# AWARD: ANSWER-LEVEL WEIGHTED AGGREGATION OF RETRIEVAL-FREE DIAGNOSTICS FOR HALLUCINATION DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Hallucination in large language model (LLM) outputs is often flagged by inference-time signals that are then used directly as answer-level scores, despite measuring a different quantity. This paper presents AWARD, a retrieval-free hallucination detector that separates a formal sequential source diagnostic from a supervised answer-level score. The core signal is Speculative Cross-Model Divergence (SCMD), the pointwise log ratio between an instruction-tuned model and its own pretrained base. Under the base-sampling null, the SCMD product is an e-process satisfying Ville's anytime type-I bound. SCMD is one of four inference-time primitives, joined by primary-model entropy, Attention-Grounding Entropy (AGE), and Semantic Refutation via NLI (SRN). Across five model families and three benchmarks (15 cells), nested cross-validation yields mean AUROCs of $0.735$, $0.689$, and $0.652$ on HaluEvalQA, MatSciQA, and TruthfulQA, respectively. In comparisons of 10-seed paired-bootstrap with the strongest of DoLa, SelfCheckGPT, and Semantic Entropy, AWARD achieves mean $\Delta\mathrm{AUROC} = +0.089$, with positive differences in $14/15$ cells; two survive Holm–Bonferroni correction in family $\alpha$=0.05. AWARD wins $12/15$ cells with mean $\Delta = +0.080$ against the stronger of two classifier- and label-matched supervised fusions. Split-conformal calibration provides marginal false-positive-rate control under exchangeability. At the lowest non-degenerate tier, $\alpha$=0.05, $12/15$ cells meet the seed-averaged target and $8/15$ meet an item-clustered $95\%$ upper-confidence-bound criterion; the $\alpha$=0.01 tier is degenerate in every cell because the calibration sets are too small. All four primary feature blocks have positive mean contributions in leave-one-block-out ablations.

## 1 INTRODUCTION

Hallucination in large language models is a well-documented obstacle to their reliable deployment (Ji et al., 2023; Huang et al., 2025). Existing detectors for hallucination during generation fall into two categories, each with a distinct limitation. Detectors based on sequential testing provide formal anytime-valid guarantees on false-positive rate under Ville's inequality (Ramdas et al., 2023; Howard et al., 2021; Ville, 1939; Shafer & Vovk, 2019), but they typically require a retrieval corpus that itself encodes the target misconception on benchmarks such as TruthfulQA (Lin et al., 2022). Empirical detectors, by contrast, rely on model behavior rather than formal statistical guarantees. DoLa contrasts layers to sharpen factual output during decoding, and it doubles as a hallucination score (Chuang et al., 2024). Semantic Entropy samples multiple generations, clusters them by semantic equivalence, and reports the entropy over clusters (Kuhn et al., 2023). SelfCheckGPT samples generations and measures inter-sample disagreement (Manakul et al., 2023). The empirical detectors achieve competitive AUROC on public benchmarks but provide no formal FPR guarantee. No prior framework unifies both properties in a single retrieval-free detector. We close this gap by constructing an e-process directly from model-internal signals, with no external retrieval and no additional training. Our starting point is the pointwise log-ratio between an instruction-tuned model $P_I$ and its own pretrained base $P_B$ at the sampled token, which we call Speculative Cross-Model Divergence (SCMD). Because $P_I$ and $P_B$ share a tokenizer, vocabulary, and pretraining corpus and differ only through the tuning step, this ratio is a valid e-value under the null hypoth-



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.

esis $H_0 : P_I = P_B$. Its product across tokens is a martingale, and Ville's inequality yields the anytime-valid bound $\Pr(\sup_t M_t \geq 1/\alpha \mid H_0) \leq \alpha$.

SCMD alone is insufficient because its signal-to-noise ratio varies across families and benchmarks, leaving usable signal unexploited. Three complementary inference-time primitives stabilize detection. Primary-model entropy separates confident divergence from uncertain divergence. Attention-Grounding Entropy (AGE) tracks how sharply the primary model attends to the answer span it is producing. Semantic Refutation via NLI (SRN) uses a frozen off-the-shelf natural language inference (NLI) model to score the generation against a fixed set of refutation templates. These primitives, and the fused classifier that combines them, are not conditional e-values. They are empirical features whose false-positive control is evaluated via split conformal.

During the experiments, we AWARD evaluate across five model families spanning three parameter scales and two architectures: Mistral-7B-Instruct (Jiang et al., 2023), Qwen-2.5-7B and Qwen-2.5-3B (Qwen et al., 2025), OLMo-7B (Groeneveld et al., 2024), and DeepSeek-V2-Lite (Liu et al., 2024), on HaluEvalQA (Li et al., 2023a), TruthfulQA, and MatSciQA (Wang et al., 2023).

We formulate an explicit base-sampling null $H_0^B$ under which SCMD is a likelihood-ratio e-process (Proposition 1, Theorem 2), and we state its non-transfer to greedy $P_I$ decoding and to factual correctness. Two primitives complementary to SCMD are introduced: Attention-Grounding Entropy (AGE) and Semantic Refutation via NLI (SRN). We evaluate AWARD under nested 5-fold cross-validation on 15 model–benchmark cells. Mean AUROC reaches $0.735/0.689/0.652$ on HaluEvalQA/MatSciQA/TruthfulQA, respectively, with peaks of $0.790/0.735/0.679$. Against the strongest of DoLa, SelfCheckGPT, and Semantic Entropy under a 10-seed paired bootstrap, the mean $\Delta = +0.089$ is positive in $14/15$ cells, and two cells survive Holm–Bonferroni family-wise $\alpha$=0.05. We further construct a classifier- and label-matched supervised baseline that fuses DoLa, SelfCheckGPT, and Semantic Entropy through AWARD's own classifier. AWARD still wins $12/15$ cells with mean $\Delta = +0.080$, indicating that the gain is not merely a supervised-versus-unsupervised artifact. On the calibration side, split-conformal thresholding delivers marginal $\mathbb{E}[\mathrm{FPR}] \leq \alpha$ under exchangeability. Empirically, at the operationally meaningful $\alpha$=0.05, $12/15$ cells hold the seed-averaged bound and $8/15$ hold an item-clustered $95\%$ upper-CI bound. Coverage loosens at higher $\alpha$. Finally, we quantify each primitive's unique value with a leave-one-block-out ablation, and a leave-one-model-family-out analysis reports a mean transfer drop of $+0.046$ AUROC, with $9/15$ cells within $0.05$.

## 2 Related Work

DoLa (Chuang et al., 2024) adaptively selects an early layer that maximizes Jensen-Shannon divergence against the final layer, using the resulting contrast as both a decoding correction and a hallucination score. Contrastive Decoding (Li et al., 2023b) employs a related expert-vs-amateur contrast. SCMD reuses the log-ratio object but treats it as a detection statistic with a formal e-value null, not as a decoding correction. SelfCheckGPT samples $K$ generations and measures inter-sample disagreement through NLI or through BERTScore (Manakul et al., 2023). Semantic Entropy samples $K$ generations, clusters them by bidirectional NLI equivalence, and reports the entropy of the cluster distribution (Kuhn et al., 2023). Both detectors incur a compute cost of $K$ forward passes per query. AWARD runs on a single greedy decode.

Recent work extracts hallucination signal from hidden-state or attention statistics: INSIDE uses eigenvalues of last-layer hidden-state covariance across samples (Chen et al., 2024). LLM-Check inspects single-pass attention and hidden-state statistics (Sriramanan et al., 2024). AGE is related to the attention component of LLM-Check but restricts entropy to the answer span rather than the full prefix. A separate line trains probes on internal states for factuality (Azaria & Mitchell, 2023), and Halo predicts hallucination from model-internal uncertainty (Elaraby et al., 2023).

Combining multiple hallucination signals through a supervised classifier is not new. (Valentin et al., 2024) concatenate eight scoring methods (perplexity, P(True) variants, NLI, verbalized probabilities, SelfCheckGPT-NLI, similarity, guardrails) and fit logistic regression on top, reporting gains on TriviaQA, FEVER, HaluEvalQA, and BIG-bench. AWARD adopts the same broad recipe but differs in three respects: (i) our primitives are a formally anchored SCMD e-process plus AGE and SRN rather than a catalog of black-box scores; (ii) we report paired-bootstrap AUROC intervals

with item-level clustering and family-wise multiplicity control, whereas (Valentin et al., 2024) give point estimates only; (iii) we explicitly delimit the scope of the guarantee: SCMD provides anytime-valid type-I error control under the base-sampling null, whereas the learned fusion carries no such guarantee.

AWARD's theoretical scaffold is game-theoretic statistics and safe anytime-valid inference (Ramdas et al., 2023; Shafer & Vovk, 2019; Waudby-Smith & Ramdas, 2022; Ville, 1939), with tighter time-uniform boundaries available through confidence sequences (Howard et al., 2021). Our composition result in Proposition 3 is standard, and we state it in the form that we need for the four-primitive fusion.

# 3 METHOD

## 3.1 PRELIMINARIES: E-VALUES AND VILLE'S INEQUALITY

On a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, a non-negative $E$ is an e-value for $H_0$ if $\mathbb{E}_{H_0}[E] \leq 1$. A non-negative process $(M_t)$ with $M_0{=}1$ is a supermartingale under $H_0$ if $\mathbb{E}_{H_0}[M_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}$. Ville's inequality (Ville, 1939) states that for any such supermartingale and $c > 0$, $\mathbb{P}(\sup_{t\geq 0} M_t \geq c) \leq 1/c$; setting $c{=}1/\alpha$ yields an anytime-valid type-I bound.

## 3.2 SPECULATIVE CROSS-MODEL DIVERGENCE (SCMD)

Let $P_I(\cdot \mid c_t)$ and $P_B(\cdot \mid c_t)$ denote the instruction-tuned and pretrained base distributions over the shared vocabulary $\mathcal{V}$ given context $c_t{=}(x, y_1, \ldots, y_{t-1})$, with $P_I \ll P_B$. The SCMD increment and product are

$$E_t^{\mathrm{SCMD}} = \frac{P_I(y_t \mid c_t)}{P_B(y_t \mid c_t)}, \quad M_t^{\mathrm{SCMD}} = \prod_{i=1}^{t} E_i^{\mathrm{SCMD}}. \tag{1}$$

The log-e-value $\log E_t^{\mathrm{SCMD}}$ is the same log-ratio used by Contrastive Decoding (Li et al., 2023b). We treat it as a detection statistic and derive its properties under an explicit null.

**Proposition 1** (SCMD under base sampling). *Under $H_0^B$: $Y_t \mid \mathcal{F}_{t-1} \sim P_B(\cdot \mid c_t)$. For every $t \geq 1$, $\mathbb{E}_{H_0^B}[E_t^{\mathrm{SCMD}} \mid \mathcal{F}_{t-1}] = 1$. Consequently $M_t^{\mathrm{SCMD}}$ is a non-negative $\mathcal{F}_t$-martingale with $M_0^{\mathrm{SCMD}} = 1$.*

*Proof.* $\sum_v P_B(v \mid c_t) \frac{P_I(v|c_t)}{P_B(v|c_t)} = \sum_v P_I(v \mid c_t) = 1$; multiplying by $M_{t-1}^{\mathrm{SCMD}}$ gives the martingale property.

**Theorem 2** (Anytime bound under base sampling). *Under $H_0^B$, for every $\alpha \in (0, 1)$, $\Pr(\sup_{t\geq 0} M_t^{\mathrm{SCMD}} \geq 1/\alpha \mid H_0^B) \leq \alpha$. The bound holds at every stopping time.*

**Scope of the guarantee:** Proposition 1 concerns the source law of the tokens. Under $P_I{=}P_B$ the ratio equals one for every token under any decoder, so the process carries no detection information. Our experiments use greedy $P_I$ decoding, which is neither sampling from $P_B$ nor a sampling scheme at all; the martingale guarantee does not transfer. The Ville bound is therefore a formal source-shift diagnostic, not a theorem about hallucination. Appendix A.1 gives the full distinction.

**Behaviour under $P_I$ sampling:** If $Y_t \mid \mathcal{F}_{t-1} \sim P_I$, then $\mathbb{E}_{P_I}[\log E_t^{\mathrm{SCMD}} \mid \mathcal{F}_{t-1}] = \mathrm{KL}(P_I(\cdot \mid c_t) \| P_B(\cdot \mid c_t))$, so the log-process drifts positively whenever the two conditional distributions differ.

**Numerical safety:** The unclipped ratio is e-valid, and upper clipping $\min(E, C)$ for any fixed $C > 0$ preserves the conditional e-value inequality. A positive floor or two-sided clipping need not preserve e-validity without an explicit normalizer. Our experiments use unclipped log-ratios.

## 3.3 COMPLEMENTARY PRIMITIVES

**Confidence-gated entropy:** Let $H_I(c_t) = -\sum_v P_I(v \mid c_t) \log P_I(v \mid c_t)$. We compute the confidence-gated log-ratio

$$G_t^{\mathrm{SCMD}} = \log E_t^{\mathrm{SCMD}} \cdot (1 - H_I(c_t)/\log |\mathcal{V}|) \tag{2}$$

plus the entropy gap $H_B - H_I$ (with $H_B$ the base-model entropy, defined analogously to $H_I$) and other entropy aggregates. These are features, not e-values.

**Attention-Grounding Entropy (AGE):** For attention weight $A_{l,h,i,j}$, at the final query position $i = t$ define

$$\mathrm{AGE}_t = \frac{1}{LH} \sum_{l,h} \frac{H(\tilde{A}^{\mathrm{ans}}_{l,h,t,:})}{\log t}, \quad t \geq 2, \tag{3}$$

where $\tilde{A}^{\mathrm{ans}}$ is the answer-span row re-normalized, $\log t$ normalizes to $[0, 1]$. Low AGE means concentrated answer-span attention. We also track peak attention and BOS-token mass, an attention-sink indicator (Xiao et al., 2024). AGE alone does not certify factual grounding.

**Semantic Refutation via NLI (SRN):** Let $g$ be the generated answer, $\mathcal{T}$ a fixed set of five refutation templates (sentences that assert the claim in $g$ is false, a misconception, unsupported or factually incorrect, i.e. hypotheses a fact-checker would use to refute $g$). Define

$$S^{\mathrm{SRN}}(g) = \max_{t \in \mathcal{T}} P_{\mathrm{NLI}}(\mathrm{entail} \mid g, t). \tag{4}$$

We use frozen `DeBERTa-v3-base-mnli-fever-anli` (He et al., 2021), trained on MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), and ANLI (Nie et al., 2020). The same NLI weights score TruthfulQA labels, §section 5 reports an independent-judge robustness check. The five templates, fixed across all configurations and benchmarks, are listed in Appendix A.3. Their individual entailment probabilities also enter the feature vector as `srn_t0`–`srn_t4`.

### 3.4 Composition and supervised fusion

**Non-e-value primitives and their consequence:** AGE and SRN are bounded but not conditional e-values without an independently estimated calibration map and a common null; we make no such construction here. The fusion is therefore not a convex combination of e-values, and its held-out threshold does not inherit Theorem 2. We report empirical operating characteristics for the fusion in §section 5.4.

**Feature vector and classifier:** Each generation is summarized by a 30-dimensional feature vector: seven SCMD aggregates, seven entropy features (including Eq. 2), five AGE attention statistics, six SRN features (one score plus five per-template entailments), three per-step KL aggregates, and two length features (content-token count and predicted length).

**Nested validation:** An outer 5-fold stratified split yields held-out predictions. Within each outer-training fold, an inner 4-fold split selects among L2 logistic regression ($C$=1), L1 logistic regression ($C$=0.3), and gradient boosting (300 trees, depth 3, learning rate 0.05) (Hastie et al., 2009; Tibshirani, 1996; Friedman, 2001). Feature standardization is fit only on the relevant inner-training or outer-training partition. This nested procedure avoids the model-selection bias that arises when the test fold is used to choose the classifier (Cawley & Talbot, 2010).

## 4 Experimental Setup

### 4.1 Large Language Models

In the experiment, we evaluate five model families. Mistral-7B pairs `Mistral-7B-Instruct-v0.2` with `Mistral-7B-v0.1` (Jiang et al., 2023). Qwen-2.5-7B pairs `Qwen2.5-7B-Instruct` with `Qwen2.5-7B`, and Qwen-2.5-3B does the same at 3B scale (Qwen et al., 2025). OLMo-7B pairs `OLMo-7B-0724-Instruct-hf` with `OLMo-7B-0724-hf` (Groeneveld et al., 2024). DeepSeek-V2-Lite pairs `DeepSeek-V2-Lite-Chat` with `DeepSeek-V2-Lite`. This last pair is a 16B mixture-of-experts model with 2.4B active parameters per forward pass (Liu et al., 2024).

| Model | Bench | DoLa | Self | SemE | BF | CBF | **AWARD** | $\Delta$ [95% CI] |
|---|---|---|---|---|---|---|---|---|
| Mistral | HaluEvalQA | .636 | .604 | .611 | .666 | .668 | **.723** | $+.088\,[+.044, +.125]^*$ |
| Mistral | MatSciQA | .566 | .577 | .528 | .565 | .505 | **.713** | $+.136\,[+.035, +.230]^*$ |
| Mistral | TruthfulQA | .574 | .577 | .505 | .520 | .620 | **.678** | $+.112\,[-.005, +.237]$ |
| Qwen-7B | HaluEvalQA | .501 | .629 | .611 | .602 | .585 | **.742** | $+.113\,[+.065, +.162]^*$ |
| Qwen-7B | MatSciQA | .569 | .505 | .517 | .555 | .503 | **.669** | $+.099\,[-.002, +.197]$ |
| Qwen-7B | TruthfulQA | .550 | .534 | .560 | .608 | .588 | **.679** | $+.141\,[-.004, +.254]$ |
| Qwen-3B | HaluEvalQA | .567 | .671 | .659 | **.723** | .678 | .712 | $+.041\,[+.001, +.085]^*$ |
| Qwen-3B | MatSciQA | .534 | .531 | .566 | .431 | .457 | **.735** | $+.169\,[+.084, +.322]^*$ |
| Qwen-3B | TruthfulQA | .599 | .515 | .542 | .470 | .455 | **.628** | $-.090\,[-.216, +.031]$ |
| OLMo | HaluEvalQA | .540 | .714 | .745 | .750 | .592 | **.790** | $+.045\,[-.007, +.081]$ |
| OLMo | MatSciQA | .558 | .538 | .523 | .498 | **.601** | .596 | $+.037\,[-.041, +.115]$ |
| OLMo | TruthfulQA | .613 | .566 | .644 | .440 | .641 | **.679** | $+.204\,[+.061, +.367]^*$ |
| DeepSeek | HaluEvalQA | .575 | .698 | .698 | **.718** | .703 | .706 | $+.005\,[-.039, +.061]$ |
| DeepSeek | MatSciQA | .578 | .524 | .601 | .482 | .668 | **.734** | $+.143\,[+.060, +.254]^*$ |
| DeepSeek | TruthfulQA | .515 | .534 | .521 | .451 | .539 | **.595** | $+.098\,[-.033, +.228]$ |
| Mean HaluEvalQA | | .564 | .663 | .665 | .692 | .645 | **.735** | |
| Mean MatSciQA | | .561 | .535 | .547 | .506 | .547 | **.689** | |
| Mean TruthfulQA | | .570 | .545 | .554 | .498 | .569 | **.652** | |

Table 1: Cross-model AUROC across 15 model–benchmark cells. AUROC columns are nested-CV out-of-fold estimates (outer 5, inner 4, seed 42), shown without leading zeros (e.g., $.723 = 0.723$). The $\Delta$ column is the 10-seed mean AWARD-minus-strongest-unsupervised gap with a paired-bootstrap 95% CI ($B$=1000), $^*$ marks unadjusted $\alpha$=0.05 significance. Discussion of the seed-42 vs 10-seed gap and Holm–Bonferroni adjustment appears in § 5.1.Columns: Self = SelfCheckGPT, SemE = Semantic Entropy, BF = supervised Baseline-Fusion, CBF = Cheap-Baseline-Fusion.

### 4.2 Inference configuration

Primary models run in `bfloat16` with eager attention so we can extract attention weights per layer. Base models run in `int8` via bitsandbytes. DeepSeek-V2-Lite requires both models in `int4` to fit on A100-40GB. We use greedy decoding with an 80-token limit. The content-token mask keeps tokens whose decoded string has length $\geq 3$ and has at least three alphabetic characters.

### 4.3 Independent-judge validation

Because SRN and the TruthfulQA labels share the DeBERTa judge, we relabel with `roberta-large-mnli` (different architecture, MNLI-only training). Item-level raw agreement is 0.81–0.90 across families (mean 0.85). Refitting the fusion on RoBERTa labels shifts TruthfulQA AUROC by mean $-0.025$, worst $-0.069$ (Qwen-2.5-7B), best $+0.026$ (Qwen-2.5-3B).

### 4.4 Benchmarks

We use $n = 1000$ question-answer items from HaluEvalQA (Li et al., 2023a), with reference correct and hallucinated answers. MatSciQA is $n = 389$ chemistry problems adapted from SciBench (Wang et al., 2023). TruthfulQA (Lin et al., 2022) contains $n = 817$ adversarial-misconception questions. We use substring matching for HaluEvalQA and a numerical parser with 5% tolerance for MatSciQA. For TruthfulQA, the DeBERTa model used by SRN assigns correctness labels from the benchmark reference answers. Across the five model families, this procedure yields 641 labeled outputs, with 114–172 retained per family.

### 4.5 Baseline Models

We compare against three retrieval-free detectors. DoLa (Chuang et al., 2024) is our self-implementation with candidate early layers $\{2, 4, 6, \ldots, 30\}$, adaptive selection by maximum Jensen-Shannon divergence, and pointwise log-odds at the greedy token. SelfCheckGPT-NLI (Manakul et al., 2023) uses $K = 3$ sampled generations at temperature 0.7, and it scores the maximum disagreement between sampled generations and the greedy answer. Semantic Entropy (Kuhn et al.,

2023; Farquhar et al., 2024) uses $K = 5$ samples, bidirectional NLI clustering at threshold 0.5 and the entropy of the cluster distribution. We also report two naive baselines of our own construction. The Mean-KL single-model detector computes the KL divergence between the primary model's final-layer distribution and a fixed early-layer projection at each content token, then takes the mean. It uses only intra-model information from the primary and does not consult the pretrained base. The Perplexity baseline is the negative mean log-probability of the greedy token under the primary model.

# 5 Results and Analysis

## 5.1 Cross-model AUROC

Table 1 reports AUROC for AWARD against three retrieval-free baselines and the two classifier- and label-matched supervised baselines across five families and three benchmarks. The paired $\Delta$ column reports the AWARD-minus-strongest-unsupervised gap with a 10-seed paired-bootstrap 95% CI (B=1000).

AWARD holds the highest AUROC in 12 of 15 cells. The three cells where a supervised baseline edges ahead are Qwen-3B/HaluEvalQA ($-0.011$ vs BF), OLMo-7B/MatSciQA ($-0.005$ vs CBF), and DeepSeek/HaluEvalQA ($-0.012$ vs BF), all within a $\pm 0.02$ AUROC band. Against the strongest unsupervised baseline, AWARD's mean $\Delta$AUROC over 15 cells is $+0.089$ under 10-seed paired bootstrap. It is positive in 14 cells and reverses on Qwen-3B/TruthfulQA. Averaging seed-42 per-cell differences within each benchmark, AWARD's mean lift over the strongest unsupervised baseline is $+0.059$ on HaluEvalQA, $+0.115$ on MatSciQA, and $+0.069$ on TruthfulQA. Under Holm–Bonferroni family-wise correction at $\alpha$=0.05 across the 15 cells (using Wald $p$-values derived from the paired-bootstrap CIs, per the caption of Table 1), two cells remain significant, both on HaluEvalQA (Qwen-2.5-7B and Mistral-7B). The seven unadjusted intervals that exclude zero should not be interpreted as individually significant after multiplicity correction.

## 5.2 Cross-family transfer

We quantify family portability by leave-one-model-family-out (LOMFO). For each held-out family we pool items from the other four, inner-4-fold select the classifier on the pool, refit, and apply to the held-out family. Appendix B gives per-cell results. Across 15 (benchmark, held-out) cells, mean LOMFO AUROC is 0.646 against within-family 0.692, a mean drop $+0.046$; $9/15$ cells transfer within 0.05 AUROC. Worst-case is OLMo-7B on TruthfulQA ($+0.123$); Qwen-7B on MatSciQA and DeepSeek on TruthfulQA show negative drops (pooled cross-family training slightly improves over the held-out family's own training data). Because the same benchmark items appear across training families, this experiment tests new-family transfer within a known benchmark distribution, not simultaneous new-family and new-question.

## 5.3 Leave-one-block-out ablation

Table 2 reports, per (model, benchmark) cell, the AUROC drop when each primitive block is removed and the inner CV re-selects the classifier. Across the 15 cells, mean drop by block is AGE $+.026$, Entropy $+.025$, SCMD $+.019$, SRN $+.017$, Length $+.017$, KL $+.005$; every block except a legacy KL block contributes on average. The dominant block varies: Entropy and SCMD each lead in five cells, AGE and SRN in two each, Length in one, KL in none. This supports the design's use of fusion over a single-primitive detector.

## 5.4 Split-conformal FPR

We evaluate split-conformal FPR under the operational null that an answer is correct. We stratify all labeled items by class into 60% training, 20% calibration, and 20% test partitions. The classifier is selected by inner 4-fold cross-validation and fit on both classes in the training partition. Only correctly answered calibration items form the conformal threshold. Writing $n_c$ for their number and $S_{(1)} \le \cdots \le S_{(n_c)}$ for their sorted classifier scores, we set $k = \lceil (n_c + 1)(1 - \alpha) \rceil$ and $\hat{q}_\alpha = S_{(k)}$, with the non-randomized convention $S_{(n_c+1)} \equiv +\infty$ when $k > n_c$. A test answer is flagged when

| Model | Benchmark | SCMD | Entropy | AGE | SRN | KL | Length | Full |
|---|---|---|---|---|---|---|---|---|
| Mistral | HaluEvalQA | +.001 | +.026 | +.010 | +.016 | −.004 | +.016 | .723 |
| Mistral | MatSciQA | +.086 | −.009 | +.013 | +.045 | +.000 | +.039 | .713 |
| Mistral | TruthfulQA | +.016 | −.004 | +.126 | −.017 | +.046 | +.015 | .678 |
| Qwen-7B | HaluEvalQA | +.010 | +.066 | +.030 | +.018 | −.006 | −.001 | .742 |
| Qwen-7B | MatSciQA | +.018 | −.006 | +.040 | −.045 | +.043 | +.049 | .669 |
| Qwen-7B | TruthfulQA | +.033 | +.027 | −.014 | −.018 | −.000 | +.021 | .678 |
| Qwen-3B | HaluEvalQA | +.002 | +.043 | +.009 | −.004 | −.003 | +.011 | .712 |
| Qwen-3B | MatSciQA | −.018 | −.013 | +.143 | +.013 | −.010 | +.040 | .735 |
| Qwen-3B | TruthfulQA | +.077 | +.045 | −.001 | +.058 | +.018 | +.015 | .628 |
| OLMo | HaluEvalQA | +.035 | +.031 | +.005 | +.017 | −.004 | +.007 | .790 |
| OLMo | MatSciQA | +.018 | +.004 | +.029 | +.042 | +.000 | +.038 | .596 |
| OLMo | TruthfulQA | −.036 | +.078 | −.002 | +.064 | −.002 | −.003 | .679 |
| DeepSeek | HaluEvalQA | −.017 | +.033 | −.005 | −.006 | −.004 | −.006 | .704 |
| DeepSeek | MatSciQA | −.005 | +.009 | −.002 | +.056 | +.001 | +.009 | .734 |
| DeepSeek | TruthfulQA | +.068 | +.048 | +.013 | +.012 | +.002 | −.001 | .592 |
| Mean | across 15 cells | +.019 | +.025 | +.026 | +.017 | +.005 | +.017 | .692 |

Table 2: Leave-one-block-out ablation. Each entry is AUROC(full) − AUROC(full − block); positive means the block contributed in that cell. Full is the nested-CV AUROC of the full 30-dim vector, computed independently within the ablation harness. It may differ from Table 1 by $\leq .003$ on three cells (DeepSeek/HaluEvalQA, DeepSeek/TruthfulQA, Qwen-7B/TruthfulQA) due to independent fold construction at seed 42.

its score exceeds $\hat{q}_\alpha$. Empirical FPR is the flagged fraction among correctly answered test items, and power is the flagged fraction among incorrect test items. Under item-level exchangeability of correct calibration and test examples, split-conformal theory gives $\Pr(S_{\text{test}} > \hat{q}_\alpha) \leq \alpha$, marginally over the calibration and test draw. This is a property of the calibration procedure; the empirical FPR distribution is a diagnostic, not a proof.

Table 3 summarizes 10 random splits for each of the 15 cells. Column 2 counts cells whose seed-averaged FPR is $\leq \alpha$. Column 3 counts cells whose item-clustered nonparametric bootstrap ($B$=1000) 95% upper CI is $\leq \alpha$ (a conservative sufficient condition for empirical control at that $\alpha$, with item-level clustering that correctly accounts for item recurrence across the 10 seeds). Per-cell correct-labeled calibration sizes are $n_c \in [9, 73]$ (minimum OLMo/TruthfulQA, maximum Mistral-7B/HaluEvalQA). At $\alpha$=0.01 we have $\lceil (n_c + 1) \cdot 0.99 \rceil > n_c$ in every cell, so the non-randomized threshold is $+\infty$ for all 15 cells and empirical FPR and power are both mechanically zero. The $\alpha$=0.01 row of Table 3 is therefore a degenerate-threshold artifact, not evidence about AWARD. Non-trivial detection is only available at $\alpha \geq 0.05$. At $\alpha$=0.05, 12/15 cells hold the mean bound and 8/15 hold the conservative upper-CI bound. Coverage is looser at higher $\alpha$. At $\alpha$=0.20 only 1/15 cell (DeepSeek/MatSciQA) has an upper CI strictly below the target, though 7/15 hold in expectation. The theoretical marginal guarantee is a statement about the calibration procedure under exchangeability, not something proven by any empirical count. DeepSeek-V2-Lite on HaluEvalQA shows the largest observed per-seed FPR at high $\alpha$. We report the excess without attributing a specific cause. Mean power at $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$ is 0.000/0.099/0.249/0.427. Appendix C reports per-cell diagnostics and the limitations of the bootstrap upper endpoints, particularly when no false positives are observed.

| $\alpha$ | Cells (mean $\leq \alpha$) | Cells (upper CI $\leq \alpha$) | Mean FPR | Mean power |
|---|---|---|---|---|
| 0.01 | 15/15 | 15/15 | 0.000 | 0.000 |
| 0.05 | 12/15 | 8/15 | 0.025 | 0.099 |
| 0.10 | 9/15 | 2/15 | 0.097 | 0.249 |
| 0.20 | 7/15 | 1/15 | 0.199 | 0.427 |

Table 3: Split-conformal FPR summary (all 15 cells, 10 seeds). Column 2: cells whose 10-seed mean FPR $\leq \alpha$. Column 3: cells whose item-clustered 95% bootstrap upper CI $\leq \alpha$ (conservative empirical sufficient condition). The $\alpha$=0.01 row is degenerate, see §section 5.4.

### 5.5 PERFORMANCE VARIATION

HaluEvalQA has mean AWARD AUROC $0.735$ and the narrowest reported paired intervals, with widths at most $0.100$ in Table 1. MatSciQA AUROCs range from $0.596$ to $0.735$; its mean AGE and Length ablation drops are approximately $0.045$ and $0.035$. TruthfulQA has the lowest mean AUROC ($0.652$), smaller retained subsets, and dependence between its NLI labels and SRN features. Qwen-3B/TruthfulQA has a positive seed-42 gap ($+0.029$) but a negative 10-seed mean gap ($-0.090$). This difference does not by itself identify a mechanism. The three seed-42 losses to supervised fusions are at most $0.012$. Model-specific observations and feature contributions appear in Appendix B.

### 5.6 COMPUTE COST

On Mistral-7B using an A100-40GB GPU under the bfloat16 primary model configuration, AWARD requires two language-model (LM) passes and five one-shot NLI queries, for a per-query latency of $1.9$ s. SelfCheckGPT ($K$=3) requires four LM passes and three NLI queries ($3.8$ s; $2.0\times$ AWARD), while Semantic Entropy ($K$=5) requires six LM passes and 20 NLI queries ($7.6$ s; $4.0\times$). DoLa uses a single LM pass with hidden-state caching across 15 layers ($2.3$ s; $1.2\times$).

## 6 DISCUSSION AND LIMITATIONS

Two complementary analyses support AWARD's empirical positioning. AWARD achieves higher paired point estimates than the strongest unsupervised baseline in $14/15$ settings (mean $\Delta = +0.089$), although only $2/15$ comparisons remain significant after Holm–Bonferroni correction at family-wise $\alpha$=0.05. It also outperforms the classifier- and label-matched supervised baseline in $12/15$ settings (mean $\Delta = +0.080$). At $\alpha$=0.05, the lowest tier at which the split-conformal threshold is non-degenerate, the mean empirical FPR does not exceed the target in $12/15$ cells, while the item-clustered upper confidence bound remains below the target in $8/15$ cells (Table 3). The LOO ablation indicates that every primitive except the legacy KL block contributes positively on average, whereas the LOMFO analysis provides evidence of partial portability across model families. OLMo achieves the highest AUROC in our evaluation ($0.790$) despite having the lowest task accuracy among the five model families. One plausible explanation is that greater divergence between OLMo's pretraining and instruction-tuning distributions yields stronger SCMD and entropy-gap signals. This interpretation is observational and should not be construed as causal.

AWARD is intended as one component of a defense-in-depth pipeline, rather than as a stand-alone decision system. Even at its most permissive operating point ($\alpha$=0.20), its mean empirical power across the 15 cells is $0.427$, meaning that more than half of the incorrect answers remain undetected. Accordingly, a below-threshold AWARD score should be interpreted as "no strong evidence of hallucination," not as evidence that an answer is correct. At $\alpha$=0.05, mean power decreases to $9.9\%$. This reflects the expected trade-off of split-conformal calibration. Under the exchangeability assumption, stronger marginal false-alarm control is obtained at the cost of additional missed detections in this operating range. Systems requiring higher recall may use a larger value of $\alpha$ while accounting for the corresponding empirical.

### 6.1 LIMITATIONS

The AWARD has six main structural limitations. (i) SCMD tests a base-sampling source null, not answer correctness, and the experimental greedy decoding does not satisfy that null. (ii) The supervised fusion and fixed-split thresholds have no anytime guarantee. (iii) TruthfulQA retains only 114–172 retained items per model and exhibits dependence between the NLI judge and NLI-derived features. The RoBERTa-MNLI robustness analysis does not eliminate this concern. (iv) MatSciQA uses a numerical parser with a $5\%$ tolerance. The item list, source mapping, and parser code are included in the supplementary archive. (v) In the three cells where AWARD does not attain the highest AUROC, its score is within $\pm 0.02$ of the strongest baseline. (vi) DeepSeek/HaluEvalQA exhibits the largest observed mean FPR at higher values of $\alpha$. This empirical excess alone does not establish a violation of exchangeability, and we therefore restrict our conclusion to the observed behavior.

## 7 CONCLUSION

AWARD combines four inference-time primitives (SCMD, primary-model entropy, Attention-Grounding Entropy, and NLI-based Semantic Refutation) into a unified, retrieval-free hallucination detector. The SCMD e-process provides an anytime-valid type-I error bound only under an explicit base-sampling null. In paired-bootstrap analyses across 15 model–benchmark cells, AWARD improves mean AUROC by $\Delta = +0.089$ relative to the strongest unsupervised baseline, with positive point estimates in $14/15$ cells and Holm–Bonferroni-adjusted significance in $2/15$. AWARD also outperforms a classifier- and label-matched supervised baseline in $12/15$ cells. At $\alpha$=0.05, the item-clustered $95\%$ upper confidence bound on empirical FPR remains at or below the target in $8/15$ cells, with mean power of $0.099$. The $\alpha$=0.01 operating tier is degenerate in every cell because the finite-sample calibration sets are too small to support a nontrivial threshold. A key direction for future work is to construct conditional e-values for an operational correctness null and combine them using predictable betting weights.

### AI USE STATEMENT

Generative AI was used to assist with language editing, grammar correction, and reorganization of material in the main text. This assistance did not rerun the experiments or independently validate the reported predictions. The authors are responsible for the final text, mathematical claims, references, and experimental results.

### ETHICS STATEMENT

AWARD can miss incorrect answers and flag correct ones. Its score should support review rather than serve as evidence of factual correctness. The evaluation does not establish safety for high-stakes deployment, and model-based labels may introduce systematic errors.

### REPRODUCIBILITY STATEMENT

Section 4 describes model configurations, generation settings, benchmarks, and baseline scoring. Section 3.4 describes classifier selection and nested validation, and Section 5.4 gives the split-conformal rule. Appendix A provides the additional theoretical discussion; Appendices B and C give transfer and calibration diagnostics.

### ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

## REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024.

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

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*, 2023.

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

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, 2024.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

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, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

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

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6449–6464, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023b.

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

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4885–4901, 2020.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Glenn Shafer and Vladimir Vovk. *Game-theoretic foundations for probability and finance*. John Wiley & Sons, 2019.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Simon Valentin, Jinmiao Fu, Gianluca Detommaso, Shaoyuan Xu, Giovanni Zappella, and Bryan Wang. Cost-effective hallucination detection for llms. *arXiv preprint arXiv:2407.21424*, 2024.

Jean Ville. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting, 2022. URL https://arxiv.org/abs/2010.09686.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

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

# A Additional theory and primitive details

This appendix expands the scope statement in Section 3.2. The predictable-mixture result below is standard and is not the fusion implemented by AWARD.

## A.1 Scope and Non-Transfer of the Theoretical Guarantee

Theorem 2 establishes an anytime-valid guarantee for the SCMD process under an explicitly specified base-sampling null. Clarifying this scope is essential because the deployed hallucination detector operates under a different data-generating mechanism and targets factual correctness rather than source-distribution consistency.

*Guarantee under the base-sampling null.* Under $H_0^B$, the next token satisfies $Y_t \mid \mathcal{F}_{t-1} \sim P_B(\cdot \mid c_t)$, and $M_t^{\mathrm{SCMD}}$ is a non-negative martingale. Consequently, Ville's inequality gives

$\Pr\left(\sup_t M_t^{\mathrm{SCMD}} \geq 1/\alpha\right) \leq \alpha$. The bound holds uniformly over time and therefore remains valid at arbitrary stopping times. Crossing the threshold provides evidence against $H_0^B$. The observed token trajectory is inconsistent with sampling from the base model. Thus, SCMD is formally a sequential source-distribution diagnostic.

*Non-transfer to factual correctness and greedy decoding.* The null $H_0^B$ does not represent the hypothesis that an answer is factually correct. In particular, if $P_I = P_B$, then every SCMD likelihood-ratio increment equals one, irrespective of whether the generated answer is correct, and the resulting process has no discriminative power. Moreover, our experiments use greedy instruction-model decoding, $Y_t = \arg\max_v P_I(v \mid c_t)$, which is deterministic conditional on $c_t$ and does not satisfy the base-sampling assumption. The SCMD increments therefore need not have conditional expectation one, and neither the martingale property nor Ville's bound applies to the generated sequences used in our experiments. Accordingly, the deployed system treats SCMD only as an empirical feature. We make no anytime-valid claim for either this feature under greedy decoding or the resulting supervised fusion.

*Why an e-process for factual correctness is not constructed.* An anytime-valid test of factual correctness would require a well-defined correctness null and conditional e-values constructed under that null. Such a construction could, for example, use a known likelihood ratio between correct and hallucinated answer distributions or a calibration procedure based on data independent of both model fitting and the evaluation stream. Our setting provides neither known class-conditional generative distributions nor an independent sequential-calibration corpus sufficient to justify such a construction. We therefore treat the fused score as the output of an ordinary supervised classifier. Its operating threshold is calibrated using split conformal prediction on a disjoint calibration set, as described in §section 5.4. This procedure targets false-positive control for factual correctness but does not convert the fused score into an e-process.

*Relation to conformal $p$-values.* Let $S_1, \ldots, S_{n_c}$ denote calibration scores from correctly answered items, where larger scores indicate stronger evidence of hallucination. For a test item $x$, define $p$-value $p(x) = (1 + \sum_i \mathbf{1}\{S_i \geq S(x)\})/(n_c + 1)$. If the calibration items and a new correctly answered test item are exchangeable, then $p(x)$ is super-uniform, and hence $\Pr(p(x) \leq \alpha) \leq \alpha$. This provides marginal false-positive control, jointly over the random calibration set and test item, and motivates the evaluation in Table 3. The conformal and Ville guarantees address different null hypotheses and rely on different assumptions. Ville's inequality provides time-uniform control under the base-sampling null, whereas split conformal prediction provides marginal control for the correctness null under exchangeability and does not generally remain valid under arbitrary optional stopping. The two guarantees are therefore complementary but should not be conflated.

### A.2 Predictable combinations of conditional e-values

**Proposition 3** (Convex combination of e-values)**.** *If $E_t^{(1)}, E_t^{(2)}$ are conditional e-values under a common null $H_0$ and $w_t \in [0,1]$ is $\mathcal{F}_{t-1}$-measurable, then $\mathbb{E}[w_t E_t^{(1)} + (1-w_t)E_t^{(2)} \mid \mathcal{F}_{t-1}] \leq 1$, and $M_t = \prod_{i \leq t}(w_i E_i^{(1)} + (1-w_i)E_i^{(2)})$ is a non-negative $H_0$-supermartingale.*

*Proof.* Linearity of conditional expectation gives $\mathbb{E}[w_t E_t^{(1)} + (1-w_t)E_t^{(2)} \mid \mathcal{F}_{t-1}] = w_t\,\mathbb{E}[E_t^{(1)} \mid \mathcal{F}_{t-1}] + (1-w_t)\,\mathbb{E}[E_t^{(2)} \mid \mathcal{F}_{t-1}] \leq w_t + (1-w_t) = 1$, using $\mathcal{F}_{t-1}$-measurability of $w_t$ to move it outside the conditional expectation and the e-value bound on each increment. For the supermartingale property, $M_{t-1} \geq 0$ is $\mathcal{F}_{t-1}$-measurable, so $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1}\,\mathbb{E}[w_t E_t^{(1)} + (1-w_t)E_t^{(2)} \mid \mathcal{F}_{t-1}] \leq M_{t-1}$; $M_0 = 1$ by convention. Ville's inequality then applies to $M_t$ and preserves the anytime type-I bound. The construction is a special case of the mixture testing framework in (Howard et al., 2021; Ramdas et al., 2023).

### A.3 Fixed SRN templates

The templates are: (1) "This claim is false."; (2) "This is a common misconception."; (3) "There is no evidence supporting this claim."; (4) "The stated claim is factually incorrect."; and (5) "This statement is a myth or a misconception." The premise is the generated answer. These generic hypotheses provide no external factual evidence.

# B TRANSFER AND MODEL-SPECIFIC RESULTS

Table 4 reports configuration holdout. Qwen-3B and Qwen-7B are separate configurations of the same family; the other Qwen configuration remains in the training pool when one is held out. The experiment also reuses benchmark questions across configurations.

| Model | Benchmark | LOMCO | Within | Drop |
|---|---|---|---|---|
| Mistral | HaluEvalQA | .686 | .723 | +.037 |
| Qwen-7B | HaluEvalQA | .641 | .742 | +.101 |
| Qwen-3B | HaluEvalQA | .655 | .712 | +.057 |
| OLMo | HaluEvalQA | .741 | .790 | +.049 |
| DeepSeek | HaluEvalQA | .696 | .706 | +.010 |
| Mistral | MatSciQA | .618 | .713 | +.095 |
| Qwen-7B | MatSciQA | .692 | .669 | −.023 |
| Qwen-3B | MatSciQA | .670 | .735 | +.065 |
| OLMo | MatSciQA | .584 | .596 | +.012 |
| DeepSeek | MatSciQA | .696 | .734 | +.038 |
| Mistral | TruthfulQA | .643 | .678 | +.035 |
| Qwen-7B | TruthfulQA | .588 | .679 | +.091 |
| Qwen-3B | TruthfulQA | .594 | .628 | +.034 |
| OLMo | TruthfulQA | .556 | .679 | +.123 |
| DeepSeek | TruthfulQA | .627 | .595 | −.032 |
| Mean drop | | | | +.046 |

Table 4: LOMCO transfer by cell. LOMCO is the AUROC obtained by training on the other four model configurations and testing on the held-out configuration; Within is the same-configuration nested-CV AUROC from Table 1. Drop is Within minus LOMCO.

## B.1 FEATURE CONTRIBUTIONS AND MODEL VARIATION

Entropy and AGE each have positive ablation drops in at least $10/15$ cells. SCMD is the largest contributor in five cells: Mistral/MatSciQA (0.086), Qwen-3B/TruthfulQA (0.077), OLMo/HaluEvalQA (0.035), DeepSeek/TruthfulQA (0.068), and Qwen-7B/TruthfulQA (0.033). On HaluEvalQA, its drops range from $-0.017$ to $0.035$. SRN's largest drop is $0.064$ for OLMo/TruthfulQA. These results describe feature contributions and do not establish tuning-divergence or refusal-rate mechanisms. KL has the smallest mean contribution ($0.005$), with five strictly positive displayed entries.

Mistral has the narrowest HaluEvalQA paired interval, $[0.044, 0.125]$. OLMo reaches the highest seed-42 AUROC ($0.790$, HaluEvalQA) and the largest TruthfulQA paired mean gain ($0.204$). Qwen-7B's TruthfulQA gain is $0.141$. DeepSeek has the smallest HaluEvalQA transfer drop ($0.010$) and the largest displayed mean FPR at $\alpha = 0.10$ and $0.20$ on that benchmark. Across all cells, the largest bootstrap upper endpoint at $\alpha = 0.05$ is Qwen-7B/TruthfulQA ($0.124$); at $0.10$ it is OLMo/TruthfulQA ($0.230$), and at $0.20$ it is Qwen-7B/MatSciQA ($0.358$).

# C CALIBRATION DIAGNOSTICS

Table 5 preserves the reported per-cell means and item-clustered bootstrap upper endpoints ($B = 1000$). The corresponding counts of endpoints at or below the target are $15/15$, $8/15$, $2/15$, and $1/15$ for $\alpha = 0.01, 0.05, 0.10, 0.20$. These are descriptive counts, not confidence certificates. Clustering repeated appearances of an item addresses one source of dependence but does not by itself establish coverage for the complete model-fitting and calibration procedure.

When all observed false-positive indicators are zero, resampling those indicators yields a degenerate bootstrap distribution. Its zero upper endpoint does not bound an unknown population FPR at zero. Separately, if $\alpha < 1/(n_c + 1)$, the threshold is structurally infinite and the detector never flags an answer. At $\alpha = 0.05$, calibration sets with fewer than 19 correctly answered items remain in this

regime. Rounded zero entries alone do not identify which mechanism applies. The table cannot replace a validated uncertainty analysis that distinguishes these cases.

| | | $\alpha$=0.05 | | $\alpha$=0.10 | | $\alpha$=0.20 | |
|---|---|---|---|---|---|---|---|
| Model | Benchmark | mean | upper | mean | upper | mean | upper |
| Mistral | HaluEvalQA | .041 | .063 | .096 | .132 | .196 | .238 |
| Mistral | MatSciQA | .000 | .000 | .060 | .113 | .180 | .261 |
| Mistral | TruthfulQA | .048 | .074 | .081 | .120 | .162 | .232 |
| Qwen-7B | HaluEvalQA | .065 | .090 | .125 | .160 | .231 | .274 |
| Qwen-7B | MatSciQA | .000 | .000 | .114 | .213 | .236 | .358 |
| Qwen-7B | TruthfulQA | .074 | .124 | .116 | .172 | .189 | .258 |
| Qwen-3B | HaluEvalQA | .050 | .074 | .085 | .118 | .203 | .251 |
| Qwen-3B | MatSciQA | .000 | .000 | .130 | .223 | .230 | .347 |
| Qwen-3B | TruthfulQA | .000 | .000 | .100 | .162 | .193 | .276 |
| OLMo | HaluEvalQA | .047 | .071 | .090 | .126 | .203 | .254 |
| OLMo | MatSciQA | .014 | .033 | .052 | .086 | .148 | .203 |
| OLMo | TruthfulQA | .000 | .000 | .144 | .230 | .200 | .286 |
| DeepSeek | HaluEvalQA | .060 | .088 | .137 | .178 | .250 | .303 |
| DeepSeek | MatSciQA | .000 | .000 | .033 | .079 | .100 | .161 |
| DeepSeek | TruthfulQA | .000 | .000 | .127 | .190 | .240 | .327 |

Table 5: Per-cell split-conformal FPR. Each $\alpha$-pair shows seed-averaged mean FPR and reported item-clustered 95% bootstrap upper endpoint ($B$=1000). Zero endpoints from all-zero resamples do not certify a zero population FPR.