between them, we repeat the analysis within each benchmark. Recovery correlations are negative in five of the 7 benchmarks. Size correlations are negative in five and positive in two (Figure 2d). These comparisons also provide no consistent support for selecting readers by size or recovery accuracy.

Reader–Target Family Match. To test whether readers from the target's family perform better, we compare their AUROCs with those of readers from other families. The distributions substantially overlap, and correlations with family match range from -0.200 to $+0.141$ across benchmarks (Figures 2e and 2d). Sharing the target's family therefore offers no consistent advantage in this pool.

Selection by Model Attributes. To test whether these attributes help select readers, we compare sets of four readers chosen by size, recovery accuracy, or random sampling. We combine each set's readers using the same additive logistic model. Their cell-macro AUROCs are 0.877, 0.877, and 0.887, respectively (Table 2). Neither size nor recovery accuracy improves on the mean for random sets.

Answer Recovery versus Detection. To test whether detection requires successful answer recovery, we restrict evaluation to responses on which all readers fail the recovery evaluation defined in Appendix E.2. Detection remains informative on this subset. Readers can therefore provide evidence for hallucination detection even when they fail to answer the question correctly in these tests.

Table 2: *Distinct readers outperform repeated views, conditional gain improves selection, and additive fusion performs comparably to the MLP.* All panels report cell-macro AUROC over 42 target–benchmark pairs, with s.d. across pairs. The same-family rule covers 28 open-weight cells; its 14 API cells are N/A because neither API target has a same-family candidate reader. Other rows cover all 42 cells. Compare within (b)–(d), where **bold**/underlining mark best/second-best AUROCs. Control details and sample and feature comparability limits are in Appendix D.5.

Variant	AUROC↑	Log loss↓	AUROC s.d.	Variant	AUROC↑	Log loss↓	AUROC s.d.
(a) Single-reader references ($K = 1$)				*(b) Repeated-view controls (four views)*			
Benchmark dev-selected reader	0.839	0.499	0.041	Four layers, one reader	0.864	0.466	0.037
Same-family reader (28/42 cells)	0.767	0.550	0.051	Four resampled probes, one reader	0.842	0.495	0.042
				Four distinct readers	**0.911**	0.384	0.031
(c) Reader-set construction ($K = 4$)				*(d) Evidence integration ($K = 4$)*			
Largest-reader set	0.877	0.445	0.044	Uniform score average	0.869	0.578	0.051
Random reader set	0.887	0.427	0.026	Learned hard router	0.746	0.587	0.083
Best-answerer set	0.877	0.445	0.044	Scores-only logistic stack	0.871	0.449	0.054
Fixed reader set	0.870	0.452	0.034	Likelihood-only logistic stack	0.756	0.586	0.029
Individual-signal selection	0.896	0.409	0.041	Linear score–covariate interactions	**0.912**	0.378	0.031
Label-free diversity	0.880	0.439	0.035	Softmax gate + direct evidence	0.890	0.423	0.047
				MLP (4 hidden units)	0.911	0.376	0.032
Conditional gain (READ)	**0.911**	0.384	0.031	**Fixed additive stack (READ)**	0.911	0.384	0.031

Implication. These results show a *reader suitability mismatch*. In the evaluated pool, size, measured answering ability, and family match do not reliably guide reader selection: reader sets chosen by size or recovery accuracy do not outperform randomly chosen sets. The trends with size and capability reported for uncertainty-based detection (Zhang et al., 2023) therefore do not necessarily extend to supervised detection using reader representations and likelihood features.

3.2 Reader Variability and Evidence Complementarity

We next ask whether selecting a reader by its detection performance is sufficient, or whether combining readers improves detection over using that reader alone. Figure 3 examines how rankings change across environments and whether combinations outperform individual readers.

Variation across Environments. To test whether the same readers perform best across environments, we rank the readers by AUROC in each column of $\mathbf{M}$. We then count each reader's wins and compare the rankings across environments. Thirteen readers rank first in at least one of the 42 environments, and the highest win rate is 0.286 (Figure 3a). The mean correlation between rankings is 0.118 (Figure 3b). Thus, no reader consistently ranks first across the evaluated environments.

Single-Reader Selection versus Combination. To test whether better selection can match combination, we compare several single-reader selection rules with a four-reader detector. One rule selects each benchmark's best reader by development performance. An oracle selects $\arg\max_r M_{r,e}$ per environment using evaluation labels. Selection on development data reaches 0.839 cell-macro AUROC. The oracle reaches 0.870, 0.041 below the combined detector's 0.911 (Figure 3c). Even with evaluation labels for selection, the best individual reader does not match the combined detector.

Complementary Detection Evidence. To test whether combining readers helps more when their errors differ, we measure each pair's AUROC gain over its better member in the same environment,

$$\Delta_e(r, r') = A_e(\{r, r'\}) - \max\{M_{r,e}, M_{r',e}\}. \tag{5}$$

We compare gains with readers' residual-error correlation. Across 3,822 comparisons, Spearman's $\rho_s = -0.21$ (Figure 3d). Pairs with less correlated errors tend to benefit more from combination.

Distinct Readers versus Repeated Views. To test whether multiple views of one reader offer a similar benefit, we compare four layers of one reader, four resampled probes of one reader, and four distinct readers. All three use additive logistic fusion. Their AUROCs are 0.864, 0.842, and 0.911, respectively (Figure 3e). The advantage of distinct readers remains when all three variants are evaluated on the same responses. Appendix D.5 describes the controls and the limits of this comparison.

Implication. The combined detector outperforms oracle selection, and reader pairs with less correlated errors tend to yield larger gains. Under a fixed budget, this motivates selecting readers by their additional evidence beyond selected readers. Section 4 formalizes this conditional detection value.

4 READ: READER EVIDENCE SELECTION AND INTEGRATION

Section 3 examines pairwise reader complementarity. Selecting a reader set requires assessing each candidate's contribution given all selected readers. READ uses this conditional contribution to select readers under a fixed budget and combines the selected evidence with an additive logistic model.

Reader Evidence. For each reader r, we combine a linear-probe score s_r with two likelihood features derived from the token probabilities in Equation 3 to form its evidence vector,

$$\mathbf{z}_r = \left(s_r, \bar{\ell}_r, \ell_r^{\min}\right)^\top, \qquad \bar{\ell}_r = \frac{1}{l_r}\sum_{t=1}^{l_r} \log p_{r,t}, \quad \ell_r^{\min} = \min_{1\le t\le l_r} \log p_{r,t}. \tag{6}$$

At a selected hidden layer, we average the states used to predict the answer tokens. A logistic probe trained on hallucination labels maps this mean representation to s_r. We write Z_r for the corresponding random vector and X_S for the concatenated evidence from reader set S. The shared covariate U is answer length, measured identically for all readers. Extraction details are in Appendix D.4.

4.1 CONDITIONAL VALUE OF A READER

A reader may perform well alone yet provide little new evidence when added to a set. We therefore assess its contribution *given the selected readers*. Fix a target–benchmark environment e with distribution $\mathbb{P}_e$ and hallucination label $Y \in \{0,1\}$. For analysis, let $\mathbb{Q}_e$ preserve the class-conditional distributions of $\mathbb{P}_e$ while assigning probability $1/2$ to each class. With binary log loss $\ell_{\log}(y,p) = -y\log p - (1-y)\log(1-p)$, define $R_e^*(S) = \inf_f \mathbb{E}_{\mathbb{Q}_e}[\ell_{\log}\{Y, f(U, X_S)\}]$, where f ranges over all measurable predictors. Since optimal log loss equals conditional entropy,

$$R_e^*(S) - R_e^*(S \cup \{r\}) = I_{\mathbb{Q}_e}(Y; Z_r \mid U, X_S) =: V_e(r \mid S) \ge 0. \tag{7}$$

The optimal reduction in log loss therefore measures the label information a reader adds beyond the current evidence (Brown et al., 2012). Defining $R_e(r \mid S) := I_{\mathbb{Q}_e}(Y; Z_r \mid U) - V_e(r \mid S)$ gives

$$V_e(r \mid S) = \underbrace{I_{\mathbb{Q}_e}(Y; Z_r \mid U)}_{\text{individual signal}} - \underbrace{R_e(r \mid S)}_{\text{signed redundancy}}. \tag{8}$$

Positive redundancy means a reader provides less label information conditional on the selected evidence. Negative redundancy indicates synergy, where conditioning makes the reader more informative. This decomposition motivates evaluating conditional contribution beyond individual performance or score correlation alone. For selection, we evaluate predictive log-loss reduction achievable by the additive logistic model. Appendix A relates this model-specific gain to conditional information through the change in approximation error, and connects cross-validated selection to this gain.

4.2 READER SELECTION

We select readers by their drop in cross-validated log loss, using the additive logistic model in Section 4.3. In each outer training split, four-fold cross-validation fits probes on three folds and produces scores for the remaining fold. For each reader, we select one of four candidate relative depths by its inner validation loss and use its out-of-fold (OOF) scores for reader selection.

Let N_{tr} denote the outer training sample size and $b(i)$ the inner fold containing example i, labeled y_i. Let $\widehat{p}_{i,S}^{(-b(i))}$ denote the prediction from a combiner trained on other folds using these OOF features. Logistic fits use class-balanced weights, while candidate selection uses unweighted log loss,

$$\begin{aligned} \widehat{L}_{\text{dev}}(S) &= \frac{1}{N_{\text{tr}}}\sum_{i=1}^{N_{\text{tr}}} \ell_{\log}\left(y_i, \widehat{p}_{i,S}^{(-b(i))}\right), \\ \widehat{\Delta}_{\text{dev}}(r \mid S) &= \widehat{L}_{\text{dev}}(S) - \widehat{L}_{\text{dev}}(S \cup \{r\}). \end{aligned} \tag{9}$$

Starting from $S = \varnothing$, we add the remaining candidate maximizing $\widehat{\Delta}_{\text{dev}}(r \mid S)$ until $|S| = K = 4$. The budget K is a hyperparameter, fixed across targets and splits; Appendix E.1 sweeps it from one reader to the full pool. The empty-set baseline predicts development-set prevalence, and every nonempty candidate includes answer length. Layer and reader selection share inner validation data. Outer evaluation labels are not used. Appendix D.4 details this reuse and baseline convention.

4.3 Evidence Integration

For a nonempty reader set, define $\mathbf{g}_{\mathcal{S}} = (\mathrm{sd}_{j\in\mathcal{S}}\, s_j, \mathrm{sd}_{j\in\mathcal{S}}\, \bar{\ell}_j)^\top$, with both entries set to zero for a single reader. Let $\mathbf{z}_{\mathcal{S}}$ concatenate the reader features $\{\mathbf{z}_j\}_{j\in\mathcal{S}}$, the observed answer length u, and the dispersion summaries $\mathbf{g}_{\mathcal{S}}$. Writing $\widetilde{\mathbf{z}}_{\mathcal{S}}$ for the standardized features, the detector is

$$\widehat{p}_{\mathcal{S}}(\mathbf{Q}, \mathbf{A}) = \mathrm{sigmoid}\big(\beta_0 + \boldsymbol{\beta}^\top \widetilde{\mathbf{z}}_{\mathcal{S}}\big)\,. \tag{10}$$

We fit this model with ℓ_2 regularization using the selected OOF evidence from the outer training data. We then refit the selected probes on all outer training responses for inference. For each fit, standardization uses only its training responses, with the resulting parameters kept fixed during evaluation.

The dispersion summaries are deterministic functions of the reader evidence. They add no information for an unrestricted predictor, but can help a logistic model use the available evidence (Appendix A). The selected readers, standardization, and coefficients remain fixed at inference. Section 5 compares this integration with alternative fusion models using the same selected evidence.

5 Experiments

Experimental Setup. We evaluate READ on 6 targets (four open-weight, two API), using 14 non-target readers. 7,560 responses cover 42 target–benchmark pairs across TruthfulQA, HaluEval-QA, TriviaQA, CoQA, GSM8K, MBPP, and xLAM (Lin et al., 2022b; Li et al., 2023; Joshi et al., 2017; Reddy et al., 2019; Cobbe et al., 2021; Austin et al., 2021; Liu et al., 2024). Baselines: white-box (target internals), black-box (extra target responses or external judges), and NI-BHD (existing question–answer pair). Access remains unchanged; some use target information unavailable to READ. We fit one detector across targets, grouping outer and inner splits by source question. Cell-macro AUROC weights pairs equally. Appendix D details datasets, labeling, baselines, and training.

Main Results. Without target internals or additional target queries, READ outperforms all evaluated NI-BHD baselines in overall AUROC and, with strictly less access, surpasses the per-environment best of the 7 evaluated white-box baselines in 19 of 28 matched environments (Table 1). It achieves 0.911 cell-macro AUROC over all 42 target–benchmark pairs and 0.907 over the 28 matched pairs involving open-weight targets. The gain over the strongest NI-BHD baseline, ELECTRA-large supervised classifier (0.795), is 0.116 (95% source-question cluster-bootstrap CI [0.106, 0.127]).

Ablation Study. Selection comparisons use four readers and the same additive model (Table 2). Individual-signal selection uses standalone development performance; label-free diversity favors low score correlations. All probes are supervised. Mean AUROCs are 0.911 (conditional gain), 0.896 (individual signal), 0.880 (diversity), and 0.887 (random sets). This supports selecting readers for conditional contributions. With selected readers fixed, cell-macro AUROCs are 0.911 (additive stacking), 0.869 (uniform averaging), 0.746 (hard routing), 0.911 (MLP), and 0.912 (score–covariate interactions). Score-only reaches 0.871; likelihood-only reaches 0.756. Additive fusion performs comparably to the MLP and interaction model. The full feature set improves AUROC over both restricted variants. Appendices D.5 and E.1 provide controls and feature analyses.

Further Analysis. READ also outperforms a four-encoder control under the same supervision and additive stacking protocol (0.911 versus 0.825 cell-macro AUROC). This supports the value of reader evidence over the evaluated text features (Appendix D.5). To test transfer to unseen benchmarks, we select readers and train the detector on six benchmarks, then evaluate it on the remaining one. Repeating this with each benchmark held out yields 0.862 cell-macro AUROC, showing that READ remains effective beyond the training benchmarks (Appendix E.3). Additional experiments on answer recovery, reader-family removal, and factual perturbations appear in Appendix E.

6 Conclusion

We study NI-BHD, assessing existing responses without target internals or additional target queries. Reader utility varies across targets and benchmarks; model attributes are unreliable selection guides in the evaluated pool. READ selects complementary readers by conditional detection value and uses fixed additive integration. Using evaluated targets' labeled responses, it surpasses all evaluated NI-BHD baselines in overall AUROC and the white-box per-environment best in 19 of 28 environments.

AI USE STATEMENT

Large language models are research subjects in this work, and their use is described in the relevant methodological and experimental sections. Beyond this research use, generative AI assistance was limited to language polishing. All AI-assisted edits were reviewed by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

To the best of our knowledge, this work does not raise specific ethical concerns requiring additional discussion under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Appendix D documents the datasets, response collection and labeling procedures, baseline implementations, training protocol, and hyperparameter settings. Details of the reader comparisons, ablations, and extended experiments are provided in Appendices D.5 and E. Appendix A presents the assumptions, derivations, and proofs underlying our theoretical analysis. Our implementation is available in the anonymous repository at https://anonymous.4open.science/r/read-C024.

REFERENCES

Anthropic. Introducing Claude 4, 2025. URL https://www.anthropic.com/news/claude-4.

Jacob Austin, Augustus Odena, Maxwell Nye, et al. Program synthesis with large language models, 2021. arXiv preprint arXiv:2108.07732.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Proc. of EMNLP Findings*, pp. 967–976, 2023.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, et al. Hallucination detection in LLMs using spectral features of attention maps. In *Proc. of EMNLP*, pp. 24354–24385, 2025.

Gavin Brown, Adam Pocock, Ming-Jie Zhao, et al. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.*, pp. 27–66, 2012.

Chao Chen, Kai Liu, Ze Chen, et al. INSIDE: llms' internal states retain the power of hallucination detection. In *Proc. of ICLR*, 2024.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, et al. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proc. of ICLR*, 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems, 2021. arXiv preprint arXiv:2110.14168.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, et al. Detecting hallucinations in large language models using semantic entropy. *Nature*, pp. 625–630, 2024.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, et al. RARR: Researching and revising what language models say, using language models. In *Proc. of ACL*, pp. 16477–16508, 2023.

Google DeepMind. Gemini 3.6 Flash: Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/.

Emma Harvey, Allison Koenecke, and René F. Kizilcec. "don't forget the teachers": Towards an educator-centered understanding of harms from large language models in education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pp. 1064:1–1064:19, 2025.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, et al. Deberta: decoding-enhanced bert with disentangled attention. In *Proc. of ICLR*, 2021.

Lei Huang, Weijiang Yu, Weitao Ma, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems*, pp. 1–55, 2025.

Mandar Joshi, Eunsol Choi, Daniel Weld, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proc. of ACL*, pp. 1601–1611, 2017.

Haoqiang Kang and Xiao-Yang Liu. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv*, 2023.

Seungone Kim, Juyoung Suk, Shayne Longpre, et al. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proc. of EMNLP*, pp. 4334–4353, 2024.

Jannik Kossen, Jiatong Han, Muhammed Razzak, et al. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. arXiv preprint arXiv:2406.15927.

Junyi Li, Xiaoxue Cheng, Xin Zhao, et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proc. of EMNLP*, pp. 6449–6464, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022a.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. of ACL*, pp. 3214–3252, 2022b.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Zuxin Liu, Thai Hoang, Jianguo Zhang, et al. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. In *Proc. of NeurIPS*, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proc. of EMNLP*, pp. 9004–9017, 2023.

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, Kalpesh Krishna, Xinxi Lyu, et al. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. of EMNLP*, pp. 12076–12100, 2023.

OpenAI. Introducing GPT-4.1 in the API, 2025. URL https://openai.com/index/gpt-4-1/.

Hadas Orgad, Michael Toker, Zorik Gekhman, et al. Llms know more than they show: On the intrinsic representation of LLM hallucinations. In *Proc. of ICLR*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *TACL*, pp. 249–266, 2019.

Dimitri Roustan, François Bastardot, et al. The clinicians' guide to large language models: A general perspective with a focus on hallucinations. *Interactive Journal of Medical Research*, 2025.

Ola Shorinwa, Zhiting Mei, Justin Lidard, et al. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions, 2024. arXiv preprint arXiv:2412.05563.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, et al. Llm-check: Investigating detection of hallucinations in large language models. In *Proc. of NeurIPS*, 2024.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proc. of ACL*, pp. 2526–2547, 2025.

Jerry Wei, Chengrun Yang, Xinying Song, et al. Long-form factuality in large language models. In *Proc. of NeurIPS*, 2024.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, et al. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In *Proc. of ACL*, pp. 9333–9347, 2024.

Tianhang Zhang, Lin Qiu, Qipeng Guo, et al. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proc. of EMNLP*, pp. 915–932, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proc. of NeurIPS*, 2023.

APPENDIX CONTENTS

A Theoretical Analysis of Conditional Reader Value **14**

A.1 Reference Distribution and Fixed Evidence Maps 14

A.2 Log Loss and Conditional Information 14

A.3 Individual Signal, Redundancy, and Synergy 15

A.4 Feature Compression and Set Summaries 15

A.5 Conditional Value under Additive Prediction 16

A.6 A Gaussian Illustration of Reader Complementarity 17

B Related Work **18**

C Limitations **19**

D Experimental Setup Details and Hyperparameter Selection **19**

D.1 Baseline Implementation Details 19

D.2 LLM-as-Judge Labeling Protocol 21

D.3 Dataset Details 21

D.4 Training Protocol and Hyperparameter Selection 23

D.5 Details of Reader Comparisons and Ablations 24

E Extended Experiments **26**

E.1 Reader Budget and Representation Choices 26

E.2 Detection under Answer-Recovery Failure 29

E.3 Generalization across Targets and Benchmarks 31

E.4 Robustness to Reader-Family Removal 34

E.5 Sensitivity to Factual Perturbations 34

A Theoretical Analysis of Conditional Reader Value

This appendix derives conditional reader value under log loss and analyzes how evidence compression and dependence affect the contribution of additional readers.

A.1 Reference Distribution and Fixed Evidence Maps

Fix an environment $e = (m, b)$ with distribution $\mathbb{P}_e$, where $Y = 1$ denotes hallucination. For a reader set $\mathcal{S}$, let $X_\mathcal{S}$ collect the evidence blocks defined in Equation 6, and let U denote the reader-independent answer-length covariate. The analysis conditions on the training data and fitted feature extractors, so all evidence maps are fixed.

Proposition 2 (Class-balanced reference measure)**.** *Suppose* $0 < \pi_e := \mathbb{P}_e(Y = 1) < 1$. *The unique measure assigning probability* $1/2$ *to each label while preserving both class-conditional laws is*

$$\mathbb{Q}_e(A) = \frac{1}{2}\mathbb{P}_e(A \mid Y = 0) + \frac{1}{2}\mathbb{P}_e(A \mid Y = 1). \tag{11}$$

Its density with respect to $\mathbb{P}_e$ *satisfies*

$$\frac{d\mathbb{Q}_e}{d\mathbb{P}_e} = \frac{1}{2\mathbb{P}_e(Y = y)} \qquad \text{on } \{Y = y\}, \quad y \in \{0, 1\}.$$

Proof. The mixture in Equation 11 is a probability measure with the required class masses and conditional laws. Conversely, the law of total probability implies that every measure with these properties equals this mixture. On $\{Y = y\}$, the reweighting factor is $1/\{2\mathbb{P}_e(Y = y)\}$, yielding the density formula. □

The following proposition characterizes the effect of class balancing on conditional independence.

Proposition 3 (Conditional independence under label reweighting)**.** *For* $r \notin \mathcal{S}$,

$$Y \perp Z_r \mid U, X_\mathcal{S}$$

holds under $\mathbb{P}_e$ *if and only if it holds under* $\mathbb{Q}_e$.

Proof. Write $Z = (U, X_\mathcal{S})$ and $a_y = 1/\{2\mathbb{P}_e(Y = y)\} > 0$. For either $W = Z$ or $W = (Z, Z_r)$,

$$\mathbb{Q}_e(Y = 1 \mid W) = F\{\mathbb{P}_e(Y = 1 \mid W)\}, \qquad F(u) = \frac{a_1 u}{a_1 u + a_0(1 - u)}.$$

The function F is one-to-one on $[0, 1]$. For a binary label, conditional independence is equivalent to equality between the posterior given (Z, Z_r) and the posterior given Z. The transformation F preserves this equality in both directions. Both reweighting factors are strictly positive, so $\mathbb{P}_e$ and $\mathbb{Q}_e$ have the same null sets. □

A.2 Log Loss and Conditional Information

We use natural logarithms, so information is measured in nats. For binary log loss

$$\ell_{\log}(y, p) = -y \log p - (1 - y) \log(1 - p),$$

define

$$R_e^*(\mathcal{S}) = \inf_f \mathbb{E}_{\mathbb{Q}_e}\left[\ell_{\log}\{Y, f(U, X_\mathcal{S})\}\right],$$

where f ranges over all measurable predictors into $[0, 1]$, with the usual limiting conventions at the endpoints. The entropy characterization of optimal log loss gives the following reader-value identity (Brown et al., 2012).

Theorem 4 (Conditional reader value)**.** *For any* $r \notin \mathcal{S}$,

$$R_e^*(\mathcal{S}) - R_e^*(\mathcal{S} \cup \{r\}) = I_{\mathbb{Q}_e}(Y; Z_r \mid U, X_\mathcal{S}) = V_e(r \mid \mathcal{S}) \geq 0. \tag{12}$$

The value is zero exactly when

$$Y \perp Z_r \mid U, X_\mathcal{S},$$

up to null events.

Proof. Let $Z = (U, X_{\mathcal{S}})$ and $\eta(Z) = \mathbb{Q}_e(Y = 1 \mid Z)$. Conditioning on Z gives

$$\mathbb{E}_{\mathbb{Q}_e}[\ell_{\log}\{Y, f(Z)\}] = H_{\mathbb{Q}_e}(Y \mid Z) + \mathbb{E}_{\mathbb{Q}_e}\left[\mathrm{KL}\left\{\mathrm{Bern}(\eta(Z)) \,\|\, \mathrm{Bern}(f(Z))\right\}\right].$$

The KL term is nonnegative and vanishes at $f = \eta$. Hence

$$R_e^*(\mathcal{S}) = H_{\mathbb{Q}_e}(Y \mid U, X_{\mathcal{S}}).$$

Applying the same argument after adding Z_r and subtracting the two conditional entropies yields Equation 12. Conditional mutual information is nonnegative and equals zero exactly under the stated conditional independence. □

A.3 Individual Signal, Redundancy, and Synergy

Define the signed redundancy of reader r relative to $\mathcal{S}$ as

$$R_e(r \mid \mathcal{S}) := I_{\mathbb{Q}_e}(Y; Z_r \mid U) - I_{\mathbb{Q}_e}(Y; Z_r \mid U, X_{\mathcal{S}}). \tag{13}$$

Corollary 5 (Signal minus signed redundancy)**.** *For $r \notin \mathcal{S}$,*

$$\begin{aligned} V_e(r \mid \mathcal{S}) &= I_{\mathbb{Q}_e}(Y; Z_r \mid U) - R_e(r \mid \mathcal{S}), \\ R_e(r \mid \mathcal{S}) &\le I_{\mathbb{Q}_e}(Y; Z_r \mid U). \end{aligned} \tag{14}$$

Positive redundancy means that conditioning on the selected evidence reduces the candidate's label information. Negative redundancy represents synergy: the candidate becomes more informative given the selected evidence.

Proof. The first identity follows by rearranging Equation 13. The inequality follows from $V_e(r \mid \mathcal{S}) \ge 0$. The sign interpretation follows by comparing the two information quantities in the definition. □

When the two feature–feature conditional mutual informations below are finite, the chain rule gives the equivalent expression

$$R_e(r \mid \mathcal{S}) = I_{\mathbb{Q}_e}(Z_r; X_{\mathcal{S}} \mid U) - I_{\mathbb{Q}_e}(Z_r; X_{\mathcal{S}} \mid U, Y). \tag{15}$$

This form relates signed redundancy to the change in dependence between reader evidence blocks after conditioning on the label.

A synergistic example. Take U constant and let B_1, B_2 be independent bits with $\mathbb{P}(B_i = 1) = q$, and $Y = B_1 \oplus B_2$. Writing H for binary entropy in bits,

$$I(Y; B_2) = H(2q(1-q)) - H(q), \qquad I(Y; B_2 \mid B_1) = H(q).$$

At $q = 0.15$ these are 0.209 and 0.610 bits. Each bit alone is weakly informative, while the pair determines the label, so the contribution of B_2 nearly triples once B_1 is observed and the signed redundancy is -0.401 bits. At $q = 0.15$, both bits carry positive marginal information, and the contribution of either bit increases after the other is observed.

A.4 Feature Compression and Set Summaries

Lemma 6 (Compression under fixed conditioning)**.** *Let X be a candidate reader's full output, let $E = g(X)$ be a deterministic measurable summary, and fix the conditioning variables Z. Then*

$$I_{\mathbb{Q}_e}(Y; E \mid Z) \le I_{\mathbb{Q}_e}(Y; X \mid Z),$$

with equality exactly when $Y \perp X \mid Z, E$.

Proof. Since E is a function of X, the chain rule gives

$$I_{\mathbb{Q}_e}(Y; X \mid Z) = I_{\mathbb{Q}_e}(Y; E \mid Z) + I_{\mathbb{Q}_e}(Y; X \mid Z, E).$$

The final term is nonnegative and vanishes exactly under the stated conditional independence. □

For fixed incumbent evidence $Z = (U, X_S)$, Lemma 6 bounds the conditional information retained by the candidate's three-dimensional evidence block. Equality holds when this block preserves all label-relevant information in the candidate's full output given Z.

Lemma 7 (Deterministic feature augmentation)**.** *Let $G_S = g_S(U, X_S)$ be measurable. For either $\nu = \mathbb{P}_e$ or $\nu = \mathbb{Q}_e$,*

$$\begin{aligned} H_\nu(Y \mid U, X_S, G_S) &= H_\nu(Y \mid U, X_S), \\ I_\nu(Y; Z_r \mid U, X_S, G_S) &= I_\nu(Y; Z_r \mid U, X_S). \end{aligned} \tag{16}$$

Proof. Measurability implies

$$\sigma(U, X_S, G_S) = \sigma(U, X_S).$$

The conditional quantities on both sides therefore use the same information. □

The lemma also applies after adding a reader and recomputing the set summaries. In the logistic model, dispersion summaries supply additional basis functions of the reader evidence. The classifier remains linear in the resulting constructed feature vector.

A.5 Conditional Value under Additive Prediction

Fix a reference distribution ν and the fitted evidence maps. Taking $\nu = \mathbb{Q}_e$ gives the environment-specific analysis above; for target-wise selection, take ν to be the pooled benchmark distribution for that target. Let $\mathcal{G}_S$ be the class of logistic predictors using the feature construction in Equation 10, with freely varying intercept and coefficients. For a nonempty set, these features comprise the reader evidence, answer length, and deterministic dispersion summaries, with standardization held fixed. For the empty set, $\mathcal{G}_\varnothing$ consists of constant predictors, matching the selection baseline.

Define the optimal predictive log loss within this class and its approximation gap by

$$L_\nu^{\mathrm{add}}(S) = \inf_{f \in \mathcal{G}_S} \mathbb{E}_\nu \left[\ell_{\log}\{Y, f(U, X_S)\}\right], \tag{17}$$

$$\mathcal{A}_\nu(S) = L_\nu^{\mathrm{add}}(S) - H_\nu(Y \mid U, X_S) \geq 0. \tag{18}$$

Proposition 8 (Conditional gain within the additive class)**.** *For $r \notin S$, let*

$$\Delta_\nu^{\mathrm{add}}(r \mid S) = L_\nu^{\mathrm{add}}(S) - L_\nu^{\mathrm{add}}(S \cup \{r\}).$$

Then

$$\begin{aligned} \Delta_\nu^{\mathrm{add}}(r \mid S) = I_\nu(Y; Z_r \mid U, X_S) \\ + \mathcal{A}_\nu(S) - \mathcal{A}_\nu(S \cup \{r\}). \end{aligned} \tag{19}$$

For $\nu = \mathbb{Q}_e$, the conditional-information term is $V_e(r \mid S)$.

Proof. The log-loss decomposition in the proof of Theorem 4 applies under ν. Writing $\eta_S = \nu(Y = 1 \mid U, X_S)$, taking the infimum over $\mathcal{G}_S$ gives

$$\mathcal{A}_\nu(S) = \inf_{f \in \mathcal{G}_S} \mathbb{E}_\nu \left[\mathrm{KL}\left\{\mathrm{Bern}(\eta_S) \,\|\, \mathrm{Bern}\big(f(U, X_S)\big)\right\}\right] \geq 0.$$

Thus each class-optimal loss equals conditional entropy plus its approximation gap. Subtracting this expression for the two reader sets, and using

$$H_\nu(Y \mid U, X_S) - H_\nu(Y \mid U, X_S, Z_r) = I_\nu(Y; Z_r \mid U, X_S),$$

proves Equation 19. □

The additive model's conditional gain therefore depends on both the information supplied by the candidate and the change in how well the model class represents the label posterior. If the posterior belongs to the corresponding logistic class for both sets, both approximation gaps vanish and the gain equals conditional information.

Connection to cross-validated selection. For the pooled reference distribution corresponding to the selection data, define

$$\varepsilon_\nu(\mathcal{S}) = \widehat{L}_{\mathrm{dev}}(\mathcal{S}) - L_\nu^{\mathrm{add}}(\mathcal{S}), \tag{20}$$

where $\widehat{L}_{\mathrm{dev}}$ is the cross-validated loss in Equation 9. This discrepancy includes the effects of fitting and regularization as well as validation relative to the class-optimal population loss. Substituting Equation 20 into the empirical gain gives

$$\begin{aligned}\widehat{\Delta}_{\mathrm{dev}}(r \mid \mathcal{S}) = \Delta_\nu^{\mathrm{add}}(r \mid \mathcal{S}) \\ + \varepsilon_\nu(\mathcal{S}) - \varepsilon_\nu(\mathcal{S} \cup \{r\}).\end{aligned} \tag{21}$$

Together, Equations 19 and 21 connect the implemented selection score to conditional information through the additive model's approximation gaps and the fitting and validation discrepancies.

A.6 A Gaussian Illustration of Reader Complementarity

Consider a Gaussian evidence model with constant U and

$$X \mid Y = y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \qquad \boldsymbol{\Sigma}_y \succ 0.$$

Write

$$\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \qquad \bar{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1}{2},$$

and let Φ denote the standard normal cumulative distribution function.

Proposition 9 (Optimal linear AUROC)**.** *For any* $\mathbf{w} \neq \mathbf{0}$*,*

$$\mathrm{AUROC}(\mathbf{w}^\top X) = \Phi\left(\frac{\mathbf{w}^\top \boldsymbol{\delta}}{\sqrt{2\mathbf{w}^\top \bar{\boldsymbol{\Sigma}} \mathbf{w}}}\right). \tag{22}$$

If $\boldsymbol{\delta} \neq \mathbf{0}$*, the maximizing directions are*

$$\mathbf{w} = c\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta}, \qquad c > 0,$$

and the maximum is $\Phi(\sqrt{d^2/2})$*, where*

$$d^2 = \boldsymbol{\delta}^\top \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\delta}.$$

If $\boldsymbol{\delta} = \mathbf{0}$*, every nonzero linear score has AUROC* $1/2$*.*

Proof. For independent class-conditional draws X_1, X_0, the score difference $\mathbf{w}^\top(X_1 - X_0)$ is Gaussian with mean $\mathbf{w}^\top\boldsymbol{\delta}$ and variance $2\mathbf{w}^\top\bar{\boldsymbol{\Sigma}}\mathbf{w}$. Its probability of being positive yields Equation 22.

Furthermore,

$$\frac{\mathbf{w}^\top \boldsymbol{\delta}}{\sqrt{\mathbf{w}^\top \bar{\boldsymbol{\Sigma}} \mathbf{w}}} = \frac{(\bar{\boldsymbol{\Sigma}}^{1/2}\mathbf{w})^\top (\bar{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\delta})}{\|\bar{\boldsymbol{\Sigma}}^{1/2}\mathbf{w}\|}.$$

Cauchy–Schwarz bounds this quantity by $\sqrt{d^2}$, with equality when $\mathbf{w} = c\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta}$ for $c > 0$. The case $\boldsymbol{\delta} = \mathbf{0}$ follows directly from Equation 22. □

Proposition 10 (Schur-complement increment)**.** *Partition* $X = (X_{\mathcal{S}}^\top, Z_r^\top)^\top$ *and write*

$$\boldsymbol{\delta} = \begin{pmatrix}\boldsymbol{\delta}_{\mathcal{S}} \\ \boldsymbol{\delta}_r\end{pmatrix}, \qquad \bar{\boldsymbol{\Sigma}} = \begin{pmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B}\end{pmatrix}.$$

Define

$$\mathbf{q} = \boldsymbol{\delta}_r - \mathbf{C}^\top \mathbf{A}^{-1}\boldsymbol{\delta}_{\mathcal{S}}, \qquad \mathbf{V} = \mathbf{B} - \mathbf{C}^\top \mathbf{A}^{-1}\mathbf{C},$$

and

$$d_{\mathcal{S}}^2 = \boldsymbol{\delta}_{\mathcal{S}}^\top \mathbf{A}^{-1}\boldsymbol{\delta}_{\mathcal{S}}.$$

Then

$$d_{\mathcal{S}\cup\{r\}}^2 - d_{\mathcal{S}}^2 = \mathbf{q}^\top \mathbf{V}^{-1}\mathbf{q} =: \Delta_e^{\mathrm{lin}}(r \mid \mathcal{S}) \geq 0, \tag{23}$$

with equality exactly when $\mathbf{q} = \mathbf{0}$*. A positive increment strictly increases optimal linear AUROC. For the empty set, take* $d_\varnothing^2 = 0$*,* $\mathbf{q} = \boldsymbol{\delta}_r$*, and* $\mathbf{V} = \mathbf{B}$*.*

Proof. Positive definiteness implies $\mathbf{A} \succ 0$ and $\mathbf{V} \succ 0$. The block inverse is

$$\bar{\mathbf{\Sigma}}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}\mathbf{V}^{-1}\mathbf{C}^\top\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{V}^{-1} \\ -\mathbf{V}^{-1}\mathbf{C}^\top\mathbf{A}^{-1} & \mathbf{V}^{-1} \end{pmatrix}.$$

Substitution into $\boldsymbol{\delta}^\top \bar{\mathbf{\Sigma}}^{-1}\boldsymbol{\delta}$ gives

$$d^2_{\mathcal{S}\cup\{r\}} = d^2_{\mathcal{S}} + \mathbf{q}^\top \mathbf{V}^{-1}\mathbf{q}.$$

Since $\mathbf{V} \succ 0$, the increment is nonnegative and vanishes exactly when $\mathbf{q} = \mathbf{0}$. The AUROC statement follows from the strict monotonicity of $\Phi(\sqrt{d^2/2})$ in d^2. □

Common covariance and conditional information. Suppose $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \mathbf{\Sigma}$. The conditional distribution of Z_r given $X_{\mathcal{S}} = x$ and $Y = y$ has covariance $\mathbf{V}$ and mean

$$\boldsymbol{\mu}_{r,y} + \mathbf{C}^\top \mathbf{A}^{-1}(x - \boldsymbol{\mu}_{\mathcal{S},y}).$$

The difference between its two class-conditional means is $\mathbf{q}$. Consequently,

$$\mathbf{q} = \mathbf{0} \iff Y \perp Z_r \mid X_{\mathcal{S}} \iff I_{\mathbb{Q}_e}(Y; Z_r \mid X_{\mathcal{S}}) = 0.$$

Under this model, the Schur increment and conditional mutual information therefore have the same zero condition.

A candidate with positive conditional value. Consider scalar incumbent and candidate features with balanced labels, $d \neq 0$, and common covariance

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad 0 < |\rho| < 1.$$

Let $\boldsymbol{\delta} = (d, 0)^\top$. The candidate has identical marginal distributions under the two labels, so $I_{\mathbb{Q}_e}(Y; Z_r) = 0$. Its conditional increment is nevertheless

$$\mathbf{q} = -\rho d, \qquad \Delta^{\mathrm{lin}}_e(r \mid \mathcal{S}) = \frac{\rho^2 d^2}{1 - \rho^2} > 0.$$

The common-covariance relation gives positive conditional mutual information and negative signed redundancy. Here the candidate helps resolve correlated variation in the incumbent feature.

A candidate with redundant marginal signal. Under the same covariance, let $\boldsymbol{\delta} = (d, \rho d)^\top$. The candidate has positive marginal separation $\rho^2 d^2$, but

$$\mathbf{q} = \rho d - \rho d = 0.$$

Its conditional distribution given the incumbent is identical across labels. Hence

$$I_{\mathbb{Q}_e}(Y; Z_r \mid X_{\mathcal{S}}) = 0, \qquad \Delta^{\mathrm{lin}}_e(r \mid \mathcal{S}) = 0.$$

B Related Work

White-Box Hallucination Detection. White-box methods detect hallucinations using information from the target model, including token probabilities, hidden states, and attention maps. Supervised probes learn to predict truthfulness from internal activations, while semantic entropy probes estimate semantic uncertainty from the hidden states of a single response (Azaria & Mitchell, 2023; Kossen et al., 2024). Other studies examine how factual information is represented internally or derive detection signals from attention structures (Orgad et al., 2025; Binkowski et al., 2025). These approaches provide useful detection signals but require access to target-model computations, which is unavailable when the target exposes only a text-generation interface.

Black-Box Hallucination Detection. Black-box methods assess model responses without accessing the target's internal states. SelfCheckGPT compares an answer with additional stochastic generations, while InterrogateLLM uses auxiliary queries to assess the reliability of generated content (Manakul et al., 2023; Yehuda et al., 2024). Such methods require further interactions with the target. Alternatively, an accessible auxiliary model can evaluate a fixed response through prompted judgments or through its own probabilities and representations (Zheng et al., 2023; Zhang et al., 2023; Sriramanan et al., 2024). In particular, LLM-Check uses teacher forcing to obtain detection signals from an auxiliary model when target internals are unavailable. Our NI-BHD setting focuses on existing responses without further target queries. Within this setting, we study the detection value of different readers and select complementary reader evidence for supervised integration.

Retrieval-Based Factual Verification. Another class of black-box methods retrieves external evidence to verify generated content. FActScore decomposes a response into atomic facts and estimates their support using retrieval and a language model (Min et al., 2023). SAFE verifies individual facts through web search, while RARR retrieves attribution evidence and revises unsupported content (Wei et al., 2024; Gao et al., 2023). These methods introduce question-specific evidence beyond the input supplied to the detector. On publicly released benchmarks, retrieved material may include the original question and its reference answer or solution. When this occurs, the verifier gains privileged access to the answer key used for evaluation and can identify errors through direct comparison. Measured performance then reflects access to benchmark solutions as well as detection quality. We therefore exclude external retrieval from our comparisons to avoid this source of answer leakage and keep reference answers on the annotation side of the evaluation.

C Limitations

Our theoretical analysis does not introduce a new theory. It builds on established information-theoretic feature selection principles (Brown et al., 2012) to characterize the contribution of additional reader evidence. The main contribution lies in the NI-BHD formulation, the empirical findings about reader utility, and the detection framework developed from these findings. In the evaluated pool, model size, answering ability, and family match do not reliably guide reader selection, while different readers provide complementary evidence. These observations motivate READ's selection of readers by conditional predictive gain and its additive integration of their evidence. The theoretical analysis explains this design, while controlled experiments assess its effectiveness.

In the primary experimental setting, READ learns reader probes and a fusion classifier from labeled responses pooled across the evaluated targets. This introduces a supervision requirement compared with detectors that can be applied without task-specific training. Future work could reduce the required annotation effort through more sample-efficient learning. Semi-supervised training could also use unlabeled response archives alongside a smaller labeled set when adapting the detector to new targets.

The current detector assigns a hallucination score to a complete response. For long answers containing multiple claims, identifying the specific statements that require verification would provide more useful feedback. A natural extension is to use complementary reader evidence for sentence- or claim-level localization, helping users identify which parts of an answer need further review and correction.

D Experimental Setup Details and Hyperparameter Selection

D.1 Baseline Implementation Details

All baselines are evaluated on the same fixed question–answer pairs and hallucination labels as READ. Trainable baselines use the same source-question-grouped splits. Each method retains the access required by its scoring rule: target-internal methods are evaluated on the four open-weight targets, while answer-resampling methods use a budget of ten additional generations per question.

Target Likelihood. We use the length-normalized log-likelihood of the observed answer as a confidence baseline (Kossen et al., 2024). The fixed answer is scored by teacher forcing with the target model's own tokenizer, and its negative mean token log-likelihood serves as the hallucination score.

Semantic Entropy. Semantic entropy measures uncertainty across semantically distinct answers (Farquhar et al., 2024). We include both probability-weighted and discrete variants. The former uses target likelihoods to estimate semantic-cluster masses, whereas the latter uses the empirical frequencies of sampled answers within each cluster. These methods produce a question-level uncertainty score, which is evaluated against the label of the fixed response.

EigenScore. EigenScore measures semantic diversity through the covariance spectrum of target-model representations of sampled answers (Chen et al., 2024). Like semantic entropy, it provides a question-level score that is evaluated against the fixed response's label.

Attention-Laplacian. This baseline uses spectral features of graph Laplacians constructed from the target's attention maps (Binkowski et al., 2025). The features are extracted from the fixed question–answer sequence and used to score hallucination.

SAPLMA and Mass-Mean Probes. SAPLMA predicts truthfulness from target hidden states (Azaria & Mitchell, 2023), while the mass-mean probe uses a direction derived from the class-conditional mean representations (Marks & Tegmark, 2024). We fit these probes separately for each target using the training responses and labels in our evaluation. This applies their probing constructions to the response population studied here.

Semantic-Entropy Probe. The semantic-entropy probe predicts semantic uncertainty from a target model's hidden states (Kossen et al., 2024). For each target, we construct its training signal from discrete semantic entropy on the training questions. The fitted probe then scores individual fixed responses using their hidden states. Its supervision therefore comes from the target's sampled answer distribution rather than the hallucination labels used to train truthfulness probes.

Lexical Similarity and SelfCheckGPT. Both methods compare the fixed answer with additional answers sampled from the same target. Lexical similarity measures surface-form agreement, whereas SelfCheckGPT evaluates consistency between the response and the sampled alternatives (Manakul et al., 2023). Lower agreement indicates greater hallucination risk.

InterrogateLLM. InterrogateLLM reconstructs the original question from the generated answer and measures agreement between the reconstructed questions and the original (Yehuda et al., 2024). We apply this reverse-questioning procedure to the fixed response under evaluation.

Verbalized Confidence. This baseline prompts the target model to report its confidence in the fixed answer (Lin et al., 2022a; 2024). The reported confidence is converted to a hallucination score by reversing its direction.

External LLM Critics. We evaluate Gemini 3.6 Flash (Google DeepMind, 2026), Qwen3.5-27B and Qwen3.5-9B (Qwen Team, 2026), and Prometheus-2 (7B) (Kim et al., 2024) as prompted critics. Each receives the fixed question and answer, including any passage supplied as part of the question, and returns a numerical assessment of hallucination. Reference answers are reserved for constructing the evaluation labels.

Text-Based Baselines. Answer length provides a simple surface-feature baseline, and TF–IDF logistic regression learns a classifier over the question–answer text. We also evaluate supervised classifiers based on DeBERTa-v3-large, ModernBERT-large, and ELECTRA-large (He et al., 2021; Warner et al., 2025; Clark et al., 2020). These classifiers use the same question–answer inputs and training labels as READ.

LLM-Check. We use the reader-side configuration of LLM-Check (Sriramanan et al., 2024). An auxiliary model processes the fixed question–answer pair by teacher forcing, and its internal fea-

tures provide the hallucination score. The representations therefore come from the auxiliary reader, allowing this baseline to evaluate both open-weight and API targets.

Baseline tuning and capacity. We use common evaluation data and metrics while allowing each baseline to use hyperparameters appropriate to its architecture and scoring rule. For applicable methods, we tune both model capacity and other hyperparameters separately on each dataset, and report the configuration achieving the best performance among those evaluated. The capacity search includes configurations larger than READ where applicable, giving these baselines the opportunity to benefit from additional capacity. When a reported baseline uses a smaller configuration, this reflects its observed performance within the search: increasing capacity further provides no improvement and in some cases reduces performance. Together, these choices support a fair comparison by allowing each baseline to use its best-performing evaluated configuration, while keeping evaluation inputs and metrics consistent across methods.

D.2 LLM-AS-JUDGE LABELING PROTOCOL

Labeling rationale. Answer correctness and hallucination are related but distinct. A response that only states that it does not know the answer fails to provide the reference answer, but does not thereby make a false factual claim. Treating every failure to match the reference answer as a hallucination would conflate explicit abstention with factual error. We therefore distinguish correct answers, hallucinated responses, and abstentions during annotation. This distinction allows the binary labels to identify hallucinatory content while retaining explicit abstentions as non-hallucinatory responses.

Labeling procedure. We use GPT-4.1 (OpenAI, 2025) as the primary labeling model under a fixed prompt and temperature 0. The judge evaluates each response using the question, reference answer, and available task evidence. It returns a structured decision—`correct`, `hallucination`, or `abstention`—with a brief justification. For binary detection, we map `hallucination` to 1 and both `correct` and `abstention` to 0. The negative class therefore represents non-hallucinatory responses, including both correct answers and explicit abstentions. Abstentions account for 197 of the 3,780 negatives (5.2%), or 2.6% of the primary dataset.

Label consistency. Labels assigned by a single LLM may reflect judge-specific errors or biases. We therefore assess their reliability through independent model and human judgments on 490 responses, with 70 per benchmark, stratified by target family and original three-way label. Claude Sonnet 4 (Anthropic, 2025) serves as an independent judge under the same labeling criteria. Two human annotators also label these responses independently, blinded to the generating model and previous judgments. A third annotator resolves their disagreements to obtain the adjudicated human labels.

Table 3 reports agreement before and after the binary mapping. Three-way agreement ranges from 95.1% to 97.1%, with most disagreements occurring between `correct` and `abstention`. After these categories are merged, binary agreement ranges from 98.8% to 99.6%, with Cohen's κ between 0.975 and 0.992. The primary and independent judges disagree with the adjudicated human binary labels on 2 and 2 of the 490 responses, respectively. Primary-judge agreement with human adjudication is also similar on the two targets used as labeling judges (100.0%) and the remaining four targets (99.4%). These results support the consistency of the binary labels used for detection.

D.3 DATASET DETAILS

We use seven benchmarks covering factual question answering, passage-grounded QA, mathematical reasoning, code generation, and tool calling. Passages and tool specifications supplied to the target remain part of the question input available to readers and labeling judges. For HaluEval-QA and CoQA, labeling assesses semantic support from the supplied passage. For the other five benchmarks, it assesses conflicts with accepted answers or task requirements.

TruthfulQA. TruthfulQA (Lin et al., 2022b) contains questions designed to elicit common misconceptions. We use the questions as open-ended generation prompts. The accepted answers and known incorrect alternatives provide reference information for semantic labeling.

Table 3: Label agreement on 490 audited responses. Agreement is reported as a percentage. Adjudicated labels are obtained from two independent human annotations, with disagreements resolved by a third annotator.

	Three-way		Binary	
Comparison	Agreement	κ	Agreement	κ
Primary vs. independent	95.1	0.908	99.2	0.984
Human 1 vs. human 2	96.3	0.930	98.8	0.975
Primary vs. adjudicated	96.3	0.931	99.6	0.992
Independent vs. adjudicated	97.1	0.946	99.6	0.992

HaluEval-QA. The QA subset of HaluEval (Li et al., 2023) provides questions with knowledge passages and paired correct and hallucinated answers. We retain the correct answer as the reference and discard the hallucinated answer, converting the subset into a standard passage-grounded QA dataset. Each target receives the question and knowledge passage and generates its own response for labeling.

TriviaQA. TriviaQA (Joshi et al., 2017) contains open-domain factual questions with annotated answers and alternative answer forms. We present each question without an accompanying evidence passage and retain the provided answer aliases as references for semantic labeling.

CoQA. CoQA (Reddy et al., 2019) is a conversational question-answering benchmark grounded in source passages. We use the first question–answer turn of each story, with the story as input context, the first question as the query, and its corresponding answer as the reference.

GSM8K. GSM8K (Cobbe et al., 2021) contains grade-school mathematical word problems requiring multi-step reasoning. The problem statement forms the model input, and the reference solution provides the mathematical basis for assessing the generated answer.

MBPP. MBPP (Austin et al., 2021) consists of short Python programming tasks with reference implementations and tests. Models generate code to satisfy the task description. The reference code and tests specify the intended functionality used to judge the generated program.

xLAM. The xLAM function-calling dataset (Liu et al., 2024) pairs user requests with tool descriptions and reference tool calls. The request and available tool specifications form the model input. The reference call provides the expected tool selection and arguments for labeling.

Candidate response collection. We collect target-generated responses using up to nine fixed prompt forms shared across targets and benchmarks, with greedy decoding at temperature 0. Additional prompt forms are used incrementally until the class-balanced sampling requirements can be met within each target–benchmark pair, subject to the nine-form limit. This procedure produces 40,220 candidate responses. After labeling, the observed hallucination rates across target–benchmark pairs range from 4% to 34%, with a median of 15%. At these rates, one response per source question can provide too few hallucinated examples to meet the per-class quota, motivating the use of multiple prompt forms.

Primary dataset construction. For each target–benchmark pair, we select 90 hallucinated and 90 non-hallucinated responses. Sampling is stratified by binary label and prompt form, with matching prompt-form distributions between the two classes within each pair. The resulting primary dataset contains 7,560 responses across 42 target–benchmark pairs and covers 2,014 distinct source questions, averaging approximately 3.8 retained responses per question across targets and prompt forms. The candidate pool records the observed class proportions before balancing, while the primary dataset contains equal numbers of hallucinated and non-hallucinated responses.

D.4 Training Protocol and Hyperparameter Selection

Outer evaluation and inner validation. We fit one detector over the responses of all targets, and report AUROC within each target–benchmark pair. The primary dataset is partitioned into four outer folds, grouped by source question. All responses derived from the same question, including different prompt forms and targets, remain in the same group. For each outer split, three folds are used for training and the remaining fold is held out for evaluation. The outer-training data are further divided into four source-question-grouped inner folds for layer and reader selection.

Probe fitting and layer selection. Each reader uses its own tokenizer and native message template; k_r includes the template tokens preceding the answer. Reader backbones remain frozen. For each reader, we consider relative layer depths $\{0.25, 0.50, 0.75, 1.00\}$. At each candidate depth, we average the hidden states used to predict the answer tokens and fit a logistic probe to the resulting representations. The probe is fitted on three inner folds and predicts the fourth; rotating the held-out fold produces an out-of-fold (OOF) score for every outer-training response. We select the depth with the lowest unweighted inner-CV log loss for each reader within each outer split. The selected OOF scores are then used for reader-set construction and final combiner fitting.

Reader-set construction. Starting from an empty set, we add the reader that gives the largest reduction in inner-CV log loss. For each candidate set, a logistic combiner is fitted on three inner folds using the selected OOF probe scores and the accompanying features defined in Equation 6. It is evaluated on the remaining fold, and losses are averaged over all inner-validation responses. The empty-set baseline predicts the empirical hallucination proportion in the outer-training data. Every nonempty set includes the reader evidence, answer length, and across-reader dispersion features. Selection continues until the fixed budget of $K = 4$ readers is reached. Layer and reader selection reuse the inner-validation results, while the outer fold provides the final evaluation of the selected pipeline.

Final fitting and evaluation. After reader selection, we fit the final combiner on the selected OOF evidence from all outer-training responses. The selected probes are then refitted at their chosen depths on the complete outer-training partition and applied to the held-out outer fold. Their scores and accompanying features are passed through the fitted combiner to obtain hallucination probabilities. Repeating this procedure across the four outer folds gives one held-out prediction per response. For the main results, we compute AUROC within each target–benchmark pair using these predictions and average the pair-level AUROCs with equal weights.

Fitting settings and score convention. Both probes and combiners use standardized ℓ_2-regularized logistic regression with class-balanced fitting weights. Standardization parameters are estimated from the training partition of each fit and held fixed when scoring its evaluation responses. We fix the inverse regularization strength at the library default $C = 1.0$ for both stages. For weighted mean log loss with penalty $\lambda\|\boldsymbol{\beta}\|_2^2/2$, $\lambda = 1/(CW)$, where W is the sum of fitting weights in that partition; the intercept is unpenalized. Layer and reader selection use unweighted validation log loss.

The shared answer-length feature counts text segments using a fixed regular expression applied to the answer string. Probe scores supplied to the combiner are linear decision values on the log-odds scale. Across-reader score dispersion is computed on this same scale. Sigmoid is applied when probabilities are required for loss evaluation or final prediction, with no additional calibration at either stage.

Statistical reporting. For the main performance differences, we compute 95% percentile confidence intervals using $1{,}000$ paired bootstrap resamples of source questions. All responses associated with a sampled question are resampled together across targets and prompt forms. Within each resample, we recompute both methods' cell-macro AUROCs from their held-out predictions and take their difference.

D.5 Details of Reader Comparisons and Ablations

The comparisons follow the source-question-grouped evaluation and OOF training protocol in Appendix D.4, with the specific settings described below.

Pairwise residual analysis. For response i in environment e, we define

$$e_{r,i} = y_i - \text{sigmoid}(\hat{\eta}_{r,i}), \tag{24}$$

where $\hat{\eta}_{r,i}$ is the linear decision value of the complete single-reader detector used to compute $M_{r,e}$. Predictions come from outer-fold evaluation, pooled across the four folds and restricted to environment e. We compute Pearson correlations between readers' residuals within each environment. Their association with the pairwise AUROC gains in Equation 5 is measured using Spearman correlation across reader-pair–environment combinations. Residual correlations range from 0.75 to 0.97. Within this range, lower residual correlations are associated with larger pairwise gains.

Reader-selection controls. The fixed-set control uses Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, and Qwen3-8B across all targets and splits. The largest-reader and best-answerer rules select the same four readers: Llama-3.1-8B, Qwen3-8B, Qwen3.5-9B, and gemma-3-12b-it. Both therefore achieve 0.877 AUROC, below the random-set mean of 0.887 (Table 2). This comparison supports the reader suitability mismatch discussed in Section 3.1.

Label-free diversity selects the four-reader set with the lowest mean absolute Pearson correlation across its six reader pairs, computed from OOF prediction probabilities on the outer-training data. This selection criterion uses no labels, while the underlying probes remain supervised. All sets use the same additive combiner.

Random-set results average eight draws within each outer split. The same-family single-reader control samples from the target's family and averages over repeated draws. Its results cover the 28 environments with same-family candidates; the 14 environments of the GPT and Claude targets are marked N/A.

Integration controls. The integration comparisons retain the four readers selected by READ. For the hard router, the reader with the lowest Bernoulli loss on each training response supplies the routing label, using OOF probe probabilities. A multiclass logistic regression predicts this label from token-likelihood summaries, answer length, and across-reader dispersion. At evaluation, it selects one reader and returns that reader's probe probability. The MLP uses the same input features as the additive combiner, with one hidden layer containing four ReLU units and a sigmoid output.

Distinct readers and repeated views. We test whether multiple views of a single model can provide detection performance comparable to that of distinct readers. Both repeated-view controls use Qwen3-4B, selected by the benchmark development rule. One combines probes from four layers; the other combines four probes fitted on separate bootstrap resamples of the training responses. Resampling provides distinct fits of the convex probe objective. Using the development-selected reader gives both controls a single-model reference chosen for its detection performance. All three configurations use additive logistic fusion, although their auxiliary features may differ.

Comparison with stacked text encoders. We test whether combining supervised text encoders yields performance comparable to combining reader evidence. The control uses DeBERTa-v3-large, ModernBERT-large, RoBERTa-large, and ELECTRA-large, with the same labeled question–answer pairs and source-question-grouped splits as READ. Their out-of-fold scores are combined using the same additive stacking procedure. As shown in Table 4, stacking improves AUROC over each individual encoder, reaching 0.825, while READ reaches 0.911. The difference is 0.086, with a 95% source-question cluster-bootstrap confidence interval of $[0.076, 0.097]$. This comparison supports the value of reader evidence beyond stacking the evaluated text encoders under the same supervision and fusion protocol.

Comparison with a single probe on concatenated reader states. We compare the two-stage construction with a single logistic probe on the concatenated answer-span representations of the four selected readers. Both use the same readers, selected depths, and source-question-grouped

Table 4: Individual and stacked text encoders compared with READ. AUROC is the cell-macro average over the 42 target–benchmark pairs.

Method	Models	AUROC
DeBERTa-v3-large	1	0.784
ModernBERT-large	1	0.785
RoBERTa-large	1	0.728
ELECTRA-large	1	0.795
Four-encoder stack	4	0.825
READ	4	0.911

splits. We also evaluate a concatenated-state probe with the same answer-token likelihood features and answer length. Table 5 reports these controls and the two-stage feature ablations. Using hidden-state features alone, the concatenated-state probe and score-only stacking achieve similar AUROC (0.873 and 0.871, respectively). With likelihood features and answer length, the concatenated probe reaches 0.896, compared with 0.911 for the full two-stage detector. Removing dispersion features from the two-stage detector reduces AUROC to 0.897. These comparisons support the usefulness of the dispersion features in the two-stage design.

Table 5: The two-stage construction against a single probe on concatenated reader states, and the second-stage feature groups. All rows use the same four selected readers, the same selected depths, and the same splits. AUROC is the cell-macro average over the 42 target–benchmark pairs.

Configuration	AUROC↑	Log loss↓
Concatenated reader states, one probe	0.873	0.447
Concatenated states + likelihood features	0.896	0.407
Two-stage, reader scores only	0.871	0.449
Two-stage, without dispersion features	0.897	0.407
Two-stage, full feature vector (READ)	**0.911**	0.384

Paired comparisons of reader-selection rules. Table 6 reports paired source-question cluster-bootstrap intervals for the AUROC differences between conditional-gain selection and the other construction rules. Within each resample, we recompute both methods' cell-macro AUROCs from their held-out predictions and take their difference. Random-set results are averaged over their draws. All three intervals exclude zero in favor of conditional-gain selection.

Table 6: Differences between READ and the other reader-set construction rules, with 95% source-question cluster-bootstrap intervals. A positive difference favours READ.

Compared configuration (K=4)	Difference	95% interval
Individual-signal selection	+0.016	[0.010, 0.021]
Random reader set	+0.024	[0.018, 0.030]
Fixed reader set	+0.041	[0.034, 0.048]

The standalone matrix. Section 3 summarises the matrix $\mathbf{M}$ through win counts, rank agreement and correlations with model attributes. Tables 7 and 8 report it in full, so those summaries can be recomputed. A two-way decomposition of its total variation attributes 10.4% to which reader is used, 19.7% to which environment is evaluated, and 69.9% to the reader–environment interaction. The reader means span 0.060, while the gap between the best and worst reader inside a single environment averages 0.178. Which reader is stronger on average therefore explains little; which readers suit a given environment explains most of it.

Table 7: Standalone detection AUROC $M_{r,e}$ for every reader, on the first three targets. Rows are ordered by parameter count, and the panels follow Table 1.

Reader	Llama-3.3-70B							Qwen3-32B							Mistral-Small-24B						
	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM
Qwen3-0.6B	0.790	0.733	0.684	0.854	0.873	0.791	0.782	0.910	0.785	0.863	0.847	0.870	0.719	0.879	0.851	0.744	0.881	0.930	0.909	0.811	0.875
Llama-3.2-1B	0.872	0.747	0.793	0.722	0.905	0.799	0.881	0.797	0.742	0.728	0.741	0.853	0.788	0.770	0.841	0.780	0.843	0.708	0.831	0.820	0.890
Qwen3-1.7B	0.790	0.846	0.820	0.841	0.827	0.768	0.852	0.782	0.840	0.781	0.842	0.700	0.765	0.856	0.851	0.768	0.883	0.880	0.857	0.720	0.813
Ministral-3-3B	0.683	0.785	0.764	0.743	0.770	0.718	0.734	0.731	0.729	0.757	0.861	0.827	0.760	0.701	0.694	0.666	0.745	0.810	0.718	0.688	0.710
Llama-3.2-3B	0.822	0.729	0.710	0.681	0.807	0.726	0.690	0.754	0.725	0.719	0.733	0.793	0.711	0.729	0.781	0.771	0.771	0.787	0.721	0.704	0.706
Phi-4-mini-instruct	0.703	0.747	0.838	0.784	0.764	0.708	0.838	0.763	0.698	0.736	0.742	0.775	0.771	0.751	0.671	0.717	0.791	0.752	0.762	0.726	0.733
Qwen3-4B	0.841	0.794	0.867	0.754	0.829	0.759	0.816	0.826	0.825	0.836	0.846	0.917	0.783	0.867	0.875	0.774	0.871	0.867	0.848	0.750	0.843
Qwen3.5-4B	0.814	0.868	0.680	0.798	0.811	0.713	0.842	0.833	0.725	0.754	0.871	0.831	0.792	0.823	0.826	0.833	0.873	0.893	0.837	0.802	0.866
gemma-3-4b-it	0.798	0.657	0.768	0.882	0.901	0.839	0.825	0.689	0.753	0.787	0.770	0.855	0.683	0.831	0.668	0.757	0.750	0.857	0.766	0.730	0.748
Qwen3-8B	0.847	0.854	0.818	0.823	0.814	0.663	0.837	0.875	0.841	0.657	0.859	0.904	0.756	0.793	0.854	0.849	0.821	0.850	0.825	0.787	0.701
Llama-3.1-8B	0.814	0.825	0.747	0.741	0.827	0.733	0.775	0.679	0.747	0.683	0.773	0.811	0.718	0.714	0.747	0.821	0.761	0.796	0.727	0.769	0.828
Ministral-3-8B	0.735	0.849	0.620	0.726	0.850	0.778	0.705	0.750	0.763	0.694	0.754	0.834	0.735	0.693	0.658	0.785	0.797	0.873	0.768	0.751	0.717
Qwen3.5-9B	0.878	0.854	0.808	0.823	0.879	0.611	0.756	0.870	0.808	0.610	0.832	0.904	0.777	0.723	0.800	0.824	0.872	0.806	0.868	0.756	0.854
gemma-3-12b-it	0.739	0.863	0.809	0.723	0.800	0.771	0.864	0.710	0.755	0.708	0.747	0.866	0.739	0.733	0.708	0.705	0.739	0.814	0.746	0.708	0.797

Table 8: Standalone detection AUROC $M_{r,e}$ for the remaining three targets, with each reader's mean $\bar{M}_r$ over all 42 environments in the final column.

Reader	Gemma-3-27B							GPT-4.1							Claude-Sonnet-4							$\bar{M}_r$
	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	
Qwen3-0.6B	0.846	0.773	0.814	0.887	0.774	0.693	0.841	0.817	0.802	0.824	0.713	0.820	0.754	0.783	0.825	0.755	0.750	0.515	0.896	0.763	0.830	**0.806**
Llama-3.2-1B	0.796	0.776	0.862	0.759	0.720	0.855	0.836	0.835	0.769	0.824	0.760	0.882	0.898	0.928	0.879	0.795	0.708	0.627	0.835	0.837	0.862	**0.807**
Qwen3-1.7B	0.771	0.832	0.777	0.839	0.727	0.830	0.660	0.783	0.782	0.644	0.707	0.790	0.782	0.823	0.821	0.765	0.795	0.757	0.800	0.751	0.807	**0.793**
Ministral-3-3B	0.764	0.757	0.728	0.747	0.747	0.780	0.762	0.776	0.826	0.824	0.852	0.837	0.843	0.845	0.753	0.780	0.778	0.802	0.859	0.777	0.778	**0.767**
Llama-3.2-3B	0.764	0.871	0.841	0.758	0.742	0.796	0.727	0.734	0.808	0.753	0.710	0.738	0.856	0.805	0.748	0.729	0.736	0.733	0.793	0.797	0.779	**0.757**
Phi-4-mini-instruct	0.797	0.829	0.800	0.769	0.762	0.774	0.790	0.810	0.833	0.774	0.781	0.792	0.816	0.823	0.790	0.706	0.844	0.864	0.825	0.819	0.820	**0.776**
Qwen3-4B	0.872	0.755	0.854	0.828	0.854	0.700	0.856	0.780	0.779	0.753	0.795	0.768	0.772	0.827	0.849	0.831	0.804	0.758	0.903	0.738	0.831	**0.816**
Qwen3.5-4B	0.789	0.828	0.807	0.830	0.724	0.859	0.871	0.701	0.775	0.788	0.718	0.731	0.792	0.814	0.804	0.812	0.784	0.544	0.850	0.817	0.747	**0.797**
gemma-3-4b-it	0.729	0.789	0.753	0.757	0.686	0.751	0.667	0.773	0.851	0.825	0.772	0.853	0.831	0.853	0.729	0.766	0.757	0.769	0.829	0.795	0.764	**0.776**
Qwen3-8B	0.858	0.835	0.840	0.780	0.763	0.753	0.852	0.737	0.825	0.825	0.793	0.759	0.740	0.842	0.823	0.814	0.797	0.820	0.870	0.781	0.787	**0.808**
Llama-3.1-8B	0.759	0.829	0.732	0.746	0.697	0.753	0.745	0.773	0.841	0.791	0.734	0.765	0.855	0.838	0.768	0.720	0.776	0.710	0.799	0.825	0.751	**0.768**
Ministral-3-8B	0.757	0.795	0.798	0.792	0.654	0.768	0.785	0.692	0.857	0.814	0.830	0.851	0.720	0.818	0.852	0.822	0.793	0.809	0.787	0.823	0.747	**0.771**
Qwen3.5-9B	0.815	0.795	0.812	0.864	0.772	0.841	0.848	0.761	0.658	0.814	0.689	0.790	0.783	0.778	0.826	0.696	0.827	0.827	0.881	0.725	0.812	**0.798**
gemma-3-12b-it	0.761	0.745	0.722	0.707	0.709	0.734	0.752	0.736	0.819	0.840	0.766	0.811	0.808	0.827	0.708	0.718	0.739	0.702	0.781	0.762	0.739	**0.760**

E Extended Experiments

These experiments test whether the detection gains of READ depend on successful answer recovery, the training distribution, or access to a particular reader family, and whether its scores respond to factual changes in the supplied answer. We also examine how the reader budget and representation choices affect the evidence available to the detector. The experiments use the common training and evaluation procedures described earlier, with the changes needed for each comparison specified below.

E.1 Reader Budget and Representation Choices

We examine how much detection performance can be retained with a limited reader budget and how representation choices affect the evidence extracted from each reader. The budget experiment evaluates the complete detector, while the layer and pooling comparisons examine the first-stage reader probes.

Reader budget. The primary experiments use a preset budget of $K = 4$ to limit per-response reader computation. Each additional reader processes the question–answer pair with its own backbone, increasing the computation required for every prediction. We evaluate budgets from one to eight readers, together with the full pool of 14 readers, using the same conditional-gain selection rule and grouped evaluation protocol. Figure 4 reports cell-macro AUROC. Increasing the budget from one to four readers improves AUROC from 0.816 to 0.911. Further additions yield smaller improvements: eight readers reach 0.936, and the full pool reaches 0.943. Using all 14 readers achieves the highest AUROC in this sweep, but requires running ten additional reader models per response for a gain of 0.031 over the four-reader configuration. Larger reader sets therefore remain useful when more computation is available, while the preset four-reader budget captures most of the improvement over a single reader at substantially lower computational cost. Reader selection is therefore central to the performance–computation trade-off: it determines how much of the full pool's detection value can be retained when only a small subset of readers is used.

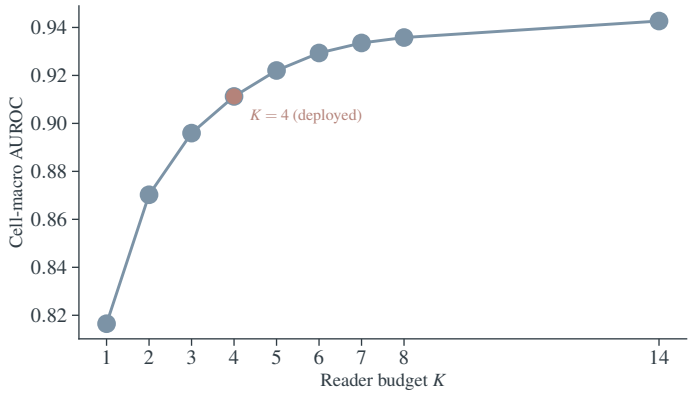


Figure 4: Detection performance across reader budgets. AUROC is the cell-macro average over the 42 target–benchmark pairs. The dashed line marks the preset budget $K = 4$ used in the primary experiments.

Table 9 lists all evaluated budgets. The largest increase occurs when moving from one reader to two, and the gains are smaller beyond four readers. The four-reader configuration retains a gain of 0.095 over the single-reader result, compared with 0.126 for the full pool. The curve continues to improve over the evaluated range, making the choice of a reader set relevant to how much computation can be devoted to each response.

Table 9: Complete reader-budget sweep. Each AUROC is the cell-macro average over the 42 target–benchmark pairs. The primary budget is $K = 4$.

Readers K	1	2	3	4	5	6	7	8	14
Mean AUROC	0.816	0.870	0.896	0.911	0.922	0.929	0.934	0.936	0.943

Layer selection. For each reader, the candidate relative depths are $\{0.25, 0.50, 0.75, 1.00\}$, and the selected candidate minimizes inner-cross-validation log loss. For this supplementary analysis, we select layers separately for each target, yielding 336 reader–target–outer-split decisions. In the primary experiments, layer selection instead uses responses pooled across targets within each outer split. Across the 56 primary selection decisions, relative depths 0.25, 0.50, 0.75, and 1.00 are selected in 8.9%, 50.0%, 41.1%, and 0.0% of cases, respectively.

Table 10 reports the target-wise supplementary analysis. The halfway layer is selected most often, accounting for 44.3% of its 336 decisions, followed by the three-quarter-depth and quarter-depth candidates at 32.1% and 18.2%. The final layer is selected in 18 of these decisions.

Figure 5 shows the overall frequencies and their distribution within reader families in this supplementary analysis. The family-level patterns differ substantially: the halfway layer dominates for Gemma, the quarter-depth layer dominates for Phi, and Mistral frequently selects the three-quarter-depth layer. These differences support choosing the depth for each reader from development data rather than imposing one relative depth on the entire candidate pool.

Table 10: Candidate-layer selection in the target-wise supplementary analysis. Selection counts and shares summarize the 336 target–outer-split–reader combinations. Validation log loss is averaged over the same combinations for each candidate depth.

Relative depth	Selections	Share (%)	Validation log loss
0.25	61	18.2	0.6518
0.50	149	44.3	0.6398
0.75	108	32.1	0.6435
1.00	18	5.4	0.6713



Figure 5: Layer-selection frequencies in the target-wise supplementary analysis. (a) Percentages over all 336 selection decisions. (b) Percentages computed within each reader family. Each family row sums to approximately 100%, subject to rounding.

Answer-span representation. We compare the selected-depth answer-span representation with fixed depths and alternative pooling rules while retaining the same readers and probe training procedure. Table 11 reports the mean AUROC of the 14 first-stage reader probes, and Figure 6 shows the corresponding variation across readers. Selecting the layer yields a mean AUROC of 0.681, compared with 0.648 and 0.672 at fixed relative depths 0.25 and 0.50.

At the selected depth, averaging the prediction states across the answer span performs better than averaging over the full sequence (0.649) or using only the state predicting the final answer token (0.627). These results favor answer-span mean pooling over the full-sequence and single-position alternatives for supervised hallucination detection. Together, the layer and pooling comparisons support the extraction choices used by the main method.

Table 11: Representation choices for the first-stage probes. AUROC is averaged over the 14 readers; the standard deviation describes variation across readers.

Representation	Mean AUROC	Reader s.d.
Answer-span mean, selected depth	0.681	0.040
Answer-span mean, fixed depth 0.25	0.648	0.045
Answer-span mean, fixed depth 0.50	0.672	0.039
Full-sequence mean, selected depth	0.649	0.039
Final answer-token prediction state, selected depth	0.627	0.036

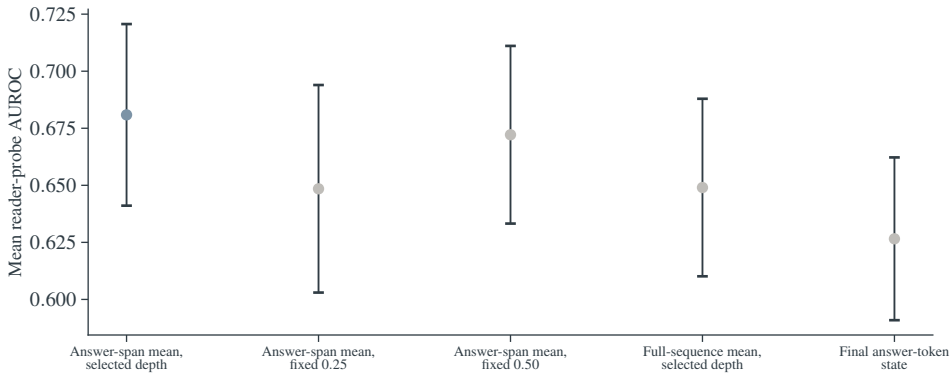


Figure 6: First-stage detection performance under different representation choices. Points show the mean AUROC over readers, and error bars show one standard deviation across readers.

E.2 Detection under Answer-Recovery Failure

We test whether useful detection evidence requires readers to recover the correct answer through generation, recognition, or verification. The main comparison evaluates detection on responses for which every candidate reader meets a common recovery-failure criterion. We report the recovery results for each reader, detection on the retained subset, and the reference signals evaluated on the same responses.

Recovery tests and subset construction. Each reader completes three tests using the benchmark question and any context supplied with it: free answer generation, four-choice recognition of the reference answer among three distractors from the same benchmark, and binary answer verification. Each test uses two fixed prompt forms with deterministic decoding. Generation is scored by normalized string match against the reference answer, with the labeling judge adjudicating responses that match no surface form, so a reader is not recorded as failing merely for answering in another format. Recognition and verification are scored by the selected option. A reader meets the failure criterion when both generation forms and both recognition forms are incorrect, and at most one verification form is correct. A response enters the final subset only when all 14 readers meet all three criteria. The recovery outcomes determine subset membership independently of the detector scores.

Table 12 gives the cumulative subset sizes. All readers fail free generation on 1,559 responses. Adding the recognition criterion retains 342, and adding the verification criterion retains 168. The final subset therefore contains responses that remain difficult for the readers across all three recovery tests. A simple calculation assuming independent outcomes across readers and prompt forms, using each reader's mean generation accuracy in Table 13 for both forms, predicts fewer than one response meeting the all-reader generation-failure criterion. The observed count of 1,559 shows that failures overlap substantially across readers and prompt forms. Figure 7 presents this progression together with the overall detection comparison.

Table 12: Cumulative construction of the all-reader recovery-failure subset. Each additional criterion is applied to the responses retained by the preceding stages.

Condition	Responses
Full primary dataset	7,560
Generation criterion	1,559
+ Recognition criterion	342
+ Verification criterion	168

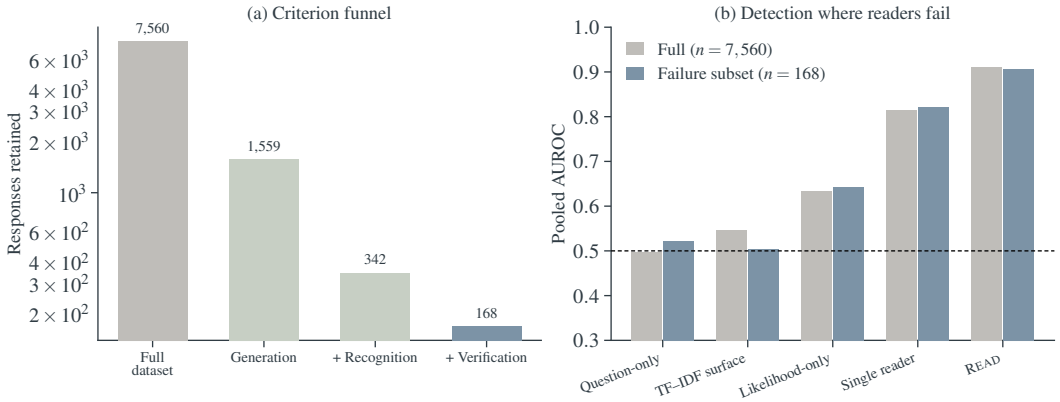


Figure 7: Answer recovery and detection. (a) Responses retained after successively applying the all-reader failure criteria; the count axis is logarithmic. (b) Pooled AUROC on the full dataset ($n = 7,560$) and the recovery-failure subset ($n = 168$), using the same held-out predictions.

Recovery performance across readers. Table 13 reports each reader's accuracy in the three tests, and Figure 8 shows how their recovery profiles differ. Generation accuracy ranges from 0.197 to

0.518, recognition accuracy from 0.292 to 0.620, and verification accuracy from 0.516 to 0.724. Recognition and verification provide additional ways for a reader to demonstrate successful recovery, which motivates requiring failure across all three tests when constructing the subset. The final subset uses the complete reader pool, including readers with the highest observed recovery accuracies.

Table 13: Accuracy of each reader in the recovery battery. Recognition uses four answer choices, and verification uses a binary decision. Values summarize each reader's results over the recovery evaluation.

Reader	Generation	Recognition	Verification
Qwen3-0.6B	0.197	0.331	0.516
Qwen3-1.7B	0.286	0.300	0.585
Qwen3-4B	0.392	0.504	0.641
Qwen3-8B	0.480	0.474	0.717
Qwen3.5-4B	0.493	0.491	0.684
Qwen3.5-9B	0.480	0.531	0.724
Llama-3.2-1B	0.329	0.292	0.585
Llama-3.2-3B	0.292	0.431	0.519
Llama-3.1-8B	0.518	0.620	0.690
gemma-3-4b-it	0.350	0.441	0.638
gemma-3-12b-it	0.500	0.574	0.708
Ministral-3-3B	0.339	0.454	0.574
Ministral-3-8B	0.408	0.522	0.648
Phi-4-mini-instruct	0.326	0.448	0.641

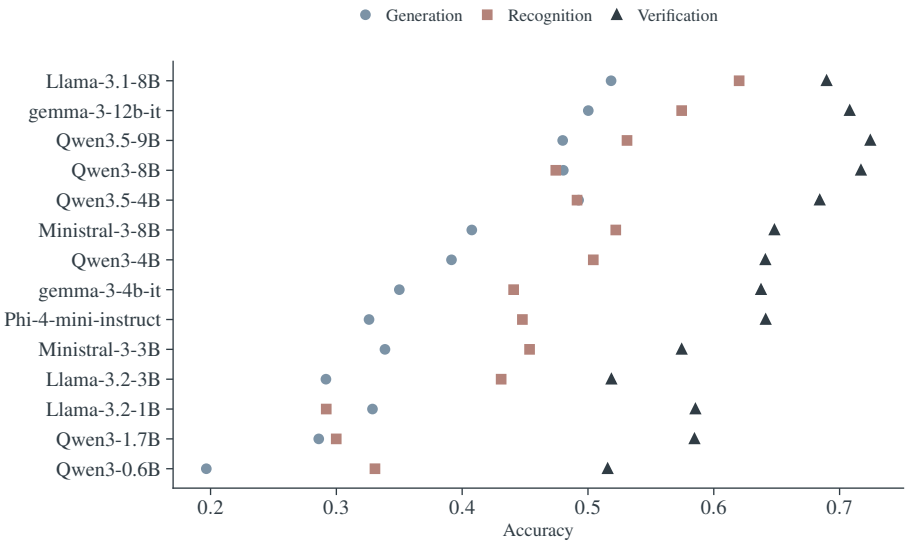


Figure 8: Reader-level recovery profiles. Readers are ordered by their mean accuracy across the three recovery tests. Colors and marker shapes distinguish generation, recognition, and verification.

Detection and controls on the failure subset. Table 14 evaluates the same held-out detector predictions on the full dataset and on the retained responses. The subset holds too few responses per target–benchmark pair for a per-pair AUROC, so this comparison is pooled. READ achieves 0.907 pooled AUROC on the failure subset, compared with 0.910 on the full dataset. The single-reader control remains informative at 0.822, while combining readers provides a higher AUROC on the same subset. Question-only, surface-text, and likelihood-only controls reach 0.522, 0.503, and 0.643, respectively.

The question-only control captures information available from the question without the supplied answer, while the surface-text and likelihood controls assess simpler signals from the response.

Their lower performance on the retained subset supports the value of the learned reader evidence for judging the answer itself. The main result is that detection remains effective when all readers fail the specified recovery tests, separating successful answer recovery from the ability to provide useful detection evidence.

Table 14: Detection on the full dataset and the all-reader recovery-failure subset. Entries are pooled AUROCs. Δ is the subset AUROC minus the full-dataset AUROC for the same method.

Method	Full dataset $n = 7{,}560$	Failure subset $n = 168$	Δ
Question-only	0.497	0.522	+0.025
TF–IDF surface	0.546	0.503	−0.042
Likelihood-only	0.634	0.643	+0.010
Single-reader control	0.815	0.822	+0.006
READ	0.910	0.907	−0.003

Target-level AUROCs range from 0.667 to 0.954, with only 14–50 responses per target. These small groups limit the precision of target-level comparisons. We therefore focus on pooled AUROC to assess detection on the recovery-failure subset as a whole.

E.3 Generalization across Targets and Benchmarks

We test whether the advantage of conditional-gain reader selection transfers to target families and benchmarks excluded from detector training. These two settings examine transfer across the model generating the answers and across the task distribution on which the detector is evaluated. We report both the overall comparison among reader-selection rules and the results for each held-out benchmark.

Transfer settings. For target-family holdout, detector training excludes responses from the evaluated target family and uses responses from the remaining target families. For benchmark holdout, we train on six benchmarks and evaluate on the remaining one, rotating the held-out benchmark over all seven sources. Layer selection, reader selection, and detector fitting are performed using the training portion of each setting. Evaluation retains the same 42 target–benchmark pairs, and AUROC is averaged equally over those pairs.

Reader-selection rules under transfer. Table 15 compares conditional-gain selection with individual-signal selection, random and fixed reader sets, a development-selected single reader, and the full reader pool. Figure 9 presents the AUROC comparison for the in-domain and two transfer settings. At a budget of four readers, conditional gain has the highest AUROC among the reported selection rules in all three settings. It reaches 0.875 under target-family holdout and 0.862 under benchmark holdout, compared with 0.865 and 0.857 for individual-signal selection. Its log loss is also lower under target-family holdout (0.461 versus 0.467). Under benchmark holdout, however, conditional-gain selection has higher log loss (0.486 versus 0.477), despite its higher AUROC. Its advantage in ranking therefore does not extend to log loss in this setting.

Here the single-reader reference is selected using the pooled training population of each setting, whereas the benchmark-selected reference in Table 2 selects separately by benchmark. The development-selected single reader reaches 0.777 and 0.751 in the two transfer settings, showing the benefit of combining evidence from multiple readers after the training distribution changes. Using all 14 readers produces higher AUROC at a larger budget, reaching 0.925 and 0.918. The four-reader results show that conditional-gain selection continues to make useful choices among the available readers under both forms of transfer.

Results for each held-out benchmark. Table 16 reports the complete benchmark-level comparison, including reader-set controls, the training-free LLM-Check baseline, and supervised text classifiers. Figure 10 visualizes the same results to show how the relative performance varies across the held-out sources. Conditional-gain selection outperforms individual-signal selection on three of the seven held-out benchmarks, and trails it on the others. Its AUROC ranges from 0.799 to 0.866 on the four QA benchmarks and from 0.852 to 0.917 on mathematics, code, and tool use. The full reader

Table 15: Reader-selection rules under in-domain evaluation, target-family holdout, and benchmark holdout. AUROC is the cell-macro average over the 42 evaluation pairs in each setting. Each setting uses its corresponding training population.

Construction rule	In-domain AUROC	In-domain Log loss	Target-family AUROC	Target-family Log loss	Benchmark AUROC	Benchmark Log loss
READ ($K = 4$)	0.911	0.384	0.875	0.461	0.862	0.486
Individual signal ($K = 4$)	0.896	0.409	0.865	0.467	0.857	0.477
Random set ($K = 4$)	0.887	0.427	0.861	0.477	0.852	0.487
Fixed set ($K = 4$)	0.870	0.452	0.862	0.469	0.858	0.471
Selected reader ($K = 1$)	0.816	0.523	0.777	0.573	0.751	0.602
Full pool ($K = 14$)	0.943	0.314	0.925	0.368	0.918	0.388



Figure 9: Detection under changes in the training distribution. The four-reader configurations use the same additive fusion model. The single-reader and full-pool configurations provide references at smaller and larger reader budgets.

pool retains the highest AUROC on every held-out benchmark, while the four-reader configuration preserves much of its detection performance.

Under benchmark holdout, mean AUROC is 0.895 on GSM8K, MBPP, and xLAM, compared with 0.837 on the four QA benchmarks. The corresponding in-domain means are 0.918 and 0.906, giving decreases of 0.023 and 0.069, respectively. Performance therefore declines more on the QA benchmarks, despite the presence of three other QA datasets in each holdout's training data.

Comparison with text encoders. The four-encoder stack reaches 0.775 cell-macro AUROC under benchmark holdout, and ELECTRA-large reaches 0.744, compared with 0.862 for READ. Their decreases from in-domain evaluation are 0.051 and 0.051, respectively, while the decrease for READ is 0.049. The gap over ELECTRA-large consequently increases from 0.116 in domain to 0.118 under benchmark holdout. The training-free LLM-Check baseline provides a further reference on the same responses, reaching 0.760. These comparisons support the value of supervised reader evidence when the evaluation benchmark is absent from detector training.

Table 16: Complete comparison on held-out benchmarks. Each benchmark column averages AUROC over the six targets. Avg. is the cell-macro AUROC over all 42 pairs, and Δ is its change from the same method's in-domain result.

Method	TQA	Halu	Triv	CoQA	GSM	MBPP	xLAM	Avg.	Δ
Full reader pool ($K = 14$)	0.927	0.906	0.922	0.909	0.937	0.882	0.942	0.918	−0.025
READ ($K = 4$)	0.866	0.799	0.859	0.822	0.917	0.852	0.917	0.862	−0.049
Individual signal ($K = 4$)	0.907	0.819	0.860	0.824	0.915	0.801	0.873	0.857	−0.039
Random set ($K = 4$)	0.863	0.833	0.838	0.843	0.881	0.831	0.875	0.852	−0.035
Fixed set ($K = 4$)	0.869	0.854	0.865	0.865	0.880	0.802	0.873	0.858	−0.012
Selected reader ($K = 1$)	0.709	0.749	0.769	0.697	0.810	0.714	0.811	0.751	−0.065
LLM-Check (reader-side)	0.740	0.795	0.715	0.738	0.740	0.800	0.794	0.760	+0.000
Four-encoder stack	0.747	0.765	0.778	0.770	0.844	0.770	0.748	0.775	−0.051
ELECTRA-large	0.769	0.732	0.747	0.783	0.749	0.762	0.662	0.744	−0.051
ModernBERT-large	0.763	0.649	0.695	0.698	0.857	0.740	0.760	0.737	−0.048
DeBERTa-v3-large	0.645	0.798	0.770	0.643	0.831	0.699	0.672	0.723	−0.061
RoBERTa-large	0.718	0.631	0.663	0.626	0.736	0.670	0.606	0.664	−0.064
TF–IDF logistic	0.532	0.535	0.613	0.551	0.562	0.645	0.544	0.569	−0.126

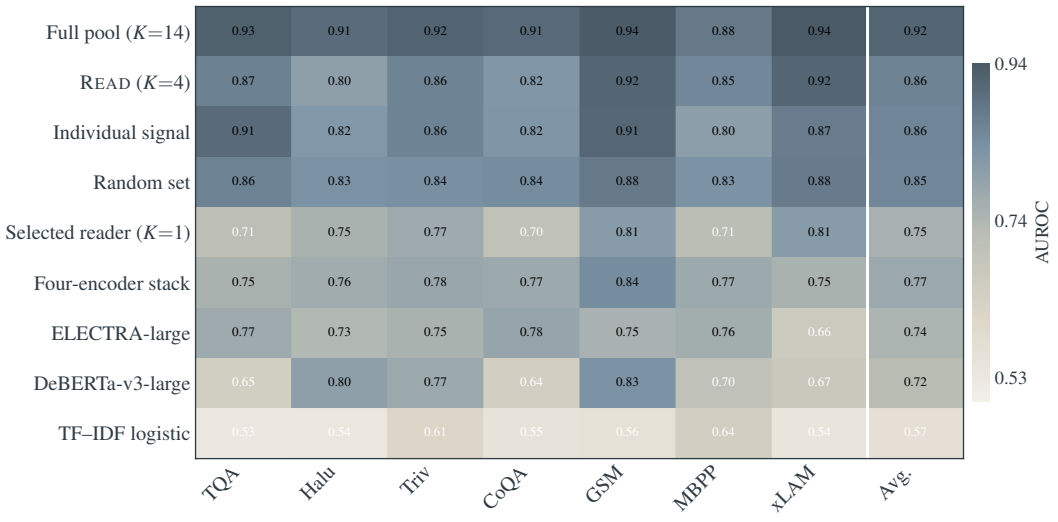


Figure 10: AUROC across held-out benchmarks. Each cell averages the six target-specific evaluation results for the indicated method and held-out benchmark. The final column averages over all benchmark columns. Bold numbers identify the READ row.

E.4 Robustness to Reader-Family Removal

We test whether useful detection can be retained when an entire reader family is unavailable in the candidate pool. For each condition, we remove that family before reader selection and detector training, then select four readers from the remaining candidates using the same conditional-gain rule. The reference selects four readers from the complete pool of 14. All conditions use the same four source-question-grouped outer splits.

Table 17 reports the remaining pool size, cell-macro AUROC, its variation across target–benchmark pairs, and cell-macro log loss. Figure 11 shows the AUROC comparison against the four-reader reference. The reference reaches 0.911 mean AUROC, and removing a family produces values between 0.889 and 0.911. Removing Gemma, Llama, or Mistral yields 0.909, 0.897, and 0.910, respectively.

The largest decrease occurs when the six Qwen readers are removed, leaving eight candidates from which to select the four-reader set. This condition reaches 0.889 AUROC and has the highest log loss among the removal conditions. Removing Phi costs 0.000: it contributes a single reader, selected in one of the four outer splits of the reference runs. The results show that the available pool affects the strength of the final detector, while the remaining families continue to supply useful evidence after any one family is removed.

Table 17: Reader-family removal with $K = 4$ in every condition. AUROC and log loss are cell-macro averages over the 42 target–benchmark pairs; the standard deviation describes variation in AUROC across those pairs.

Removed family	Removed	Remaining	Mean AUROC	AUROC s.d.	Log loss
None	0	14	0.911	0.031	0.384
Gemma	2	12	0.909	0.032	0.387
Llama	3	11	0.897	0.032	0.410
Mistral	2	12	0.910	0.033	0.385
Phi	1	13	0.911	0.032	0.385
Qwen	6	8	0.889	0.039	0.427

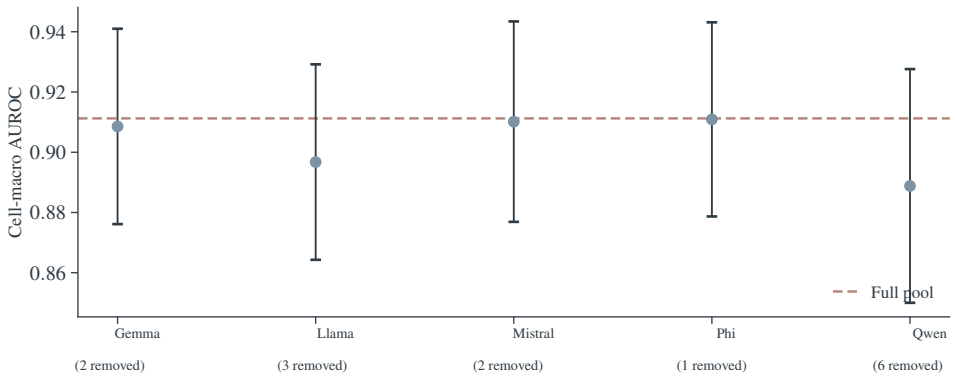


Figure 11: Detection after removing each reader family. Points show cell-macro AUROC and error bars show one standard deviation across the 42 target–benchmark pairs. The dashed line marks the mean for four readers selected from the complete pool.

E.5 Sensitivity to Factual Perturbations

We test whether the detector responds to factual changes in an answer while remaining stable under meaning-preserving rephrasing. This comparison examines whether the evidence extracted by the readers tracks changes in answer content, alongside the detection results on the original responses.

Paired edits. We construct original–edited answer pairs using entity substitution, numeral substitution, and negation insertion as factual corruptions. Paraphrases provide a meaning-preserving control. Edits are applied to answers for which the corresponding change is well defined, and each pair is checked for the intended factual corruption or preservation of meaning before evaluation. The fitted detector scores both versions of the answer with the question held fixed. The score change is the edited-answer score minus the original-answer score, so a positive value indicates increased hallucination risk.

Table 18 reports the number of retained pairs, mean score change, and fraction with a positive change. The retained sets contain 1,568 entity substitutions, 1,825 numeral substitutions, 1,712 negation insertions, and 6,842 paraphrases. Each edit family is evaluated on its corresponding retained pairs.

Score response to factual and surface changes. Entity substitutions, numeral substitutions, and negation increase the hallucination score by 0.221, 0.174, and 0.131 on average. The score increases on 87.7%, 83.5%, and 71.5% of the respective pairs. These three types of factual corruption therefore produce a similar overall response, although the amount of change varies across examples. Paraphrases produce a small mean score increase of 0.005, with positive changes in 51.9% of pairs. The average change is much smaller than those caused by factual edits.

Figure 12 shows the individual score changes as well as the means. The factual edits yield broad distributions shifted toward higher hallucination scores, while the paraphrase changes are concentrated near zero. Together, the paired statistics and distributions support sensitivity to factual content and relative stability under changes in wording.

Table 18: Paired score changes after editing an answer. The positive-shift rate is the fraction of retained pairs for which the hallucination score increases. All score statistics use the retained pairs.

Edit	Pairs	Mean score change	Positive shifts (%)
Entity substitution	1,568	0.221	87.7
Numeral substitution	1,825	0.174	83.5
Negation insertion	1,712	0.131	71.5
Paraphrase	6,842	0.005	51.9

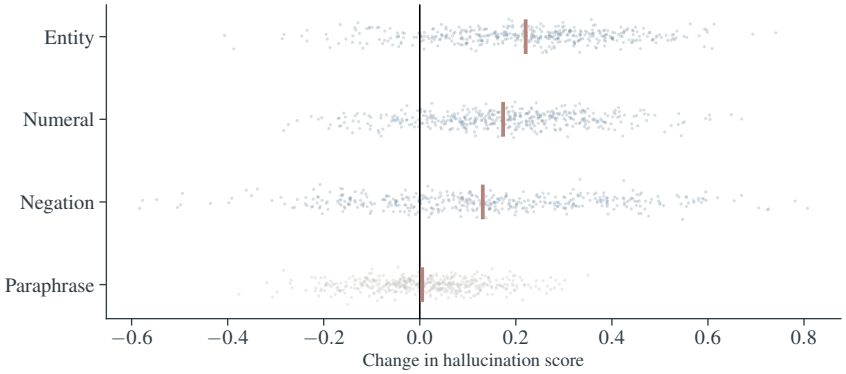


Figure 12: Changes in hallucination score after editing an answer. Points represent validated original–edited pairs, and horizontal bars indicate the mean change for each edit type. Positive changes indicate increased predicted hallucination risk.