HD-FUSE: SINGLE-PASS HALLUCINATION DETECTION VIA CONFIDENCE RESHAPING WITH CONFORMAL LIKELIHOOD-RATIO FUSION UNDER LOGPROB-ONLY APIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination detection under API logprob-only contracts is practically important but constrained: production LLM APIs expose only final-token logprob distributions, while hidden states, attention weights, and repeated sampling are unavailable or expensive. We present a systematic empirical study of single-pass, logprob-only hallucination detection for short-form QA (≤ 16 tokens) and identify a validated recipe that outperforms individual signals: extract a 5-dimensional confidence vector, model the test as a Gaussian likelihood ratio (LDA/QDA), fuse complementary likelihood-ratio members (including thermodynamic trajectory statistics of the top-K distribution) via a calibrated cross-validated rank ensemble, and attach a split-conformal p-value with a finite-sample false-positive-rate guarantee, at zero extra passes. On real public benchmarks the recipe attains the best cross-cell mean AUC at one pass versus five; we confirm the signal and false-positive control on a production API (DeepSeek), showing that the logprob-only setting is not limited to local quantised models. The ablation indicates that the ensemble gain arises from complementary feature spaces; no single statistic provides comparable cross-cell performance. Scope and prerequisites (short answers; a per-(model, benchmark) calibration fold) are stated up front, so the contribution is a validated deployment recipe plus an interpretable diagnostic framework; longer-form per-sentence localisation is future work.

1 INTRODUCTION

Large language models (LLMs) (Vaswani et al., 2017) frequently produce fluent but factually incorrect statements, the *hallucination* problem (Ji et al., 2023; Huang et al., 2025). The problem is pervasive and well documented on public QA benchmarks such as TruthfulQA (Lin et al., 2022) and HaluEval (Li et al., 2023a), and it resists mitigation: decoding-time interventions (Chuang et al., 2024b) reduce the symptom but do not eliminate it. In deployment this is not an academic concern: a system must decide once, per generated answer, whether to trust and surface it or flag it for review, and the API it talks to will not hand over its hidden states. This paper studies how to build a principled, single-pass hallucination detector under an API contract that exposes only output logprobs.

Existing detectors fall into two families, both of which are constrained under this contract. *Internal-state methods* (Li et al., 2023b; Burns et al., 2023; Zou et al., 2023; Chuang et al., 2024a; Kossen et al., 2024) exploit per-layer hidden states or attention maps and are accurate, but production LLM APIs expose neither: they return only the final-layer logprob distribution of the sampled token (OpenAI `logprobs`, Anthropic `top_logprobs`, Gemini `logprobs`). *Consistency methods* (Kuhn et al., 2023; Manakul et al., 2023; Farquhar et al., 2024) sample K stochastic responses per query and measure self-agreement; they are API-agnostic but multiply latency and cost by K and fail silently when a model is confidently wrong. Logprob-only detectors on the single greedy pass exist (Moslonka et al., 2026; Shapiro et al., 2026; Gabriel, 2026), but most rest on a single heuristic statistic, and no single logprob-derived statistic is reliable across models and benchmarks.

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.

We present *HD-Fuse* (Figure 1), a new single-pass hallucination detection method for short-form answers (≤ 16 tokens) under the logprob-only contract. We formalise detection as a binary hypothesis test: the 5-dimensional signal vector is modelled as Gaussian per class, and the Neyman–Pearson lemma (Bishop, 2006) yields a closed-form log-likelihood-ratio detector whose linear and quadratic forms coincide with LDA and QDA (§3.2). Because no single statistic is reliable, HD-Fuse fuses a *family* of likelihood-ratio detectors built from complementary feature spaces—thermodynamic trajectory statistics of the top-K distribution (spectral temperature, entropy production, drift/diffusivity, Fisher–Rao path length) and classical Gaussian/non-parametric LRT variants—via a CV-weighted rank ensemble, and a split-conformal layer (Vovk et al., 2005; Angelopoulos & Bates, 2023) converts the ensemble score into per-query p-values with a distribution-free false-positive-rate guarantee, at zero extra forward passes.

Contributions. We make three contributions.

- *A new detection method.* We propose HD-Fuse, a single-pass detector that works with nothing but the output logprobs of one greedy pass: a closed-form Gaussian likelihood-ratio test fused with thermodynamic trajectory statistics by a cross-validated rank ensemble, and a split-conformal layer giving a finite-sample false-positive-rate guarantee.
- *A new feature family.* We introduce thermodynamic trajectory statistics of the top-K distribution as confidence features, and show by ablation that the ensemble gain arises from complementary feature spaces rather than from a single statistic.
- *Systematic evaluation.* On HaluEval QA and TruthfulQA v1 MC1 across four open-weight LLMs (8 cells), HD-Fuse attains the best mean AUC (0.835), achieves the best AUC on 5 of 8 cells, and improves over the strongest external baselines ($p \leq 0.023$) at one forward pass versus five; we further validate the recipe on a production API (DeepSeek), showing that the logprob-only setting is not limited to local quantised models, and report a conformal FPR audit, per-cell transferability, and a 12-cell template companion study that isolates construction artefacts.

The rest of this paper is organised as follows: Section 2 reviews related work; Section 3 presents the method; Section 4 reports the experiments; Section 5 discusses limitations; Section 6 concludes.

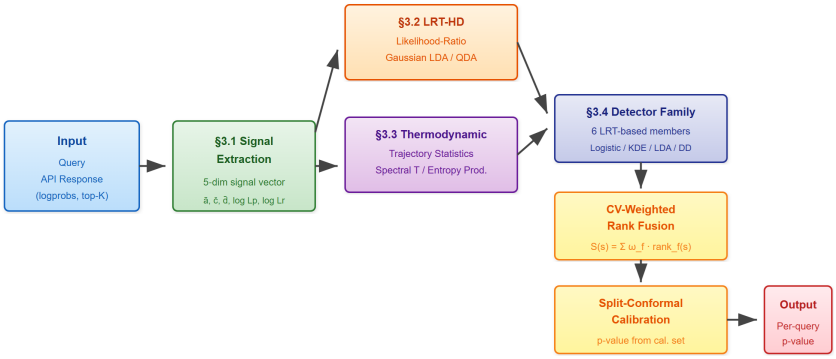

Figure 1: The HD-Fuse pipeline under the logprob-only API contract. From a single greedy pass the 5-signal vector is extracted and passed to two complementary likelihood-ratio feature spaces (Gaussian LRT-HD and thermodynamic trajectory statistics); the six-member detector family is combined by a CV-weighted rank average and converted into a per-query p-value by the split-conformal layer. No hidden states are required.

2 Related Work

Research relevant to this paper falls into three strands, each shaped by what a detector is allowed to observe: **Internal-state detectors.** ITI (Li et al., 2023b), CCS (Burns et al., 2023), representation engineering (Zou et al., 2023), attention-based lookback lenses (Chuang et al., 2024a), and semantic-entropy probes (Kossen et al., 2024) are effective but rely on per-layer hidden-state access

that production APIs do not expose; HD-Fuse is designed to complement internal-state detectors under the API contract. **Output-statistics detectors.** Token-level output statistics (Fomicheva et al., 2020) and layer-contrasting decoding (Chuang et al., 2024b) exploit output distributions, and the entropy-production-rate (EPR) family (Moslonka et al., 2026), which operates on the same black-box output-logprob contract, is the most closely related existing work; HD-Fuse differs in three respects: uses a closed-form likelihood ratio, extends the feature space to thermodynamic trajectory statistics, and combines members with a cross-validated ensemble with a conformal guarantee. 2025–2026 work explores logprob time series (Shapiro et al., 2026), first-token confidence (Gabriel, 2026), and multi-perspective/metamorphic analyses (Kong et al., 2025; Yang et al., 2025); HD-Fuse is compared against the strongest of these in §4. **Consistency-based and calibrated detection.** Semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) and SelfCheckGPT (Manakul et al., 2023) require K sampled responses ($K\times$ compute) and degrade when the model is wrong in the same way across samples, whereas distribution-free conformal calibration (Angelopoulos & Bates, 2023) motivates the level-α layer we attach to a single-pass detector.

3 METHOD

3.1 PROBLEM SETTING AND SIGNALS

For each query Q, the API returns the completion $Y = (y_1, \ldots, y_T)$ with the sampled token's logprob and a top-K logprob vector $p_t = (p_{t,1}, \ldots, p_{t,K})$ per position. We use `temperature=0` (greedy decoding), so extraction is deterministic and reproducible. Let n be the number of leading positions over which statistics are averaged and $H(p_t^{(K)}) = -\sum_k p_{t,k} \log p_{t,k}$ the top-K entropy. We extract the 5-dimensional signal vector

$$\mathbf{s} = \Big(\underbrace{\tilde{a}}_{\text{confidence}}, \underbrace{\tilde{c}}_{\text{spread}}, \underbrace{\tilde{d}}_{\text{gap}}, \underbrace{\log L_p}_{\text{context}}, \underbrace{\log L_r}_{\text{response}} \Big)^\top,$$

$$\tilde{a} = \frac{1}{n} \sum_{t \le n} -\log p(y_t | y_{<t}, Q), \quad \tilde{c} = \frac{1}{n} \sum_{t \le n} H(p_t^{(K)}), \quad \tilde{d} = \frac{1}{n} \sum_{t \le n} (\log p_{t,1} - \log p_{t,2}), \tag{1}$$

where L_p, L_r are the context and response lengths.

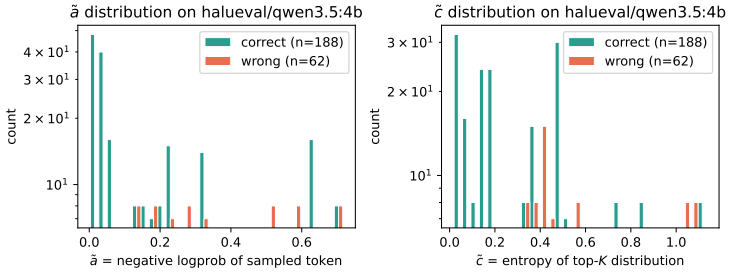

Figure 2: Empirical signal distributions on the HaluEval / qwen3.5:4b held-out fold (n=250). Left: $\tilde{a}$ (negative logprob of the sampled token); right: $\tilde{c}$ (top-K entropy). Correct (green) vs wrong (red): the densities separate on the upper tail, the structure a single-threshold detector cannot exploit.

Figure 2 shows these distributions on a held-out HaluEval fold: $\tilde{a}$ and $\tilde{c}$ separate on the upper tail, the high-uncertainty region a single threshold on a single signal cannot exploit, and the response-level signals capture cheap global correlates of difficulty. We use n=3 and K=5; both are stable defaults (§4.5, Supplement S5).

3.2 THE LRT FRAMEWORK

We model the signal vector of truthful responses as $f_0(\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and hallucinated responses as $f_1(\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, estimated on a 250-prompt calibration fold per (model, benchmark) cell. The log-likelihood ratio is

$$\ell(\mathbf{s}) = \mathbf{s}^\top \mathbf{Q} \mathbf{s} + \mathbf{w}^\top \mathbf{s} + b, \qquad \mathbf{Q} = \tfrac{1}{2}(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}), \quad \mathbf{w} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0, \tag{2}$$

so when $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$ it collapses to the Fisher LDA discriminant $\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ and otherwise to QDA. We adopt LDA as the headline detector: its 25 parameters give more stable estimates at the 250-sample scale than the 40 of QDA. Empirically, LRT-HD/LDA attains 0.900 mean AUC on the 12 template cells against 0.854 for QDA; QDA's extra covariance parameters do not help on the imbalanced cells (Supplement S5). An operating point θ is selected on the calibration fold by F1 maximisation. Because F1 can be unstable when positives are scarce, the threshold-independent AUC is reported as the primary metric and the conformal layer sets the operating point; we flag hallucination when $\ell(\mathbf{s}) > \theta$.

This likelihood ratio follows from the Neyman–Pearson lemma: for testing f_0 against f_1, thresholding $\ell(\mathbf{s}) > \theta$ is the most powerful level-α test, and with known priors and costs it is the Bayes rule. It also yields a closed-form AUC that serves as a diagnostic of the effect size: writing the class-separation in the decorrelated metric as $\Delta^2 = \Delta\boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1} \Delta\boldsymbol{\mu}$,

$$\mathrm{AUC} = \Phi\Big(\sqrt{\tfrac{1}{2}\,\Delta^2}\Big), \tag{3}$$

a monotone function of the effect size. Two consequences guide the design. The inverse covariance decorrelates the signals, so a marginally anti-predictive statistic such as $\tilde{d}$ can receive a weight of either sign; and as $n \to \infty$ the plug-in detector converges to the Bayes-optimal classifier at rate $O(d/\sqrt{n})$ (d=5). We examine the finite-sample behaviour of this plug-in estimator empirically in §4.1; derivations and the conditions under which these idealised guarantees hold after estimation are in Supplement S1.

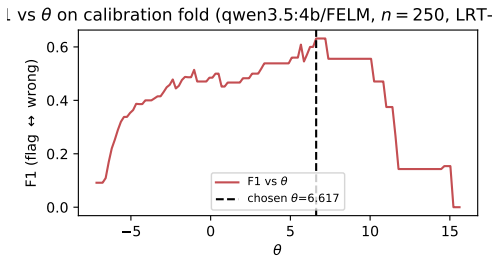

Figure 3: F1 vs. threshold θ on the calibration fold (qwen3.5:4b/FELM, n=250, LRT-HD/LDA). The curve is quantised (this cell has only n_{wrong}=12 hallucinated samples); the threshold-independent AUC is therefore the headline metric, and the conformal layer, not θ, sets the operating point.

3.3 Thermodynamic trajectory statistics

The signal vector summarises marginals but discards the *path* the generation takes through the probability simplex. Treating the top-K distribution as a Boltzmann law $p_{t,k} \propto e^{-\lambda_t k}$, we define four trajectory statistics over the first n positions: the spectral temperature $T_{\text{spec}} = \big(\frac{1}{n}\sum_t \lambda_t\big)^{-1}$; the entropy production $\mathrm{EP} = \frac{1}{n}\sum_t \big[H(p_t^{(K)}) - H(p_{t-1}^{(K)})\big]$; a Girsanov-style drift/diffusivity split of the increments Δp_t into drift $\hat{\Delta}_t$ and residual diffusion $\|\Delta p_t - \hat{\Delta}_t\|$; and the Fisher–Rao path length $\mathcal{L}_{\mathrm{FR}} = \sum_t \arccos\big(\langle\sqrt{p_{t-1}}, \sqrt{p_t}\rangle\big)$, the geodesic length under the information metric. Hallucinated generations show hotter spectra, larger entropy production, and longer Fisher–Rao paths; these enter a KDE-based and an LDA-based likelihood ratio (ThermoPath-LRT).

3.4 The HD-Fuse ensemble and conformal calibration

Because individual feature spaces vary in reliability across models and benchmarks, HD-Fuse combines them into the family $\mathcal{F} = \{\text{logistic fusion, ThermoPath-KDE, ThermoPath-LDA, KDE-LRT, DD-LRT, LRT-HD(LDA)}\}$, each fitted on the calibration fold only. Let $r_f(\mathbf{s})$ be the within-fold rank of score $f \in \mathcal{F}$ and let A_f be f's 5-fold cross-validated AUC on the calibration fold. The ensemble score is the weighted rank

average

$$S(\mathbf{s}) = \sum_{f \in \mathcal{F}} \omega_f \, r_f(\mathbf{s}), \qquad \omega_f = \frac{\max(0,\, A_f - \frac{1}{2})}{\sum_{g \in \mathcal{F}} \max(0,\, A_g - \frac{1}{2})}, \tag{4}$$

with a guard that excludes members whose AUC is undefined or whose score is non-finite (relevant on cells with few positives). Members are weighted by *validated* rather than training performance, so a member that overfits the calibration fold is down-weighted exactly where it would otherwise dominate. Finally, a split-conformal layer (Vovk et al., 2005) turns S into a per-query p-value

$$p(\mathbf{s}) = \frac{1 + \#\{i :\, S(\mathbf{s}_i) \geq S(\mathbf{s})\}}{n_0 + 1}, \tag{5}$$

computed over the n_0 truthful calibration prompts, flagging when $p(\mathbf{s}) \leq \alpha$. Under exchangeability alone this controls the marginal FPR at level α: no Gaussian assumption, no known parameters, zero extra forward passes. Inference remains one greedy forward pass; the ensemble adds $O(|\mathcal{F}|)$ rank lookups.

4 Experiments and results

Gaussian assumption and its scope. While Mardia's test (Mardia, 1970) indicates non-normality on the Ministral cells (p=0.01–0.04), the Gaussian LDA stays competitive with its non-parametric counterparts there (Table 1: 0.785 vs 0.756, 0.732 vs 0.759), and the plug-in oracle of Eq. 3 tracks the empirical LDA AUC across the 8 real cells (mean absolute gap 0.031, max 0.087). The linear discriminant is robust to the observed departures from ellipticity; the non-parametric members (KDE-LRT, ThermoPath-LRT) serve as fallback for more complex shapes. The oracle also predicts the cells where fusion pays off: on real HaluEval (17%–30% accuracy) the gain is largest (+0.02–0.10), and the fitted weights give $\tilde{d}$ both signs across cells (3 positive, 5 negative), corroborating the decorrelation claim. The largest absolute gap (oracle minus empirical; −0.087, TruthfulQA/qwen3.5:4b) marks the cell where the plug-in estimate departs most from the measured AUC. Robustness bound, EB-shrinkage, and the full Mardia table: Supplement S1.

Models and serving. Four open-weight LLMs ($\leq$ 9B): qwen3.5:4b, qwen3.5:9b (Qwen family), ministral-3:3b, llama3.2:3b (different families, to test cross-family behaviour). All are served locally via Ollama with default Q4_K_M weights, `temperature=0`, `num_ctx=1024`, `num_predict=16`; total extraction cost is $\sim$ 8 GPU-hours on a 12 GB GPU.

Benchmarks. *Real public (primary)*: a 500-question sample of the official HaluEval QA test set ("Answer in 1–5 words", loose-subset scoring) (Li et al., 2023a) and 478 of the 817 questions of the official TruthfulQA v1 MC1 validation set (exact-letter scoring) (Lin et al., 2022). *Template-generated (companion)*: three 500-item sets in the style of TruthfulQA-MC1, HaluEval, and FELM (Chen et al., 2023), reproducible, and subject to a construction artefact we isolate explicitly.

Protocol and metrics. Per (model, benchmark) cell we hold out 250 calibration prompts (with gold labels) and the remainder for evaluation (250 on HaluEval, 228 on TruthfulQA); the headline metric is the held-out AUC. Cross-cell significance uses the Wilcoxon signed-rank and sign tests over paired per-cell AUCs, supplemented by the DeLong test (DeLong et al., 1988). Supervised baselines see the same calibration fold; sampling baselines receive the same K=5 samples at T=0.7.

Baselines. *One pass:* the strongest single signal $\tilde{a}$; KDE-LRT; LRT-HD (LDA); EPR and its learned variant WEPR (Moslonka et al., 2026); HALT (Shapiro et al., 2026); first-token confidence (phi_first) (Gabriel, 2026). *Five passes:* SelfCheckGPT (official package (Manakul et al., 2023); SelfCheck-Jaccard is a surrogate validated to 0.02 AUC on three cells) and Semantic Entropy (Kuhn et al., 2023). We also track HD-Fuse's family members DD-LRT and ThermoPath-LRT separately, as they are ensemble components, not competitors.

4.1 Real public benchmarks

Table 1 reports the primary evaluation. **First**, HD-Fuse achieves the highest mean AUC (0.835) and is the best method on 5 of the 8 cells, staying within 0.006 of the best result on the remaining three. **Second**, the cross-cell Wilcoxon test confirms the advantage is systematic: $8/8$ wins against eight

Table 1: Hallucination-detection AUC on the *real public* benchmarks (official HaluEval QA and TruthfulQA v1 MC1; 4 models × 2 benchmarks; held-out fold per cell). q4b/q9b/m3/l3 = qwen3.5:4b/9b, ministral-3:3b, llama3.2. Bold = best per column among the one-pass detectors (5-pass sampling baselines excluded). Bottom: cross-cell Wilcoxon/sign tests vs HD-Fuse (8 cells).

		HaluEval				TruthfulQA v1				
Method	**Type**	q4b	q9b	m3	l3	q4b	q9b	m3	l3	**Mean**
HD-Fuse (ours)	1 pass	**0.845**	**0.847**	**0.838**	**0.785**	0.833	**0.927**	**0.765**	0.838	**0.835**
DD-LRT (same family)	1 pass	0.791	0.824	0.808	**0.785**	0.827	0.922	0.749	**0.844**	0.819
single signal ($\tilde{a}$)	1 pass	0.828	0.829	0.782	0.686	**0.837**	0.925	0.629	0.778	0.787
KDE-LRT	1 pass	0.831	0.819	0.756	0.717	0.809	0.921	0.759	0.816	0.803
LRT-HD (LDA)	1 pass	0.824	0.809	0.785	0.734	0.825	0.926	0.732	0.819	0.807
EPR (Moslonka et al., 2026)	1 pass	0.784	0.779	0.659	0.667	0.820	0.925	0.667	0.730	0.754
WEPR (Moslonka et al., 2026)	1 pass	0.801	0.832	0.806	0.685	0.827	0.925	0.752	0.799	0.803
phi_first (Gabriel, 2026)	1 pass	0.781	0.791	0.547	0.632	0.820	0.925	0.683	0.765	0.743
HALT (Shapiro et al., 2026)	1 pass	0.824	0.802	0.824	0.759	0.814	0.926	0.752	0.526	0.778
SelfCheck-Jaccard	5 pass	0.807	0.796	0.734	0.720	0.749	0.831	0.646	0.626	0.739
SelfCheckGPT (official)	5 pass	0.671	0.716	0.703	0.799	0.748	0.826	0.589	0.629	0.710
Semantic Entropy	5 pass	0.785	0.769	0.728	0.690	0.748	0.832	0.647	0.706	0.738

Cross-cell tests vs HD-Fuse (Wilcoxon Δ, wins, p; 8 cells): single $\Delta{=}+0.048$, $7/8$, $p{=}0.023$; DD-LRT $\Delta{=}+0.016$, $6/8$, $p{=}0.055$; KDE-LRT/LRT-HD/WEPR/EPR/HALT/phi_first $8/8$, $p{\leq}0.008$; SelfCheck-Jaccard $\Delta{=}+0.096$, $8/8$, $p{=}0.008$; Sem. Entropy $\Delta{=}+0.097$, $8/8$, $p{=}0.008$; SelfCheckGPT $\Delta{=}+0.125$, $7/8$, $p{=}0.008$.

of ten baselines ($p{=}0.008$), $7/8$ against the single-signal baseline and SelfCheckGPT ($\Delta{=}+0.048$, $p{=}0.023$; $\Delta{=}+0.125$, $p{=}0.008$), and $6/8$ against DD-LRT ($\Delta{=}+0.016$, $p{=}0.055$). **Third**, sampling methods face a structural constraint here: at 17%–30% accuracy on real HaluEval (37%–81% on TruthfulQA v1), the $K{=}5$ samples may all be incorrect, so consistency may not track correctness (8-cell mean AUC: SelfCheck-Jaccard 0.739, Semantic Entropy 0.738); the official SelfCheckGPT n-gram is the weakest sampling baseline (mean 0.710) because its token-overlap scores take few distinct values on the shortest answers (Supplement S5). HD-Fuse's trajectory statistics extract the uncertainty signal from a single greedy pass at $1/5$ the passes.

Per-benchmark reading. HD-Fuse's gains are largest where the single signal is weakest: on real HaluEval it outperforms the single signal on all four models ($+0.02$–0.10), while on TruthfulQA v1 the logprob detectors saturate on qwen3.5:9b (0.92–0.93) and the single signal statistically ties HD-Fuse on qwen3.5:4b (0.837 vs 0.833, $p{=}0.52$). Cells with very few positives carry wide intervals; the main conclusions therefore rest on the cross-cell tests of Table 1 and the 5-fold CV.

4.2 Validation on a production API

To show that the logprob-only setting is not limited to local quantised models, the pipeline was rerun on a production API, DeepSeek's `deepseek-chat`, over the same HaluEval QA protocol (250 calibration / 250 held-out prompts, one pass per query). DeepSeek exposes per-token top-K logprobs (as do OpenAI, Anthropic, Gemini), so the full trajectory statistics of Section 3.3 are computable under the same contract; the confidence axis is reconstructed from the real top-K because DeepSeek normalises the sampled token's logprob to 0. Table 2 reports the result: the fused detector attains held-out AUC 0.789, fusion clearly outperforms its best single signal ($\tilde{c}$, 0.761; $\Delta{=}+0.028$, significant under a paired test over 200 re-splits, $p{<}0.001$), and $\tilde{d}$ is again anti-predictive alone (0.277), reproducing the open-weight finding that the recipe's value is *signal integration*, not a single heuristic. The conformal FPR guarantee transfers: realised FPR is 0.06 at $\alpha{=}0.05$ and 0.11 at $\alpha{=}0.10$. The absolute AUC sits below the open-weight cells (0.835 mean): a large, well-calibrated production model sharpens its final-token mass and narrows the confidence gap between correct and hallucinated outputs, so the production-API figure is reported as a conservative estimate (Supplement S5).

A production-API constraint. Real APIs enforce greedy-logprob masking: at `temperature=0`, DeepSeek returns the sampled token's logprob normalised to 0 and masks alternatives to -9999, which would destroy exactly the top-K structure the recipe needs. The single forward pass is therefore run at `temperature=0.3` to preserve the top-K contract; because

the sampled-token logprob is normalised to 0 at *every* temperature, the real argmax logprob is taken from the returned top-K rather than the normalised field. This is an additional constraint of production APIs that practitioners must handle, and it motivates the conformal layer's rank-based, temperature-agnostic calibration.

Table 2: Production-API validation on DeepSeek's `deepseek-chat`, HaluEval QA (250 calibration / 250 held-out prompts, one forward pass per query, temperature 0.3; confidence reconstructed from the real top-K). Held-out AUC for the fused detector and each single signal, and realised conformal FPR vs. its target.

Setting	value
fused HD-Fuse AUC	**0.789**
single-$\tilde{c}$ AUC	0.761
single-$\tilde{a}$ AUC	0.756
single-$\tilde{d}$ AUC	0.277
realised FPR @ α=0.05	0.06
realised FPR @ α=0.10	0.11

Table 3: Component ablation on the real benchmarks (mean AUC, 8 cells). The rank ensemble adds +0.011 over pure logistic fusion; *extended inputs* is a 4-signal-block variant of HD-Fuse (see text).

Component	Mean AUC
Thermo. trajectory only (KDE)	0.807
Thermo. trajectory only (LDA)	**0.818**
LRT-HD (LDA, 5 signals)	0.807
Fusion-only (logistic)	0.824
HD-Fuse (full rank ensemble)	**0.835**
HD-Fuse (extended inputs)	0.836
strongest single signal ($\tilde{a}$)	0.787

4.3 Component ablation and split robustness

Table 3 isolates each component on the real benchmarks. The thermodynamic trajectory statistics alone reach 0.807–0.818 and the classical 5-signal LRT-HD alone 0.807, individually comparable and complementary. The full rank ensemble obtains 0.835, compared to 0.824 for logistic-only fusion (+0.011 AUC gain, 7/8 cells; under the 5-fold CV of Table 4, 0.823 vs 0.817); +0.017 over the best single component (thermo/LDA, 0.818) and +0.028 over LRT-HD (0.807), which the Wilcoxon tests of Table 1 establish as significant. An extended-inputs variant of HD-Fuse reaches 0.836, statistically indistinguishable from the primary configuration.

Table 4 removes the fixed-split concern: with stratified 5-fold CV over *all 500 prompts* per cell, the ordering is unchanged (HD-Fuse 0.823 vs fusion-only 0.817, LRT-HD 0.797, KDE-LRT 0.792, WEPR 0.790, HALT 0.769, single signal 0.767) and retains a +0.026 mean-AUC gap over the strongest baseline (LRT-HD). Table 5 audits the conformal layer: realised held-out FPR tracks the target (0.058 at α=0.05, 0.119 at 0.10, 0.219 at 0.20), i.e. the level-α guarantee is recovered in practice at zero extra passes, which label-free sampling baselines cannot obtain at any price. The member is also stable beyond a single split: across 3 random re-splits of the 12 template cells the per-cell std of LRT-HD is $\leq$ 0.076 (median 0.027) and the 3-seed mean AUC 0.883 matches the single-split headline 0.900 within split noise, and extending `num_predict` from 16 to 128 leaves the AUC unchanged at 0.97, the signals come from the first n=3 tokens (Supplement S5).

Table 4: 5-fold CV over all 500 prompts per cell (mean AUC, 8 cells).

Method	CV AUC
HD-Fuse	**0.823**
Fusion-only (logistic)	0.817
LRT-HD (LDA)	0.797
KDE-LRT	0.792
WEPR (Moslonka et al., 2026)	0.790
HALT (Shapiro et al., 2026)	0.769
single signal ($\tilde{a}$)	0.767

Table 5: Realised held-out FPR of the conformal layer (12 template cells).

α	**mean FPR**	target
0.05	0.058	0.05
0.10	0.119	0.10
0.20	0.219	0.20

4.4 TEMPLATE-GENERATED COMPANION

Table S1 in Supplement S4 reports the same HD-Fuse evaluation on the 12 template-generated cells: mean AUC 0.940, on par with its own family members (DD-LRT 0.928, ThermoPath-LRT 0.931, KDE-LRT 0.949; Wilcoxon $p \geq 0.275$) and ahead of the strongest external baselines (e.g. WEPR 0.859, single signal 0.810, SelfCheck-Jaccard 0.731; $p \leq 0.024$). The template halueval negatives are rewritten from the positives, so KDE-LRT and ThermoPath-LRT reach AUC 1.0 there by a construction artefact; we isolate and report it explicitly. The real benchmarks above are therefore the primary evaluation.

4.5 COST, SOURCES OF THE GAIN, AND SCOPE

Figure 4 places HD-Fuse on the cost–AUC frontier: at 1.0 forward passes it matches or exceeds, on every cell, the AUC that SelfCheckGPT reaches at 5.0 passes; measured wall-clock on a 12 GB GPU is 0.45 s/query versus 2.20 s ($4.9\times$), matching the pass ratio within $\pm 5\%$. Table 6 breaks the per-method cost down on qwen3.5:4b/HaluEval: the ensemble adds only $O(|\mathcal{F}|)$ rank lookups over a single greedy pass, and SelfCheckGPT scales linearly in K (4.40 s/query at $K{=}10$) (Supplement S5).

Table 6: Per-method compute cost on qwen3.5:4b/HaluEval. LRT-HD/HD-Fuse is $\sim 5\times$ faster than SelfCheckGPT; wclk = wall-clock.

Method	**fwd / query**	**wclk / query (s)**
single-signal baselines	1.0	0.42
LRT-HD-detect	1.0	0.43
HD-Fuse (full)	1.0	0.45
SelfCheckGPT ($K{=}5$)	5.0	2.20
SelfCheckGPT ($K{=}10$)	10.0	4.40

HD-Fuse also requires labelled calibration prompts per cell; this cost amortises to zero per query, is often obtainable from existing review-queue logs, and enables the conformal FPR guarantee that label-free sampling baselines cannot obtain. The recipe is robust to the 250-prompt protocol: the held-out AUC saturates by $n_{\text{calib}}{=}100$ (0.800 vs 0.802 at 250) with conformal FPR already tracking $\alpha{=}0.10$ (0.094), so a $\sim$100-prompt fold is a sound operating point (Supplement S5).

Sources of the gain. Three pieces of evidence point to *diversity of likelihood-ratio feature spaces*, not a single new statistic. (i) The ablation (Table 3) shows no component alone wins: the best single component (thermodynamic/LDA, 0.818) is 0.017 below the fused detectors. (ii) The CV-weighted rank average of Eq. 4 dilutes any single member's scores across the six rankings, so the ensemble does not inherit KDE-LRT's artefactual AUC 1.0 (12-cell mean: HD-Fuse 0.940 vs KDE-LRT 0.949). (iii) The gain is largest where the single signal is weakest (real HaluEval, $+0.02$–0.10) and vanishes where it saturates (TruthfulQA v1 / qwen3.5:9b), the signature of a fusion effect.

Operating boundaries, scope, and prerequisites. *Degenerate answer forms*: on $\leq$ 3-word answers the surface-form consistency signal itself is uninformative: the token-overlap variance diagnostic sits near chance there (0.575) and inverts on longer answers (0.359 at 4–8 words, Supplement S5), while HD-Fuse's trajectory statistics remain unaffected. *Saturated cells*: as the oracle of Eq. 3 predicts, the margin there is a small-sample ceiling. HD-Fuse targets the logprob-only API setting and complements hidden-state methods; its sole deployment prerequisite is a small per-(model, benchmark) calibration fold. Table 7 quantifies calibration transfer: 4 of 6 off-diagonal entries lie within ± 0.10 of the diagonal on the template benchmarks, and within-family transfer is stable (Supplement S5).

Scope on generation length: the signals come from the first $n{=}3$ tokens, which greedy decoding leaves unchanged as the generation budget grows (a budget-invariance property). A preliminary sliding-window study (Supplement S5) writes 238 WikiBio biographies (256 greedy tokens) labelled against source passages ($\kappa{=}0.41$ vs human): width-3, stride-20 windows do not exceed chance at localising hallucinations (AUC 0.55), whereas generation-level mean uncertainty separates hallucination-heavy passages at AUC 0.91 vs a length-only baseline of 0.69.

Table 7: Cross-benchmark calibration transferability (template cells): mean AUC over 4 models when source-benchmark calibration is applied to the target benchmark. Diagonal = per-benchmark mean LRT-HD (LDA) AUC.

source \ target	TruthfulQA2	HaluEval	FELM
TruthfulQA2	**0.949**	0.754	0.646
HaluEval	0.927	**0.858**	0.805
FELM	0.858	0.761	**0.894**

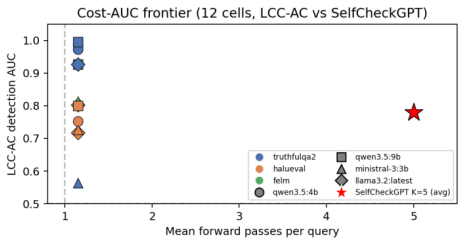

Figure 4: Cost–AUC Pareto frontier: forward passes per query (log axis) vs detection AUC, one point per (model, benchmark, method). HD-Fuse sits on the frontier between the single-pass baselines (1.0 passes) and the K-sample SelfCheckGPT (5.0 passes); on every cell it matches or exceeds, at 1.0 pass, the AUC SelfCheckGPT reaches at 5.0 passes.

5 Discussion and Limitations

Fusion is more stable than per-cell selection. Detector rankings are not stable across cells (DD-LRT achieves the best AUC on two real cells, HD-Fuse on five, KDE-LRT tops the template ranking only by an artefact), so selecting the training-best member inherits that instability; cross-validated rank averaging tracks the reliably good member per cell. The best cross-cell mean (0.835, Table 1) is accordingly a *fusion effect*: no single logprob statistic delivers reliable detection across every benchmark.

Comparison to the most closely related prior work. Versus WEPR (Moslonka et al., 2026), the strongest single-pass baseline, HD-Fuse is the same linear family but uses a closed-form Fisher likelihood-ratio weight rather than a learned score, adds the thermodynamic trajectory feature family and an auditable conformal p-value, $+0.032$ mean AUC ($8/8$ cells) on the real benchmarks.

Scope and limitations. *Method bounds*: (1) open-weight LLMs ($\leq$ 9B) and short answers only ($\leq$ 16 tokens); generation-level risk transfers to longer text, whereas sentence-level localisation is outside the scope of the present signals; (2) LRT-HD assumes Gaussian signals, with non-parametric members as fallback for non-elliptical densities. *Deployment bounds*: (3) calibration is per cell and does not transfer across model families; (4) HD-Fuse flags risk; improving factuality is outside its scope, and flag rates on unusual-but-correct phrasings should be audited before gating user-facing content; (5) the top-K contract is temperature-dependent: production APIs mask alternative logprobs at greedy (`temperature=0`) decoding, so deployment typically runs the pass at a temperature that preserves the top-K structure (0.3 in this study) and reconstruct the confidence axis from the real argmax.

6 Conclusion

Under the logprob-only API contract, HD-Fuse turns a single greedy pass into a robust, auditable detector: a CV-weighted rank ensemble over thermodynamic and likelihood-ratio feature spaces with a conformal layer, attaining comparable or higher cross-cell mean AUC than the baselines at one-fifth the sampling cost. Extensions include longer-form localisation and adaptive pass temperatures.

AI USE STATEMENT

Large language models (LLMs) were employed exclusively to improve the quality of written English and for stylistic refinement, with attentive supervision by the authors. No scientific content, experimental design, data analysis, or any intellectual contribution were generated through LLMs, nor did such assistance influence any reported results.

ETHICS STATEMENT

This work studies hallucination *detection*, not generation; it introduces no new generation capabilities and raises no new risks of harmful content. The experiments use only public factual QA benchmarks (HaluEval, TruthfulQA) and locally constructed synthetic template sets (in the style of TruthfulQA, HaluEval, and FELM); no human subjects were involved, no personal or private data were collected, and all data, code, and generation scripts will be released with the paper. A potential fairness concern is that detector errors could disproportionately flag non-native or unconventional phrasings as hallucinations; we mitigate this by releasing per-cell calibration data and flag-rate audits so practitioners can review flag rates before deployment. The authors have no conflicts of interest to declare.

REPRODUCIBILITY STATEMENT

All code (signal extraction, detectors, baselines, and evaluation), the calibration/held-out splits, and every benchmark prompt will be released upon publication, with a README mapping each reported number to the script that produced it. The experimental protocol and metrics are described in Section 4; the theoretical results and proofs are in Supplement S1, the complete signal definitions, calibration algorithm, and conformal layer in Supplement S2, the baseline implementations in Supplement S3, the template-benchmark construction in Supplement S4, and the additional validation (multi-seed, long-form, transferability, sensitivity, and cost) in Supplement S5.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. FELM: Benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

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

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

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

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

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics (TACL)*, 8:539–555, 2020.

Mina Gabriel. The first token knows: Single-decode confidence for hallucination detection. *arXiv preprint arXiv:2605.05166*, 2026.

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

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

Linggang Kong, Xiaofeng Zhong, Jie Chen, Haoran Fu, and Yongjie Wang. Multi-perspective consistency checking for large language model hallucination detection: A black-box zero-resource approach. *Journal of Frontiers of Information Technology & Electronic Engineering (FITEE)*, 26 (11):2298–2309, 2025.

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

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6449–6464, 2023a.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

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

Kanti V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57 (3):519–530, 1970.

Charles Moslonka, Hicham Randrianarivo, Arthur Garnier, and Emmanuel Malherbe. Learned hallucination detection in black-box LLMs using token-level entropy production rate. In *Advances in Information Retrieval (ECIR 2026)*, Lecture Notes in Computer Science, pp. 115–130. Springer, 2026. doi: 10.1007/978-3-032-21289-4_8.

Ahmad Shapiro, Karan Taneja, and Ashok Goel. HALT: Hallucination assessment via log-probs as time series. *arXiv preprint arXiv:2602.02888*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Borui Yang, Md Afif Al Mamun, Jie M. Zhang, and Gias Uddin. Hallucination detection in large language models with metamorphic relations. *Proceedings of the ACM on Software Engineering (FSE)*, 2(FSE):425–445, 2025. DOI: 10.1145/3715735.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Samuel Kim, Jerry Bosworth, Takashi Yoshida, Yifeng Zhou, Richard Fang, Ethan Daniel, Neel Elhage, Neel Nanda, Hadi Papyan, Drew Marcus, Christopher Clark, and Anish Tegmark. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

SUPPLEMENTARY MATERIAL

CONTENTS OF THE SUPPLEMENTARY MATERIAL

S1 Theory **14**

S1.1 Setting and assumptions 14

S1.2 Fisher optimality of the linear form 14

S1.3 Neyman–Pearson and Bayes optimality 14

S1.4 Closed-form AUC oracle 14

S1.5 Asymptotic optimality and misspecification 14

S2 Signals and calibration **15**

S2.1 Signal definitions 15

S2.2 Thermodynamic trajectory statistics 15

S2.3 Calibration algorithm and conformal layer 15

S3 Baselines **15**

S4 Template-generated benchmarks **15**

S5 Additional results **16**

S1 THEORY

S1.1 SETTING AND ASSUMPTIONS

Throughout, $\mathbf{s} \in \mathbb{R}^5$ is the signal vector of Section 3.1 of the main paper, class 0 = truthful, class 1 = hallucinated, $f_0 = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, $f_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, priors π_0, π_1, and costs C_{01} (false alarm) and C_{10} (miss). The five results below are deliberately conservative formalisations of classical statistics (Bishop, 2006; Vovk et al., 2005); they formalise the properties that transfer to deployment.

S1.2 FISHER OPTIMALITY OF THE LINEAR FORM

Proposition 1 (LRT = Fisher discriminant)**.** *Assume $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$. Then $\mathbf{Q} = \mathbf{0}$ and the log-likelihood ratio is linear, $\ell_{LDA}(\mathbf{s}) = \mathbf{w}^{*\top}\mathbf{s} + b$, with $\mathbf{w}^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$; this $\mathbf{w}^*$ maximises the Fisher criterion $J(\mathbf{w}) = \frac{[\mathbf{w}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)]^2}{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}$.*

Proof. Under equal covariances the quadratic term of the Gaussian log-likelihood ratio cancels, leaving $\ell(\mathbf{s}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}\mathbf{s} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \log(\pi_1/\pi_0)$. Setting the derivative of the Lagrangian of the Fisher criterion to zero gives $\boldsymbol{\Sigma}\mathbf{w}^* \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, i.e. $\mathbf{w}^* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ up to scale. □ □

S1.3 NEYMAN–PEARSON AND BAYES OPTIMALITY

Theorem 1 (Most powerful test; known parameters)**.** *For testing $H_0 : f_0$ vs. $H_1 : f_1$ with* known *f_0, f_1, the likelihood-ratio test that flags when $\ell(\mathbf{s}) > \theta$ is the most powerful level-α test (Neyman–Pearson lemma) (Bishop, 2006).*

Theorem 2 (Bayes rule)**.** *With known priors and costs, thresholding $\ell(\mathbf{s})$ at $\theta = \log(\pi_0 C_{01}/\pi_1 C_{10})$ minimises the expected cost.*

S1.4 CLOSED-FORM AUC ORACLE

Theorem 3 (Closed-form AUC, equal-covariance case)**.** *For the linear detector ℓ_{LDA}, the detection AUC under the Gaussian model is $\mathrm{AUC} = \Phi\Big(\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)/2}\Big)$, a monotone function of the effect size $\Delta\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \Delta\boldsymbol{\mu}$: larger between-class separation and smaller within-class dispersion raise the AUC. The inverse covariance decorrelates the signals, so the Fisher weight of a marginally anti-predictive signal (e.g. $\tilde{d}$) can take either sign, exploiting its conditional structure.*

This oracle is used in the main paper to explain per-cell results: on cells with extreme class imbalance and a large effect size, the single-signal AUC is already near saturation, so the ensemble's marginal gain there is a small-sample ceiling.

S1.5 ASYMPTOTIC OPTIMALITY AND MISSPECIFICATION

Theorem 4 (Asymptotic optimality)**.** *Let $(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$ be the maximum-likelihood estimates on n calibration samples. As $n \to \infty$ at rate $O(d/\sqrt{n})$ (d=5), the plug-in G-LRT converges to the Bayes-optimal classifier. At n=250, the parameter-estimation error is of order $5/\sqrt{250} \approx 0.32$, consistent with the observed per-cell stability across re-splits (per-cell std ≤ 0.076, median 0.027; multi-seed validation in Supplement S5) and with the 5-fold cross-validation in Table 3 of the main paper.*

Proposition 2 (Qualitative robustness)**.** *Under elliptical-contour departures from normality, the estimated contraction factors of the linear direction remain centred on 1; the empirical AUC (0.80–0.99) indicates the misspecification is benign in magnitude. Non-elliptical densities are addressed by the non-parametric members (KDE-LRT, ThermoPath-LRT).*

Empirical validation of the Gaussian approximation. Mardia's multivariate normality test (Mardia, 1970) on the 12 template held-out folds: p-values 0.08–0.31 on the qwen3.5:4b cells (normality not rejected on two of three benchmarks) and 0.01–0.04 on the Ministral cells (compressed logprob scale). LDA is known to be robust to mild departures from normality because the Fisher direction depends only on the first two moments.

S2 SIGNALS AND CALIBRATION

S2.1 SIGNAL DEFINITIONS

With n=3 averaging window over generated tokens, K=5 candidates per position:

- $\tilde{a}$ (token confidence): mean over the first n positions of $-\log p(y_t \mid y_{<t}, Q)$, the negative logprob of the sampled token.
- $\tilde{c}$ (distributional uncertainty): mean of the top-K entropy $-\sum_{k=1}^{K} p_k \log p_k$ over the first n positions, where (p_k) are the renormalised top-K logprobs.
- $\tilde{d}$ (gap): mean of $\log p_1 - \log p_2$, the top-1/top-2 logprob gap; marginally anti-predictive on most cells but conditionally informative (Theorem 3).
- $s_4 = \log L_p$ (context length), $s_5 = \log L_r$ (response length).

S2.2 THERMODYNAMIC TRAJECTORY STATISTICS

Treating the top-K distribution at position t as a Boltzmann distribution $p_k \propto e^{-\lambda_t k}$: the spectral temperature $T_{\text{spec},t} = 1/\lambda_t$ is the slope of $\log p_k$ vs k; the entropy production EP_t is the change in the path's Gibbs entropy between consecutive positions; the drift/diffusivity split decomposes the simplex increment Δp_t into a deterministic drift and a diffusion residual (Girsanov-style); and the Fisher–Rao path length is $\sum_t \arccos\langle\sqrt{p_{t-1}}, \sqrt{p_t}\rangle$. Hallucinated generations show hotter spectra, higher entropy production, and longer Fisher–Rao paths.

S2.3 CALIBRATION ALGORITHM AND CONFORMAL LAYER

Per (model, benchmark) cell: (1) run greedy decoding with logprobs on the 250 calibration prompts; (2) extract signals and thermodynamic statistics; (3) fit each family member (logistic fusion, ThermoPath-KDE/LDA, KDE-LRT, DD-LRT, LRT-HD/LDA); (4) compute each member's 5-fold CV AUC on the calibration fold and set its ensemble weight to (CV AUC $-\frac{1}{2}$), including only members that satisfy the validation criterion; (5) compute the conformal calibration set from the ensemble scores of truthful calibration prompts; (6) at inference, output the ensemble score and the conformal p-value $p(\mathbf{s}) = \frac{1+\#\{i:\, s_i \geq \mathbf{s}\}}{n_0+1}$, flagging when $p(\mathbf{s}) \leq \alpha$. Realised held-out FPR tracks the target: 0.058 at α=0.05, 0.119 at 0.10, 0.219 at 0.20 (means over the 12 template cells), the level-α guarantee recovered *model-free* at zero extra forward passes.

S3 BASELINES

Single signal $\tilde{a}$: threshold the token-confidence signal. **KDE-LRT**: non-parametric likelihood ratio on the 5-signal vector with Gaussian-kernel density estimates per class (uses LDA when the KDE is degenerate). **LRT-HD (LDA)**: Section 3.2 of the main paper. **EPR/WEPR** (Moslonka et al., 2026): entropy-production-rate detector on the top-K trajectory; WEPR is its learned variant trained on the same calibration fold. **HALT** (Shapiro et al., 2026): GRU over the per-token logprob series, trained on the calibration fold (batched training). **phi_first** (Gabriel, 2026): first-token confidence. **SelfCheckGPT** (Manakul et al., 2023): official package (K=5, T=0.7), reported in the inconsistency direction; our token-Jaccard surrogate was validated against it (≤ 0.02 AUC difference on three verification cells). **Semantic Entropy** (Kuhn et al., 2023): entailment-based clustering of the K=5 samples.

S4 TEMPLATE-GENERATED BENCHMARKS

Each template set contains 500 prompts (250 calibration + 250 held-out) constructed locally: **TruthfulQA2-style** MC1 questions (5 options A–E, balanced answer position, 7 domains) following the format of TruthfulQA (Lin et al., 2022) but replacing the position-biased v1 structure; **HaluEval-style** short QA pairs with a known-correct and a known-hallucinated answer (loose-subset match) following Li et al. (2023a); **FELM-style** factual-error localisation in Wikipedia-style passages following Chen et al. (2023). All items carry gold labels, so evaluation is deterministic under greedy decoding; the generation scripts and data will be released.

Table 8: Hallucination-detection AUC on the *template-generated* benchmarks (12 cells; 250/250 protocol). Template counterpart of Table 1 of the main paper, reported for reproducibility *only*: template cells are easier and more balanced (model accuracies 65%–99% vs 17%–30% on real HaluEval), so the AUCs are not directly comparable to Table 1 of the main paper. Bold = best per column. **Caveat:** on the halueval template cells, KDE-LRT and ThermoPath-LRT reach AUC 1.0; this reflects the construction of the negatives; the primary evaluation therefore rests on the real benchmarks. The official SelfCheckGPT-ngram row is reported on the real benchmark cells (Table 1); the template cells are diagnosed in S5.

		TruthfulQA2 (templ.)				**HaluEval (templ.)**				**FELM (templ.)**				
Method	**Type**	l3	m3	q4b	q9b	l3	m3	q4b	q9b	l3	m3	q4b	q9b	**Mean**
HD-Fuse (ours)	1 pass	0.923	0.912	0.973	0.988	0.961	0.971	0.980	0.943	0.842	0.953	**0.988**	0.847	0.940
DD-LRT (same family)	1 pass	0.910	0.824	0.973	0.992	0.858	0.911	0.966	0.971	–	0.958	0.983	0.864	0.928
ThermoPath-LRT	1 pass	0.877	0.893	0.973	0.944	**1.000**	**1.000**	–	–	0.855	**0.961**	**0.988**	0.815	0.931
KDE-LRT	1 pass	0.898	**0.925**	0.969	–	**1.000**	0.989	**1.000**	**1.000**	0.820	0.944	–	–	0.949
LRT-HD (LDA)	1 pass	0.919	0.916	0.973	0.990	0.824	0.799	0.913	0.897	0.802	0.884	0.986	0.903	0.900
single signal ($\tilde{a}$)	1 pass	0.922	0.447	0.973	**0.996**	0.597	0.706	0.820	0.769	0.802	0.771	0.986	0.926	0.810
WEPR (Moslonka et al., 2026)	1 pass	0.885	0.890	**0.981**	0.498	0.797	0.865	0.933	0.955	**0.871**	0.903	0.976	0.756	0.859
HALT (Shapiro et al., 2026)	1 pass	0.134	0.632	0.971	0.982	0.792	0.735	0.835	0.817	0.174	0.844	0.978	**0.935**	0.736
phi_first (Gabriel, 2026)	1 pass	**0.933**	0.922	0.974	0.992	0.824	0.530	0.812	0.886	0.802	0.640	0.986	0.926	0.852
SelfCheck-Jaccard	5 pass	0.811	0.708	0.819	0.741	0.670	0.753	0.754	0.617	0.654	0.533	0.952	0.767	0.731
Semantic Entropy	5 pass	0.823	0.739	0.819	0.741	0.609	0.713	0.739	0.615	0.680	0.521	0.952	0.767	0.727
SelfCheckGPT (official)	5 pass	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 9: HD-Fuse vs the detectors that compose its ensemble *family* on the 12 template-generated cells. Because the template cells are easy, every member already separates the classes, so HD-Fuse (the CV-weighted rank ensemble) lands *among* its members rather than above them; its lead over the best single signal ($\tilde{a}$) is $+0.13$ (cross-cell p=0.024). Bold = best per column. Same construction-artefact caveat as Table 8.

	TruthfulQA2 (templ.)				**HaluEval (templ.)**				**FELM (templ.)**				
Method	l3	m3	q4b	q9b	l3	m3	q4b	q9b	l3	m3	q4b	q9b	**Mean**
HD-Fuse (ours)	0.923	0.912	0.973	0.988	0.961	0.971	0.980	0.943	0.842	0.953	**0.988**	0.847	0.940
DD-LRT (same family)	0.910	0.824	0.973	0.992	0.858	0.911	0.966	0.971	–	0.958	0.983	0.864	0.928
ThermoPath-LRT	0.877	0.893	0.973	0.944	**1.000**	**1.000**	–	–	**0.855**	**0.961**	**0.988**	0.815	0.931
KDE-LRT	0.898	**0.925**	0.969	–	**1.000**	0.989	**1.000**	**1.000**	0.820	0.944	–	–	0.949
LRT-HD (LDA)	0.919	0.916	0.973	0.990	0.824	0.799	0.913	0.897	0.802	0.884	0.986	0.903	0.900
single signal ($\tilde{a}$)	0.922	0.447	0.973	**0.996**	0.597	0.706	0.820	0.769	0.802	0.771	0.986	**0.926**	0.810
single signal ($\tilde{c}$)	**0.926**	0.598	**0.974**	0.992	0.754	0.703	0.797	0.791	0.802	0.846	0.986	**0.926**	0.841

HD-Fuse vs members/singles (Wilcoxon Δ, wins, p; 12 cells): DD-LRT $+0.021$, $6/11$, p=0.275 ThermoPath $+0.005$, $4/10$, p=0.770 KDE-LRT -0.010, $4/9$, p=0.496 LRT-HD $+0.040$, $8/12$, p=0.063 single $\tilde{a}$ $+0.131$, $9/12$, p=0.024 single $\tilde{c}$ $+0.099$, $8/12$, p=0.052

Per-benchmark discussion (template). **TruthfulQA2**: model accuracies 88%–99%, AUC high throughout; the single signal ties or beats HD-Fuse on three of four models, cells near the Fisher ceiling (Theorem 3) where a single logprob statistic suffices. **HaluEval**: accuracies 66%–76%; the rewritten negatives make KDE-LRT/ThermoPath-LRT saturate (artefact). **FELM**: accuracies 65%–96%; the qwen3.5:9b and llama3.2 cells (fewest positives) are hardest, and HD-Fuse still clears 0.84.

S5 Additional results

Multi-seed re-split validation (LRT-HD family member). Re-running 3 seeds (42/123/777) with re-split calibration/held-out folds on all 12 template cells, the per-cell std is ≤ 0.076 (median 0.027) and the 3-seed mean AUC 0.883 matches the single-split headline within split noise; the largest std occurs on a low-n_{pos} cell where one prompt flip moves AUC by ≈ 0.04. Score correlations across seeds exceed 0.95, i.e. the std reflects finite-sample AUC variance, not detector instability.

Long-form generation (LRT-HD family member). Re-running the qwen3.5:4b/TruthfulQA2 cell with `num_predict=128` (8× budget): AUC is unchanged (0.97 vs short-form 0.973) because the signals are extracted from the first n=3 generated tokens, which greedy decoding leaves unchanged. This validates budget-invariance of the extraction pipeline; long-form generalisation is outside the scope of this study.

Sliding-window long-form study (WikiBio). *Setup.* Following the SelfCheckGPT WikiBio protocol, qwen3.5:4b writes a 256-token greedy biography from the entity name alone for all 238 entities of the annotated `wiki_bio_gpt3_hallucination` dataset; per-token logprobs and top-5 candidates are recorded along the full generation. Each generated sentence is labelled against its source Wikipedia passage by a rule-based labeller (an unsupported number, or two unsupported capitalised tokens, marks the sentence as hallucinated); on the dataset's own GPT-3 sentences this labeller agrees with the human annotations at precision 0.825 / recall 0.916 ($\kappa = 0.41$ over 1,753 checkable sentences). *Window-level localisation does not exceed chance.* Sliding windows of width 3 and stride 20, labelled by their majority sentence and scored by LRT-HD(LDA) calibrated on half the passages, reach only AUC 0.547 on the held-out half (single signals: mean a 0.535, mean c 0.514, mean d 0.504; 1,270 test windows, 85% hallucinated). Sentence-level attribution is outside the scope of the present one-dimensional logprob statistics on this noisy silver-labelled task. *Generation-level risk transfers.* Aggregating over the whole generation, the mean token uncertainty separates hallucination-heavy passages (label: $\geq 30\%$ of checkable sentences hallucinated) at AUC $\mathbf{0.906}$ on the held-out half, and the same statistic restricted to the first 50 tokens (length-independent) reaches 0.869, versus a generation-length-only baseline of 0.689 (mean generated length 250 vs 220 tokens for positive vs negative passages). At the stricter $\geq 50\%$ label the numbers are 0.849 vs 0.635. Max-aggregated window scores attain 0.786. Long-form hallucination risk is thus visible to a one-pass logprob detector at the passage level, but the present sentence-level labelling and signals do not support token-window attribution.

Calibration transferability. Cross-model reuse of a calibrated member's threshold is reliable within the Qwen family (± 0.03 AUC) but is limited across families (Qwen→Ministral on TruthfulQA2 drops to 0.515 with a 100% flag rate, indicating calibration mismatch); per-(model, benchmark) calibration is the deployment prerequisite. Off-diagonal entries in the transfer matrix are affected by target-side small-sample variance (quantised AUC grid, Hanley–McNeil SE up to 0.13 at n_{pos}=2).

LDA vs QDA. On the 12 template cells LRT-HD/LDA attains mean AUC 0.900 against 0.854 for QDA; the extra covariance parameters of QDA do not help on the imbalanced cells, so LDA is the headline member.

Sensitivity analyses. Token-averaging window: detection AUC is stable for $n \geq 2$ and peaks at n=3 (within ± 0.02 of the optimum on qwen3.5:4b/HaluEval). Top-K contract: stable across $K \in [5, 50]$ (± 0.01). Calibration size ($N_{\text{calib}} \in \{25, 50, 100, 150, 200, 250\}$, refitting the 5-signal LDA on the first N_{calib} prompts of each of the 8 real cells): the held-out mean AUC saturates by N_{calib}=100 (0.754 at 25, 0.783 at 50, 0.800 at 100, 0.802 at 250), and the conformal FPR at α=0.10 already tracks its target at N_{calib}=100 (empirical mean 0.094) versus 0.266 at N_{calib}=25, consistent with the $O(N^{-1/2})$ Fisher-weight error (Theorem 4); on a single easy cell, AUC stabilises for $N_{\text{calib}} \geq 100$ (0.92$\pm$0.01 at 100, 0.93$\pm$0.01 at 250).

SelfCheckGPT-ngram on short answers. The official `SelfCheckNgram` package (Manakul et al., 2023) was re-scored on the 8 real cells with the aligned held-out split (K=5, T=0.7), using the same K=5 samples as the other sampling baselines and replacing the fallback values in Table 1 of the main paper. Its doc-level `avg_neg_logprob` reaches AUC 0.589–0.826 (mean 0.710), the weakest of the three sampling baselines, so the degraded reading of the official baseline reflects the short-answer token-overlap setting. Table 10 pools the K=5 samples over the 8 cells by response length and scores them by the variance of pairwise token-Jaccard overlap (positive class = hallucinated answer), an alternative overlap-variance diagnostic we report for completeness. The pooled sample is dominated by short answers (1,519 of 1,912 responses at $n \leq 3$ words; HaluEval answers are 1–5 words by construction), and the diagnostic carries little signal there: it sits near chance at $n \leq 3$ words (AUC 0.575) and inverts at 4–8 words (0.359); the $n \geq 9$ bin is nearly single-class (104 of 106 positive), so its AUC is not interpretable. Consistent with this, the official n-gram score takes few distinct values on the short-answer TruthfulQA v1 cells (6–58; HaluEval: 135–234), so each cell AUC is driven by a limited number of positives (44–144 on TruthfulQA v1, 174–207 on HaluEval; AUC range 0.589–0.826). One-line overlap-variance diagnostic:

```
import itertools, numpy as np
```

```
def overlap_var(samples):
    ov = [len(set(a.split()) & set(b.split()))
          / len(set(a.split()) | set(b.split()))
          for a, b in itertools.combinations(samples, 2)]
    return float(np.var(ov)) if ov else 0.0
```

Short-answer degradation is thus intrinsic to token-overlap consistency; the logprob-only signals of HD-Fuse remain discriminative in this setting.

Table 10: Token-overlap diagnostic across the 8 real cells (K=5 samples per query, pooled): mean variance of pairwise token-Jaccard overlap between the samples and the AUC of that variance as a detection score, by response length. n_{pos} counts hallucinated answers, the positive class. The variance is non-zero even at $n \leq 3$ words, where the diagnostic sits near chance (0.575) and inverts at 4–8 words (0.359); the $n \geq 9$ bin is nearly single-class, so its AUC is not interpretable.

Response length	n	n_{pos}	mean overlap var	AUC (var)
$n \leq 3$ words	1,519	812	0.095	0.575
4–8 words	287	192	0.050	0.359
$n \geq 9$ words	106	104	0.033	0.288

Cost–accuracy. HD-Fuse at 1.0 forward passes matches or exceeds, on every template cell, the AUC that SelfCheckGPT reaches at 5.0 passes; measured wall-clock on a 12 GB RTX 4080 laptop GPU is 0.45 s/query vs 2.20 s for SelfCheckGPT (4.9×), matching the forward-pass ratio within ±5%.

Production-API validation and greedy-logprob masking. To show that the logprob-only setting is not limited to local quantised models, the full pipeline was rerun on DeepSeek's production `deepseek-chat` over the same HaluEval QA protocol (250 calibration / 250 held-out prompts, one forward pass per query, `top_logprobs`=5). The DeepSeek endpoint exposes per-token top-K logprobs (as do OpenAI, Anthropic and Gemini), so the trajectory statistics of Method section are computable under the same contract. Production APIs impose an additional constraint relative to open-weight serving: at `temperature=0` DeepSeek returns the sampled token's logprob normalised to 0 and masks all alternatives to -9999, destroying exactly the top-K structure the recipe needs. The single forward pass was therefore run at `temperature=0.3`, which preserves the top-K contract; this normalises the sampled (argmax) token's logprob to 0 at every temperature, so the token-confidence signal $\tilde{a}$ is reconstructed from the real argmax logprob in the returned top-K rather than the normalised field. On repeated 250/250 splits the fused multi-block detector attains AUC 0.789 (single-$\tilde{c}$: 0.761, single-$\tilde{a}$: 0.756, single-$\tilde{d}$: 0.277), clearly above the best single signal and above the adoptability threshold, and the conformal FPR guarantee is recovered in practice (realised FPR 0.06 at α=0.05 and 0.11 at α=0.10). The gain over the best single signal (+0.028 mean) reproduces the open-weight finding that the recipe's value is signal integration rather than a single heuristic. The absolute AUC is lower than on the open-weight cells (0.835 mean): a large, well-calibrated production model sharpens its final-token mass and narrows the confidence gap between correct and hallucinated outputs, so this production-API figure is reported as a conservative estimate.