SEQGUARD: LENGTH-INDEPENDENT STOPPING FOR STREAMING HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming detectors can score a language model's output for hallucinations while it is being written. The rule for *when* to interrupt is still a tuned score threshold, and a threshold has no error control that holds at every length: set to $\alpha = 0.1$ on RAGTruth, it reaches 0.19 on long outputs, and calibrated on AggreFact or TofuEval and deployed on RAGTruth it runs 1.4–$1.6\times$ over budget. We introduce SEQGUARD, a stopping rule that turns any streaming detector's scores into an anytime-valid sequential test. Each score is ranked against scores from known-faithful outputs, the rank becomes a bet, and the output is interrupted once the accumulated winnings reach $1/\alpha$. A martingale argument bounds the probability of ever interrupting a faithful output by α, for any detector and any length. The proof covers a conditionally calibrated variant that keeps 74–80% of the deployed rule's power; the simpler marginal variant we deploy is checked empirically. With a prefix-entailment scorer and a hidden-state probe, the operating configuration's realized false-alarm rate stays at or within its confidence interval of the budget in-domain (its one point estimate above budget is Data2txt at $\alpha = 0.01$, 5 of 321 episodes) and under shift among RAGTruth, AggreFact, and TofuEval, helped by slack (in-domain FAR is 0.34–0.72 of α). A far-out-of-distribution HaluEval test breaks both rules: SEQGUARD overshoots 1.2–$1.5\times$ on summarization and $2.2\times$ on dialogue, against $2.0\times$ and $5.5\times$ for the tuned threshold. Against a memoryless rule that sees the same two detectors, the power gain is concentrated on Data2txt (1.4–$1.8\times$ on three of four annotation-defined classes at matched FAR) and at $\alpha = 0.01$ on QA and Data2txt; it does not win on Summary. The cost is delay (32–73 tokens at $\alpha = 0.1$), though 72–85% of flags fire before the output ends.

1 INTRODUCTION

Language models hallucinate: they write fluent statements that their source evidence does not support (Ji et al., 2023; Alansari & Luqman, 2026). In retrieval-augmented generation and summarization, one natural fix is to watch the output as it streams and interrupt as soon as it goes wrong, then regenerate, re-retrieve, or hand off to a person. Recent work supplies the sensor: detectors that score every prefix of an output for faithfulness in real time, from prefix-tuned entailment models (Harary et al., 2025) and token-level monitors (Obeso et al., 2025; Lu et al., 2026) to fact checkers (Tang et al., 2024a), model-internal probes (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Chuang et al., 2024; Hu et al., 2026; Zhao et al., 2026), and uncertainty signals (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024).

What these systems do not supply is the decision rule: given a stream of scores $s_1, s_2, \ldots$, when should the monitor act? In practice the rule is a tuned threshold (Obeso et al., 2025; Hu et al., 2026; Zhao et al., 2026); in the simplest form, flag the first checkpoint whose score falls below 0.5. This fails in three ways. It is *uncalibrated*: on RAGTruth (Niu et al., 2024), the fixed-0.5 rule falsely interrupts 39.9% / 54.0% / 76.0% of faithful outputs across the three domains (Table 1), with no dial for a tolerable false-alarm rate. It is *length-unaware*: each checkpoint is a fresh chance to false-alarm, so the rate climbs from 0.22 on two-checkpoint outputs to 0.86 at twelve (Figure 2, right), and even a tuned threshold drifts over budget on long outputs (0.19 at $12+$ checkpoints against a budget of 0.1). It is *memoryless*: a hallucination whose evidence is spread thinly over many checkpoints cannot be caught one checkpoint at a time, whatever the threshold.

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 treat interruption as a sequential test and solve it with e-values and test supermartingales (Shafer, 2021; Vovk & Wang, 2021; Grünwald et al., 2024; Ramdas et al., 2023). Our rule, SEQGUARD, is a betting game that runs once per output. Each score is ranked against a table of scores from known-*faithful* outputs, which turns it into a p-value: how unusual would this score be if the output were faithful (Vovk et al., 2005)? A betting function fixed in advance turns the p-value into a payoff, the payoffs multiply into a running wealth, and the output is interrupted the first time wealth reaches $1/\alpha$. Ville's inequality then bounds the probability of ever interrupting a faithful output by α (Theorem 1), at every stopping time and any length. Because all test outputs share one finite calibration table, we also prove a train-conditional version (Theorem 2) and stress-test it by resampling the calibration set. Detectors fuse by averaging their wealth, the wealth values double as e-values for span-level false-discovery control (Wang & Ramdas, 2022) (Proposition 1), and a delay bound shows the rule is not trivially conservative (Theorem 3).

Contributions. **(1)** We cast stopping for streaming hallucination detection as anytime-valid sequential testing and give a detector-agnostic rule with a finite-sample, distribution-free, length-independent false-alarm guarantee for its conditionally calibrated form, plus a train-conditional version for the shared calibration table; the marginal form we deploy is checked empirically (§3). **(2)** The guarantee holds where tuning breaks: the realized false-alarm rate tracks its budget across three RAGTruth domains (0 violations over 120 calibration resamples), across output lengths, and under cross-benchmark shift, where source-tuned thresholds run 1.4–$1.6\times$ over budget; a deliberate far-OOD test breaks both rules, tuning $2.5\times$ harder (§5). **(3)** Against memoryless rules that see the same detectors, accumulation helps in two places and not elsewhere: at $\alpha = 0.01$ on QA and Data2txt, and on Data2txt's annotation-defined classes (1.4–$1.8\times$ at matched FAR). A score-defined split shows an apparent $5.4\times$ gain; it does not survive an annotation-defined control (§5.2). We also compare against CUSUM and Bonferroni, report a negative result (log-probabilities carry no onset signal), and account for interrupt-and-regenerate token savings.

2 RELATED WORK

Streaming faithfulness scoring. PrefixNLI (Harary et al., 2025) defines prefix-level entailment and trains MiniTruePrefixes, our primary detector. Other streaming detectors read tokens (Obeso et al., 2025), long chain-of-thought prefixes (Lu et al., 2026), reasoning graphs, or denoising trajectories (Hu et al., 2026; Zhao et al., 2026); Alansari & Luqman (2026) survey the area. In all of these the intervention is a tuned cutoff with no stopping-time false-alarm guarantee. RAG hallucination risk also varies with domain, retrieval, and pipeline design (Hindi et al., 2025; Wołk, 2025; Papageorgiou et al., 2025; Ngartera et al., 2026; Zhang et al., 2026), so a threshold tuned in one condition need not hold in another.

Conformal and risk-controlling generation. Conformal factuality (Mohri & Hashimoto, 2024), conformal language modeling (Quach et al., 2024), and risk-control frameworks (Angelopoulos et al., 2021; 2024) give guarantees for static objects: output sets, claim removal, batch risk. Li et al. (2025) cast hallucination detection as hypothesis testing with conformal p-values, but decide once per response at a fixed horizon. Dubey & Huo (2026) pair conformal calibration with a betting e-process for anytime-valid alarms in federated RAG, but monitor a model swarm rather than stopping one output. Neither gives a stopping time inside a growing sequence, where per-position exchangeability fails and martingale arguments replace exchangeability arguments (Barber et al., 2023).

E-values and sequential testing. We assemble standard pieces: test supermartingales and Ville's inequality (Ville, 1939; Howard et al., 2021), testing by betting (Shafer, 2021; Waudby-Smith & Ramdas, 2024), e-value calibration and combination (Vovk & Wang, 2021; Grünwald et al., 2024; Ramdas et al., 2023), e-BH (Wang & Ramdas, 2022), change detection (Page, 1954; Lorden, 1971; Shekhar & Ramdas, 2023), anytime-valid inference (Sokolova & Sokolov, 2026; Polson et al., 2026; Brannath & Fischer, 2026), and recent e-BH improvements that carry over to our span-level layer (Xu et al., 2025; Goeman et al., 2025; Ignatiadis et al., 2026; Huang & Goretzko, 2026; Sun et al., 2025). Two classical routes do not fit. Wald's SPRT (Wald, 1945) needs likelihoods under both hypotheses; here the faithful law is a finite conformal table and the hallucinated law is unknown. Group-sequential α-spending (Pocock, 1977; O'Brien & Fleming, 1979; Lan & DeMets, 1983) needs a pre-specified maximum horizon and look schedule; a streaming output's length is a random

stopping time. Test supermartingales need neither. Bonferroni over a horizon (B4) is our stand-in for the spending family. Our contribution is the assembly and its validation in this setting.

3 METHOD

3.1 SETTING

An *episode* is one model output, monitored from its first token. As it streams, a checkpointing scheme emits decision points $t = 1, 2, \ldots, T$; we use sentence boundaries and test a per-token clock in Appendix E. At each checkpoint a detector gives a score $s_t \in \mathbb{R}$, higher meaning more faithful. The monitor decides online whether to interrupt, cannot take the decision back, and looks after every checkpoint (optional stopping). The null hypothesis H_0 is that the episode is *faithful*: no hallucinated content at any checkpoint. The *false-alarm rate* $\mathrm{FAR} = \mathbb{P}_{H_0}(\tau < \infty)$ is the probability that a faithful episode is ever flagged, where τ is the (possibly infinite) flag time. We want $\mathrm{FAR} \le \alpha$ for a user-chosen budget α at every length, with high power (flagging hallucinated episodes at or after onset) and small delay.

We control per-episode FAR, not per-checkpoint error, because that is what an operator sees. On hallucinated episodes, a flag at or after the first hallucinated checkpoint t^* is a *detection*; a flag before t^* is a *premature flag*, a lucky false alarm, reported separately. Delay is counted in generated tokens.

3.2 FROM SCORES TO A CERTIFICATE

Conformal p-values. Offline, we score a calibration set of faithful episodes with the same detector and pool the scores into position buckets $\{1\}, \{2\}, \{3\}, \{4, 5\}, \{6\text{–}8\}, \{9\text{–}12\}, \{13+\}$; bucket $c(t)$ holds the calibration scores from positions in the same bucket as t. Online, each score is ranked against its bucket:

$$p_t \;=\; \frac{1 + \#\{s \in \widehat{F}_{c(t)} : s \le s_t\}}{1 + |\widehat{F}_{c(t)}|}. \tag{1}$$

In words, p_t is the fraction of faithful calibration scores at this position that are at least as low as the new score, so it is small exactly when s_t looks bad *for a faithful output at that position*. Bucketing by position absorbs the main nonstationarity (early checkpoints score differently from late ones) without assuming exchangeability across positions. If a bucket's calibration scores are i.i.d. draws from the faithful law at that position, p_t is super-uniform: $\mathbb{P}(p_t \le u) \le u$. This assumes the positions in a bucket share one faithful law, and it is only approximate because one episode contributes several dependent scores to a composite bucket; keeping one sampled checkpoint per episode per bucket removes that dependence and moves FAR by at most .003 and power by at most .02 at $\alpha = 0.1$ (Appendix Table 5).

E-values: bets against faithfulness. An *e-value* is the payoff of a fair bet against the null: a nonnegative quantity with expected value at most 1 on faithful text, so a large value is evidence of hallucination. Any $e : [0, 1] \to [0, \infty)$ with $\int_0^1 e(p)\, dp \le 1$ turns a super-uniform p-value into an e-value (Vovk & Wang, 2021). We use the power-mixture calibrator over a fixed grid $\mathcal{K} = \{0.1, 0.2, \ldots, 0.9\}$:

$$e(p) \;=\; \frac{1}{|\mathcal{K}|} \sum_{\kappa \in \mathcal{K}} \kappa\, p^{\kappa - 1}. \tag{2}$$

Each term is a valid calibrator and mixtures of e-values are e-values. The calibrator is *predictable* (the grid is fixed in advance, so nothing is tuned on test data) and *unbounded* as $p \to 0$, so one damning checkpoint can contribute a large factor. The linear bet $e(p) = 1 + \lambda(1 - 2p)$ is capped below 2 per checkpoint at $\lambda = 1$, so reaching $1/\alpha = 100$ needs about 6.6 checkpoints of perfect evidence while the median RAGTruth episode has 5; the mixture lifts power at $\alpha = 0.01$ from 0.00–0.02 to 0.11–0.18 (Summary/QA, NLI detector; Table 5) by removing that ceiling.

Wealth, stopping, and fusion. Bets compound: each detector $j \in \{1, \ldots, m\}$ reinvests its winnings, so its wealth after t checkpoints is $W_t^{(j)} = \prod_{u \le t} e(p_u^{(j)})$. Many mildly suspicious checkpoints can therefore add up to a flag, which a memoryless threshold cannot do. The flag statistic is

the average wealth,

$$W_t = \frac{1}{m}\sum_{j=1}^{m} W_t^{(j)}, \qquad \tau = \inf\{t : W_t \geq 1/\alpha\}. \tag{3}$$

We call this *wealth-mix* fusion. A convex combination of supermartingales is a supermartingale, so validity is unaffected, and a useless detector ($e \approx 1$ throughout) costs at most one factor of m in total. Averaging e-values at each step instead pays that factor at every step and erases most of the fusion gain (Table 5).

3.3 Anytime validity, and the finite calibration table

Let $(\mathcal{F}_t)$ be the filtration generated by the scores and predictable side information such as position.

Assumption 1 (Oracle conditional super-uniformity)**.** *For every t, $\mathbb{P}(p_t \leq u \mid \mathcal{F}_{t-1}) \leq u$ for all $u \in [0,1]$ under H_0.*

Theorem 1 (Anytime-valid false-alarm control)**.** *Under Assumption 1, for any predictable, non-increasing calibrator e with $\int_0^1 e(p)\,dp \leq 1$ and any fusion rule that is a convex combination of per-detector wealths, $(W_t)_{t\geq 0}$ is a test supermartingale under H_0, and for every $\alpha \in (0,1)$,*

$$\mathbb{P}_{H_0}(\tau < \infty) \;=\; \mathbb{P}_{H_0}\Big(\sup_t W_t \geq 1/\alpha\Big) \;\leq\; \alpha.$$

The bound holds at every data-dependent stopping time and for every (possibly unbounded) horizon.

The proof (Appendix A) is classical: super-uniformity plus $\int e \leq 1$ gives $\mathbb{E}[e(p_t) \mid \mathcal{F}_{t-1}] \leq 1$, products and convex combinations keep the supermartingale property, and Ville's inequality (Ville, 1939; Howard et al., 2021) bounds the crossing probability. The payoff over per-checkpoint corrections is length-independence: there is no union bound over positions, so nothing degrades as outputs grow (Figure 2) and the horizon T need not be known.

Assumption 1 idealizes two things. First, scores are autocorrelated within an episode, so the marginal position-bucketed table need not match the conditional law of s_t given the past; the conditional-calibration variant (§3.4) targets this. Second, the calibration table is one finite draw shared by all test episodes: given an unlucky table, p-values can be slightly optimistic, the optimism compounds over steps, and realized FARs are correlated across episodes. This is the sharpest objection to conformal-then-bet constructions, so we answer it directly:

Theorem 2 (Train-conditional validity)**.** *Fix $\alpha, \delta \in (0,1)$. Let $\widehat{F}$ be drawn i.i.d. from the faithful score law with n_{cell} points in each calibration cell, and let test episodes be independent of $\widehat{F}$. Run the e-process on inflated p-values $p_t^\dagger = \min\{1, p_t + \epsilon_n\}$ with $\epsilon_n = \sqrt{\log(2N_{\text{cells}}/\delta)/(2n_{\text{cell}})}$. Then, with probability at least $1-\delta$ over the calibration draw, $\mathrm{FAR}(\widehat{F}) = \mathbb{P}(\tau < \infty \mid \widehat{F}) \leq \alpha$, at every stopping time and for episodes of any length.*

The proof (Appendix A) applies the DKW inequality (Massart, 1990) in each cell, a union bound over cells, and then Theorem 1 for almost every table. The un-inflated rule we deploy has only a weaker statement: for a bounded calibrator (the linear bet, or the power mixture truncated at a fixed u_0) its FAR is at most $\alpha + \eta_n$ on episodes of at most $T_{\max}$ checkpoints with $\eta_n \to 0$; for the unbounded power mixture the DKW route gives nothing, since $e(u_{\min})$ grows like $n_{\text{cell}}^{0.9}$ while ϵ_n shrinks like $n_{\text{cell}}^{-1/2}$ (Appendix A). The one-checkpoint-per-episode table is the variant with i.i.d. cells; the resample study is the empirical companion. The inflation costs at most .019 power at $\alpha = 0.1$ on the fused rule and pulls the Data2txt $\alpha = 0.01$ point estimate from .016 to .012 (Appendix Table 6); the main tables report the un-inflated rule. Over 20 calibration resamples $\times$ two budgets $\times$ three domains, realized FAR never exceeded α ($0/120$; Appendix Table 8).

SeqGuard is not trivially conservative: if each post-onset bet has expected growth $\gamma > 1$, delay grows only logarithmically in the budget, plus the wealth a faithful prefix lost before onset.

Theorem 3 (Delay bound)**.** *Suppose Theorem 1's conditions hold and there exist $\gamma > 1$ and onset t^* such that $\mathbb{E}[\log e(p_t) \mid \mathcal{F}_{t-1}] \geq \log\gamma$ for all $t \geq t^*$ on hallucinated episodes. Then the detection delay $\Delta = \tau - t^*$ satisfies $\mathbb{E}[\Delta \mid \mathcal{F}_{t^*-1}, \tau < \infty] \leq \frac{\log(1/\alpha) - \log W_{t^*-1} + B}{\log\gamma}$, where B bounds the overshoot at the crossing step.*

The term $-\log W_{t^*-1}$ is the drawdown a faithful prefix accumulates before onset; wealth has no floor, so it grows with t^*. On RAGTruth its median at the median onset is 0.5–0.9 nats and its 90th percentile at late onsets is 2.4–5.7 nats, above $\log(1/\alpha) = 2.3$ at $\alpha = 0.1$ (Appendix A).

The same wealth values give corpus-level guarantees. For null hypotheses $H_{0,1}, \ldots, H_{0,n}$ ($H_{0,i}$: "span i is faithful"), let $E_i = W_{t_i^{\text{end}}}/W_{t_i^{\text{start}}-1}$ be the wealth earned inside span i, the product of the e-values at the checkpoints the span covers. Given the past before the span, this is a product of conditional e-values, so $\mathbb{E}[E_i] \leq 1$ under $H_{0,i}$ whatever happened earlier in the episode; the prefix wealth $W_{t_i^{\text{end}}}$ is instead an e-value for the intersection null "no hallucination up to t_i^{end}". A rejection declares span i hallucinated, and e-BH (Wang & Ramdas, 2022) controls FDR under arbitrary dependence:

Proposition 1 (Span-level FDR). *If each E_i is an e-value under its null, the e-BH procedure (reject the R largest where $R = \max\{r : E_{(r)} \geq n/(\alpha r)\}$) has* $\text{FDR} \leq \alpha$.

Corollary 1 (Interrupt-and-regenerate accounting). *Under Theorem 1, the expected tokens wasted by falsely aborting faithful episodes is at most $\alpha \cdot T_{\max}$ per episode (and at most $\alpha \cdot \mathbb{E}[T]$ when flag and length are uncorrelated); under Theorem 3, the expected tokens saved on a detected hallucination is at least $\mathbb{E}[T - t^*]_+ - \left(\frac{\log(1/\alpha) - \log W_{t^*-1} + B}{\log \gamma}\right) \bar{L}$, with $\bar{L}$ the mean tokens per checkpoint (25 / 25 / 23 on Summary / QA / Data2txt). Net savings therefore depend on the hallucination base rate (§5.3).*

3.4 Conditional calibration

Within an episode, a low score at checkpoint $t-1$ makes a low score at t more likely even on faithful text. A marginal table can then let dependence "mint" wealth: several correlated, moderately low scores each look surprising on their own but not given the past. The fix is to condition the calibration cell $c(t)$ on predictable context: we bucket on (position) $\times$ (previous score), with previous-score bins $[0, 0.5), [0.5, 0.9), [0.9, 1]$, falling back to the position-only table in cells with fewer than 50 points. In our experiments the marginal table already holds the budget, and the conditional table is more conservative (fused FAR .013–.037 at $\alpha = 0.1$) at a large power cost (.225 $\rightarrow$ .152 on Summary, .381 $\rightarrow$ .231 on QA, .420 $\rightarrow$ .230 on Data2txt; Appendix Table 7). Part of that cost is structural: a 50-point cell cannot give $p < 1/51$, so its largest payoff is $e(1/51) \approx 4.6$ and crossing $1/\alpha = 100$ takes four maximally bad checkpoints, where a marginal cell allows $p \approx 0.0003$ and a single-step crossing. Theorem 1 is a proof under Assumption 1. The conditional table targets it, but its thin cells fell back to the marginal table, which is where the guarantee leaked. Two changes close the leak: a thin cell makes no bet ($e_t = 1$), so every bet is against its own conditional table, and the previous-score split sits at each detector's calibration median. This certified rule has power .181 / .281 / .326 at $\alpha = 0.1$ against .225 / .381 / .420 for marginal (Appendix Table 7). We deploy the marginal table for that power, with three checks that dependence is not minting wealth: FAR is under budget at every length bin (Figure 2, right); the token clock holds every budget with 150–500 dependent looks (Appendix E); and with autocorrelation dialled into the p-value stream, marginal calibration holds up to $\rho = 0.9$ on every domain and crosses only on Data2txt at $\rho = 0.95$, while the real lag-1 correlation is 0.56–0.70 (Appendix J).

4 Experimental setup

Datasets. Our main benchmark is RAGTruth (Niu et al., 2024): 17,790 responses from six generators over news summarization (*Summary*), retrieval-augmented QA (*QA*), and data-to-text (*Data2txt*), with human-annotated hallucination spans. Span starts give the onset checkpoint t^*; the median onset is at 37% of response length. We calibrate on all faithful train-split episodes per domain (3,275 / 3,346 / 1,623) and evaluate on the official test split (900 / 875 / 900 episodes). For shift we use AggreFact (Tang et al., 2023) and TofuEval MeetingBank (Tang et al., 2024b) in both directions. For breadth we use HaluEval (Li et al., 2023) summarization/QA/dialogue (400 pairs each), calibrated on RAGTruth Summary with no adaptation.

Detectors. (i) *NLI*: MiniTruePrefixes (Harary et al., 2025), a 1B-parameter prefix-entailment model built for this setting; the score is P(prefix entailed by source). Our harness matches the authors' released inference stack to within 0.1 F1 (Appendix K). (ii) *Probe*: logistic regression on layer-16 hidden states of Llama-2-7b-chat (Touvron et al., 2023) at each checkpoint's last token, trained on a

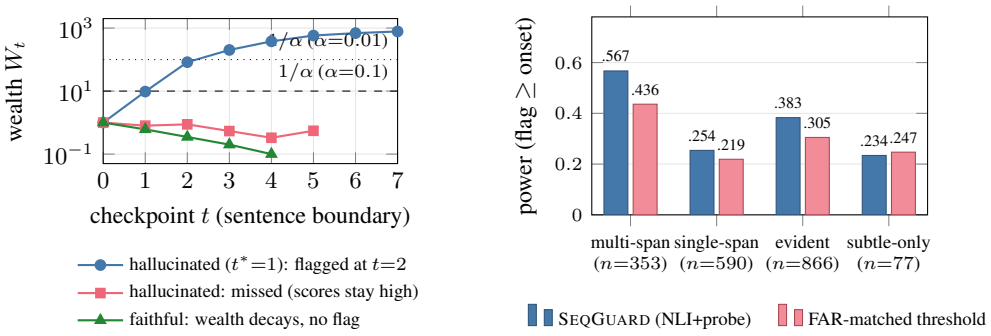

Figure 1: **Left: the mechanism.** Wealth paths of three RAGTruth Summary test episodes. On the hallucinated episode no single checkpoint is decisive at $\alpha = 0.01$, but compounded evidence crosses both budgets; on faithful text wealth decays. **Right: where accumulation pays.** Power at $\alpha = 0.1$ by annotation-defined class, pooled across domains, against a fused-Fisher threshold matched to SeqGuard's realized FAR. Pooled gains are 0.9–1.3× and are weighted toward Data2txt, which holds 61% of the hallucinated test instances; on Summary and QA the threshold is ahead in most cells (Table 4).

source-disjoint half of the RAGTruth train split; the score is P(faithful). RAGTruth responses come from several generators, so the probe reads a re-encoding rather than the generator's own states, which understates an in-situ probe. Each detector's conformal table uses only faithful episodes the detector was not trained on. (iii) A *negative result*: teacher-forced mean log-probability carries almost no onset signal on RAGTruth (power ≤ 0.05 at $\alpha = 0.1$), from a Qwen2.5-1.5B proxy (Qwen Team, 2024) or from Llama-2-7b-chat itself (Appendix F). The operating configuration is NLI + probe with wealth-mix fusion, marginal calibration, and the power-mixture calibrator; nothing is tuned on test data.

Baselines. B1: fixed threshold 0.5, the de facto standard. **B2**: a threshold tuned on our calibration set to target each α, the strongest memoryless competitor. **B3**: CUSUM (Page, 1954; Lorden, 1971) on score deficits, drift tuned on held-out hallucinated train episodes and alarm level calibrated on faithful episodes at each α. **B4**: our conformal p-values with a Bonferroni correction over checkpoints (flag if $p_t \leq \alpha/T$). B4 is given the realized length T, which a streaming monitor does not know; a causal fixed-$T_{\max}$ version is in §5.3. All methods consume the same cached scores, so comparisons are internal by construction.

Metrics. Realized FAR with exact Clopper–Pearson 95% intervals (Clopper & Pearson, 1934); power conditional on flag ≥ onset, within each domain's hallucinated episodes; premature-flag rate; mean delay in tokens. Onsets and spans differ across domains and Data2txt is majority-hallucinated (579/900), so power is not comparable across domains; FAR is.

5 Results

5.1 Validity: the certificate holds where tuning breaks

In-domain. The fixed threshold (Table 1) falsely interrupts 40–76% of faithful outputs, and 33–55% of its "detections" fire before onset: lucky false alarms. SeqGuard holds every budget at $\alpha = 0.1$ (realized FAR .034–.072). At $\alpha = 0.01$ its FAR is .007–.016; the one point estimate above budget (Data2txt, 5 of 321 faithful episodes, CI $[.005, .036]$) is finer than one split can resolve. Figure 2 (left) shows the full α range. The resample study supports Theorem 2: 0/120 violations (Appendix Table 8).

Length. Figure 2 (right) isolates the length effect predicted in §3.3. Every per-checkpoint rule spends budget with length: B1 climbs from 0.22 to 0.86, and B2, tuned to $\alpha = 0.1$ in-domain, reaches 0.11 by 8–11 checkpoints and 0.19 at 12+. The e-process's ever-flagged rate also rises with length (.017 to .092) but stays under budget at every bin. No tuning fixes this, and the direct fix, a union bound, costs power (§5.3).

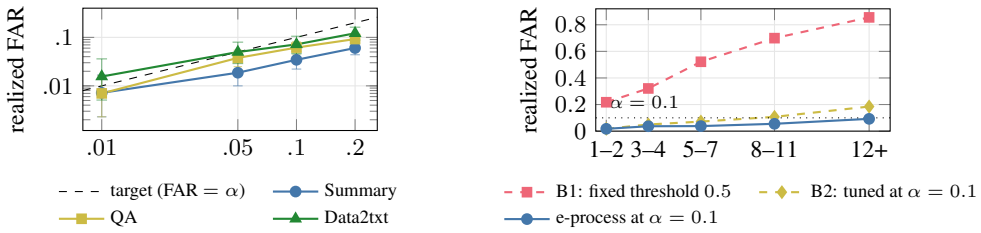

Figure 2: **Left: validity in-domain.** Realized FAR vs. target α for the operating configuration, with Clopper–Pearson 95% intervals. Every point is on or under the diagonal except Data2txt at $\alpha = 0.01$ (.016; 5 of 321 episodes), and none collapse to zero. **Right: length-independence.** Realized FAR vs. episode length on pooled faithful test episodes ($n = 115$–738 per bin). The fixed threshold degrades toward certainty; the tuned threshold crosses its budget beyond ~ 8 checkpoints; the e-process's ever-flagged rate rises with length but stays under budget at every bin.

Table 1: Main results on RAGTruth (official test split; calibration on all faithful train episodes). FAR with Clopper–Pearson 95% intervals; power conditional on flag $\geq$ onset; premature-flag rate and mean delay in tokens at $\alpha = 0.1$. B2 is run on the NLI score, the probe score, and their Fisher combination, so the memoryless rule sees the same detectors as SeqGuard. B2 on the probe at $\alpha = 0.01$ tunes to a threshold below every faithful calibration score, so it never flags. Nothing is tuned on test.

	$\alpha = 0.01$		$\alpha = 0.1$			
Method	FAR	power	FAR	power	prem.	delay
Summary (696 faithful / 204 hallucinated test episodes)						
B1: fixed threshold 0.5	FAR .399 [.363, .437] at any α			.426	.333	9
B2: threshold (NLI)	.009 [.003, .019]	.103	.085 [.065, .108]	.338	.074	10
B2: threshold (probe)	.000 [.000, .005]	.000	.111 [.088, .136]	.245	.044	32
B2: threshold (fused Fisher)	.014 [.007, .026]	.137	.085 [.065, .108]	.387	.049	8
B3: CUSUM (calibrated h)	.010 [.004, .021]	.074	.093 [.073, .117]	.324	.054	28
B4: conformal + Bonferroni	.006 [.002, .015]	.113	.057 [.041, .077]	.279	.049	8
SeqGuard (NLI only)	.009 [.003, .019]	.108	.039 [.026, .056]	.230	.020	29
SeqGuard (probe only)	.003 [.000, .010]	.020	.022 [.012, .035]	.093	.010	43
SeqGuard (NLI+probe, wealth mix)	.007 [.002, .017]	.103	.034 [.022, .051]	.225	.020	32
QA (715 faithful / 160 hallucinated test episodes)						
B1: fixed threshold 0.5	FAR .540 [.502, .577] at any α			.519	.400	11
B2: threshold (NLI)	.008 [.003, .018]	.131	.083 [.063, .105]	.394	.062	17
B2: threshold (probe)	.000 [.000, .005]	.000	.113 [.091, .139]	.406	.056	39
B2: threshold (fused Fisher)	.011 [.005, .022]	.188	.103 [.082, .128]	.469	.075	15
B3: CUSUM (calibrated h)	.001 [.000, .008]	.056	.081 [.062, .104]	.381	.044	54
B4: conformal + Bonferroni	.010 [.004, .020]	.113	.097 [.076, .120]	.281	.044	9
SeqGuard (NLI only)	.003 [.000, .010]	.175	.049 [.034, .067]	.275	.025	34
SeqGuard (probe only)	.007 [.002, .016]	.119	.055 [.039, .074]	.244	.037	52
SeqGuard (NLI+probe, wealth mix)	.007 [.002, .016]	.206	.062 [.045, .082]	.381	.037	46
Data2txt (321 faithful / 579 hallucinated test episodes)						
B1: fixed threshold 0.5	FAR .760 [.710, .806] at any α			.320	.553	19
B2: threshold (NLI)	.000 [.000, .011]	.002	.072 [.046, .106]	.067	.012	16
B2: threshold (probe)	.012 [.003, .032]	.104	.103 [.072, .141]	.456	.033	63
B2: threshold (fused Fisher)	.016 [.005, .036]	.111	.103 [.072, .141]	.351	.069	46
B3: CUSUM (calibrated h)	.009 [.002, .027]	.031	.072 [.046, .106]	.140	.005	91
B4: conformal + Bonferroni	.003 [.000, .017]	.000	.050 [.029, .080]	.078	.024	39
SeqGuard (NLI only)	.012 [.003, .032]	.017	.050 [.029, .080]	.066	.009	73
SeqGuard (probe only)	.022 [.009, .044]	.247	.072 [.046, .106]	.444	.038	69
SeqGuard (NLI+probe, wealth mix)	.016 [.005, .036]	.206	.072 [.046, .106]	.420	.019	73

Shift. Appendix Table 3 tests transfer between RAGTruth Summary and external summarization benchmarks with no recalibration. Forward transfer passes for both methods; the reverse direction is the story. Calibrated on AggreFact or TofuEval and deployed on RAGTruth, B2 exceeds its budget

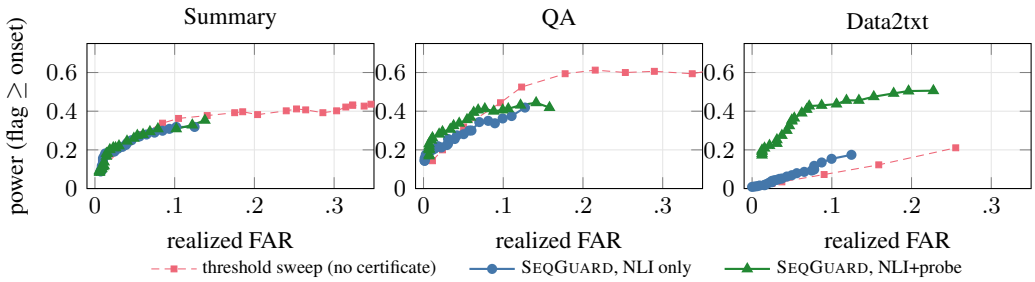

Figure 3: **Power vs. realized FAR.** Threshold sweep (red) vs. SeqGuard with NLI alone (blue) and with the NLI+probe wealth mix (green). At strict budgets (FAR $\lesssim .05$) the fused e-process matches or beats the sweep on QA and dominates on Data2txt; at loose budgets on Summary/QA the sweep is higher. Only the e-process curves carry a guarantee; the sweep uses test-set knowledge.

at every level, by 1.4–1.6× at $\alpha \in \{0.1, 0.2\}$ (realized FAR .141 at $\alpha = 0.1$; .303–.323 at $\alpha = 0.2$). SeqGuard stays under every budget in both directions (.014–.063 in the reverse direction). Part of that robustness is slack: in-domain fused FAR at $\alpha = 0.1$ is 0.34–0.72 of the budget, so a 1.4× inflation still lands under α; a rule tuned tight would not have that room. HaluEval (Li et al., 2023), calibrated on RAGTruth Summary with no adaptation, is a far-out-of-distribution test and both rules break: B2 by 5.5× on dialogue and 2.0× on summarization, the e-process by 2.2× and 1.2–1.5× (Appendix Table 12). The e-process fails soft, tuning fails hard.

5.2 Power: accumulation beats memoryless rules where it should

The strict-budget regime. A guarantee matters most at small α. Giving the threshold the same two detectors (B2 on the probe; B2 on the fused Fisher score) does not erase the $\alpha = 0.01$ gap on QA and Data2txt: fused SeqGuard has power .206 on both, against .188 and .111 for the fused Fisher threshold, at matched or lower FAR (Table 1). It does *not* win on Summary (.103 vs. .137). At loose budgets the fused threshold's frontier is higher (Figure 3), the cost of the guarantee.

Where accumulation pays. A score-defined split ("distributed" if no post-onset checkpoint has $e \geq 10$) gives a 5.4× gain over an NLI-only FAR-matched threshold, but it is close to circular: it defines the class by the threshold's failure. With classes defined from the annotations instead (multi- vs. single-span, evident vs. subtle) and the threshold matched on the fused Fisher stream, the pooled ratios fall to 0.9–1.3× (Figure 1, right; Table 4 in Appendix C). The pooled ratio leans on Data2txt (61% of hallucinated test instances); on Summary and QA the threshold is ahead or level in seven of eight class cells. The gain that survives is on Data2txt: 1.8× on single-span, 1.5× on evident, 1.4× on multi-span; on Summary and QA the two rules are within 1.1× or the threshold is ahead. Delay is 32–73 tokens at $\alpha = 0.1$; 72–85% of flags fire before the output ends, but only 22–51% before the annotated span ends.

Fusion. The probe transforms Data2txt (.066 to .420 at $\alpha = 0.1$), lifts QA (.275 to .381) and leaves Summary unchanged (.230 to .225); per-step e-averaging dilutes (Data2txt .306 vs. .420; Table 5). Part of the Data2txt margin over the fused Fisher threshold (.420 vs. .351) is fusion robustness, not accumulation: SeqGuard on the probe alone reaches .444 against .456 for the probe-alone threshold, while Fisher's combination lets the uninformative NLI stream dilute the probe.

5.3 Sequential baselines, ablations, and systems payoff

CUSUM (B3) is competitive at loose budgets (Summary .324, QA .381 at $\alpha = 0.1$) but has no distribution-free guarantee (realized FAR .081–.093), collapses at strict budgets, and its delays are as long or longer than ours (28–91 vs. 32–73 tokens). **Bonferroni over conformal p-values (B4)** is no straw man at sentence horizons ($T \approx 5$): its power (.279/.281/.078 at $\alpha = 0.1$) is respectable. Given the realized length T, it fails at scale: on a per-token clock ($T \approx 150$–500) its realized FAR is .103/.097/.131 at $\alpha = 0.1$, at budget everywhere and over it on two of three domains, while the token-clock e-process holds the budget (.035–.087) on the same inputs (Appendix Table 9). The union bound needs no independence; its input fails. The bar $\alpha/T \approx 2 \times 10^{-4}$ is a quantile that a

table of 1,600–3,300 scores cannot resolve (its smallest p-value is 3–6 $\times 10^{-4}$), and T is random. Made causal with a fixed $T_{\max}$, Bonferroni is valid but weaker: on the sentence clock ($T_{\max} = 10$ / 14 / 15) FAR is .033/.038/.025 and power .216/.231/.050; on the token clock ($T_{\max} = 500$) FAR is .024–.065 and power $\leq$ .025 (Appendix E).

Ablations. Table 5 (Appendix C): validity is insensitive to calibration size ($n = 100$ episodes is on-budget; power saturates by $n \approx 250$–500); one checkpoint per episode per bucket (i.i.d. cells) moves FAR by $\leq .003$ and power by $\leq .02$; marginal calibration wins power while the certified conditional rule keeps 74–80% of it (Appendix Table 7); shifting every onset label by ± 1 checkpoint moves the symmetric-mean power by at most .04 (a one-sided $+1$ shift by up to .09), and FAR uses no onset labels (Appendix Table 13).

e-BH on spans. With the within-span wealth ratio as each span's e-value (fused detectors), e-BH realizes FDR 0 at $\alpha \in \{.05, .1\}$ on every domain, where the uncorrected rule (reject if $E_i \geq 1/\alpha$) realizes up to .05. Power is small (.004 / .021 / .000 at $\alpha = 0.1$): almost every annotated span covers one checkpoint (1,050 of 1,054 on Data2txt), so E_i is a single e-value and cannot compound. Using prefix wealth instead, which tests the intersection null, raises power to .02–.08 on Summary/QA but lets the uncorrected rule reach FDR .19–.41 (Appendix Table 10). **Interrupt-and-regenerate.** As Corollary 1 predicts, the sign of net token savings follows the hallucination base rate: Data2txt (64% hallucinated) nets +2.8/+4.6/+6.1% of corpus tokens at $\alpha = .01/.05/.1$, Summary is net-negative at every budget (-0.3 to -1.7%), and QA is positive only at the strictest (+0.9%), because false-alarm regenerations dominate when most episodes are faithful (Appendix Table 11).

6 Discussion and limitations

The detector ceiling. SeqGuard controls false alarms by construction; power comes from the detectors. Pooled power on single-span and subtle hallucinations (.23–.25 at $\alpha = .1$) reflects today's detectors, not the stopping rule; a FAR-matched threshold extracts no more. The hidden-state probe unlocks Data2txt at the price of white-box access.

Scope of the guarantee. The proof covers the conditionally calibrated rule, deployable at 74–80% of the marginal rule's power; the marginal rule we deploy is checked empirically, not certified (§3.4). Both assume calibration data from the deployment's faithful score distribution; under shift the rule degraded softly with no formal protection. Weighted conformal calibration (Tibshirani et al., 2019) could reweight the table if the faithful score law is what shifted (the score is a scalar, so the weight is estimable); adaptive conformal inference (Gibbs & Candès, 2021) needs post-episode labels, which an audit loop could supply with delay. We tested neither. Atomic-claim benchmarks for long-form text (Min et al., 2023) need a claim extractor in the loop and are future work.

Delay and exposure. Delay is the price of accumulation, and it has a user-facing cost: at $\alpha = 0.1$ only 22–51% of flags fire before the annotated span ends, so in most detections the false claim has already streamed (median 18 / 1 / 35 tokens past the span end). SeqGuard truncates; it does not prevent. Preventing exposure needs a display buffer of the order of the delay, or a rollback to the last checkpoint with wealth below 1, which removes the whole span in 72 / 70 / 51% of detections (Appendix B).

Economics. Interrupt-and-regenerate saves compute only where hallucinations are common: net savings are positive above the prevalence $\pi^* = C/(S + C)$, with C the tokens wasted per faithful episode and S the tokens saved per hallucinated one. At $\alpha = 0.1$, π^* is .39 / .23 / .36; Data2txt (64%) clears it, Summary (23%) does not (Appendix Table 11). Below π^* the rule is a guardrail, and α prices it.

7 Conclusion

Streaming hallucination detection has good sensors and an ad hoc trigger. We replaced the trigger: scores become p-values, p-values become bets, and a supermartingale turns the bets into an interruption rule with a length-independent false-alarm guarantee, proved for the conditional form and checked empirically for the deployed marginal form, that held where tuned thresholds broke. It flags when evidence compounds, and rarely when it does not.

AI USE STATEMENT

In this work, we used generative AI tools to draft and revise sections of the text under the authors' direction, to implement plotting code, to typeset tables and figures from result files, and to search for related work, which the authors verified against the original sources. We have not used generative AI tools to formulate the mathematical claims or proofs, to design the method or the experiments, to generate or clean data, or to interpret results; synthetic data generation and translation are not applicable to this work. Additionally, we used generative AI tools for editing prose and for checking the manuscript against the result files. We have reviewed all AI-assisted work: all LLM-drafted code was read, run, and checked by the authors, and every reported number is produced by the released code from cached detector scores and was verified against the source result files by the authors. LLMs are also the object of study: the datasets consist of LLM generations, and the detectors monitored include LLM-based scorers. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments run on cached detector scores: scoring each dataset requires one GPU pass ($\sim$1 A100-day in total across all experiments), after which every number in this paper is reproducible CPU-locally in minutes. The supplemental archive `code/` (README therein) and the anonymized repository `https://anonymous.4open.science/r/seqguard-32BD/` include the scoring and evaluation code, cached score files for all detectors and datasets, the exact calibration/test splits (official RAGTruth `train`/`test`; source-disjoint probe training split with seed 0), precomputed report JSONs, and scripts (`gen_tables.py`, `gen_figdata.py`, `check_claims.py`) that regenerate every table and figure, including the pgfplots data files compiled into this document, from those caches. Theorem proofs are in Appendix A; the harness validation against the released PrefixNLI stack, including a reproduction discrepancy we document for transparency, is in Appendix K.

REFERENCES

Aisha Alansari and Hamzah Luqman. Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61:100970, 2026. doi: 10.1016/j.cosrev.2026.100970.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

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*, 2023.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Werner Brannath and Lasse Fischer. Anytime-valid testing with e-values and confirmatory adaptive designs. *arXiv preprint arXiv:2606.00878*, 2026.

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)*, 2024.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Prasanjit Dubey and Xiaoming Huo. Anytime-valid federated conformal RAG for LLM swarms. *arXiv preprint arXiv:2605.29139*, 2026.

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

Isaac Gibbs and Emmanuel J Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jelle J Goeman, Rianne de Heide, and Aldo Solari. The e-partitioning principle of false discovery rate control. *arXiv preprint arXiv:2504.15946*, 2025.

Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024.

Sapir Harary, Eran Hirsch, Aviv Slobodkin, David Wan, Mohit Bansal, and Ido Dagan. PrefixNLI: Detecting factual inconsistencies as soon as they arise. *arXiv preprint arXiv:2511.01359*, 2025.

Mahd Hindi, Linda Mohammed, Ommama Maaz, and Abdulmalik Alwarafy. Enhancing the precision and interpretability of retrieval-augmented generation (rag) in legal technology: A survey. *IEEE Access*, 13:46171–46189, 2025. doi: 10.1109/ACCESS.2025.3550145.

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.

Jianpeng Hu, Yanzeng Li, Jialun Zhong, Wenfa Qi, and Lei Zou. Detecting hallucinations in retrieval-augmented generation via semantic-level internal reasoning graph. *arXiv preprint arXiv:2601.03052*, 2026.

Shan Huang and David Goretzko. Controlling the false discovery rate in DIF detection with e-values: Evidence from multidimensional and testlet simulations. *Educational and Psychological Measurement*, 2026.

Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. Tiny but uniform improvements of adaptive BH procedures via compound e-values. *arXiv preprint arXiv:2603.21424*, 2026.

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.

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.

K K Gordon Lan and David L DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

Jiawei Li, Akshayaa Magesh, and Venugopal V Veeravalli. Principled detection of hallucinations in large language models via multiple testing. *arXiv preprint arXiv:2508.18473*, 2025.

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)*, 2023.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

Haolang Lu, Minghui Pan, Ripeng Li, Guoshun Nan, Jialin Zhuang, Zijie Zhao, Zhongxiang Sun, Kun Wang, and Yang Liu. Streaming hallucination detection in long chain-of-thought reasoning. *arXiv preprint arXiv:2601.02170*, 2026.

Potsawee Manakul, Adian Liusie, and Mark JF 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.

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

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning (ICML)*, 2024.

Lebede Ngartera, Saralees Nadarajah, and Rodoumta Koina. Bayesian RAG: uncertainty-aware retrieval for reliable financial question answering. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2026. ISSN 2624-8212. doi: 10.3389/frai.2025.1668172.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-time detection of hallucinated entities in long-form generation. *arXiv preprint arXiv:2509.03531*, 2025.

Peter C O'Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

George Papageorgiou, Vangelis Sarlis, Manolis Maragoudakis, Ioannis Magnisalis, and Christos Tjortjis. Evaluating faithfulness in agentic RAG systems for e-governance applications using LLM-based judging frameworks. *Big Data and Cognitive Computing*, 9(12):309, 2025.

Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

Nicholas G Polson, Vadim Sokolov, and Daniel Zantedeschi. Bayes, e-values and testing. *arXiv preprint arXiv:2602.04146*, 2026.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations (ICLR)*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431, 2021.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential change detection via backward confidence sequences. In *International Conference on Machine Learning (ICML)*, 2023.

Alexandra Sokolova and Vadim Sokolov. E-values for adaptive clinical trials: Anytime-valid monitoring in practice. *arXiv preprint arXiv:2602.06379*, 2026.

Jiajun Sun, Zhanrui Cai, and Wenxuan Zhong. A general stability approach to false discovery rate control. *arXiv preprint arXiv:2512.17401*, 2025.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024a.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024b.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Jean Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

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

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86(1):1–27, 2024.

Krzysztof Wołk. Evaluating retrieval-augmented generation variants for clinical decision support: Hallucination mitigation and secure on-premises deployment. *Electronics*, 14(21):4227, 2025.

Ziyu Xu, Aldo Solari, Lasse Fischer, Rianne de Heide, Aaditya Ramdas, and Jelle J Goeman. Bringing closure to false discovery rate control: A general principle for multiple testing. *arXiv preprint arXiv:2509.02517*, 2025.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. Stable-RAG: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*, 2026.

Guoshenghui Zhao, Weijie Zhao, and Tong Yu. HIVE: Hidden-evidence verification for hallucination detection in diffusion large language models. *arXiv preprint arXiv:2604.26139*, 2026.

A PROOFS

A.1 PROOF OF THEOREM 1

Fix a detector j and write $e_t = e(p_t^{(j)})$. Since e is predictable and $\int_0^1 e(u)\,du \le 1$ with $e \ge 0$, and p_t is conditionally super-uniform (Assumption 1), we have for any nonincreasing calibrator (both our calibrators are nonincreasing in p)

$$\begin{aligned}\mathbb{E}\big[e(p_t) \mid \mathcal{F}_{t-1}\big] &= \int_0^\infty \mathbb{P}\big(e(p_t) > x \mid \mathcal{F}_{t-1}\big)\,dx \\ &\le \int_0^\infty \mathrm{Leb}\{u : e(u) > x\}\,dx = \int_0^1 e(u)\,du \le 1,\end{aligned}$$

where the inequality uses $\{e(p_t) > x\} = \{p_t < u_x\}$ for the level set threshold u_x and $\mathbb{P}(p_t < u_x \mid \mathcal{F}_{t-1}) \le u_x$. Hence $W_t^{(j)} = W_{t-1}^{(j)} e_t$ satisfies $\mathbb{E}[W_t^{(j)} \mid \mathcal{F}_{t-1}] \le W_{t-1}^{(j)}$: each per-detector wealth is a nonnegative supermartingale with $W_0^{(j)} = 1$. A convex combination $W_t = \sum_j \omega_j W_t^{(j)}$ ($\omega_j \ge 0$, $\sum \omega_j = 1$; we use $\omega_j = 1/m$) is again a nonnegative supermartingale with $W_0 = 1$. Ville's inequality gives $\mathbb{P}(\sup_t W_t \ge 1/\alpha) \le \alpha W_0 = \alpha$. The event $\{\tau < \infty\}$ equals $\{\sup_t W_t \ge 1/\alpha\}$ by definition of τ. Optional stopping is free: the bound is on the supremum over all t, hence over all stopping times. □

Remark 1. Per-step e-averaging ($e_t = \frac{1}{m}\sum_j e(p_t^{(j)})$, one shared product) is also covered: $\mathbb{E}[e_t \mid \mathcal{F}_{t-1}] \le 1$ by linearity, so the shared product is a supermartingale. Both fusions are valid; §3.2 explains why wealth-mix is preferred when detector strengths are unequal.

A.2 PROOF OF THEOREM 2

Let the calibration cells be indexed $c = 1, \ldots, N_{\text{cells}}$, each an i.i.d. sample of size n_{cell} from the faithful score law F_c restricted to that cell's predictable context. By the DKW inequality with Massart's constant (Massart, 1990), for each cell,

$$\mathbb{P}\Big(\sup_x \big|\widehat{F}_c(x) - F_c(x)\big| > \epsilon\Big) \le 2e^{-2n_{\text{cell}}\epsilon^2},$$

so with $\epsilon_n = \sqrt{\log(2N_{\text{cells}}/\delta)/(2n_{\text{cell}})}$, a union bound gives $\sup_c \sup_x |\widehat{F}_c - F_c| \le \epsilon_n$ simultaneously with probability $\ge 1 - \delta$ over the calibration draw. On that event, for any test score s falling in cell c, the empirical p-value equation 1 satisfies $p_t \ge F_c(s_t) - \epsilon_n$ pointwise, hence $\mathbb{P}(p_t^\dagger \le u \mid \mathcal{F}_{t-1}, \widehat{F}) \le \mathbb{P}(p_t \le u - \epsilon_n \mid \mathcal{F}_{t-1}, \widehat{F}) \le \mathbb{P}(F_c(s_t) \le u \mid \mathcal{F}_{t-1}) \le u$: the inflated value is conditionally super-uniform given $\widehat{F}$. Assumption 1 then holds pathwise for the inflated process, and Theorem 1 applied conditionally on $\widehat{F}$ yields $\mathrm{FAR}(\widehat{F}) \le \alpha$ on the good event. For the un-inflated rule, the same uniform deviation bounds each per-step conditional expectation by $\mathbb{E}[e(p_t) \mid \mathcal{F}_{t-1}, \widehat{F}] \le 1 + \zeta_n$ with $\zeta_n = O(\epsilon_n \cdot e(u_{\min}))$ by integration by parts, $u_{\min} = 1/(n_{\text{cell}} + 1)$ being the smallest achievable p-value, so $W_t(1 + \zeta_n)^{-t}$ is a supermartingale and on episodes of length $\le T_{\max}$ Ville gives $\mathrm{FAR}(\widehat{F}) \le \alpha(1 + \zeta_n)^{T_{\max}}$. This is a finite-sample statement with $\eta_n = \alpha[(1 + \zeta_n)^{T_{\max}} - 1] \to 0$ only when e is bounded, such as the linear bet or the power mixture truncated at a fixed u_0: then $\zeta_n = O(\epsilon_n) = O(n_{\text{cell}}^{-1/2})$. For the unbounded power mixture, $e(u_{\min}) = \Theta(n_{\text{cell}}^{0.9})$ from the $\kappa = 0.1$ term, so $\zeta_n = O(n_{\text{cell}}^{0.4})$ diverges and the DKW route proves nothing about the un-inflated rule; a relative, Bernstein-type deviation bound would give $\zeta_n = O(\epsilon_n)$ and is the route we leave to future work. The resample study (Table 8) is the empirical evidence for the un-inflated rule at our table sizes. □

Remark 2. The empirical companion is the calibration-resample study (Table 8): $0/120$ budget violations, with realized-FAR spread of at most .016 across resampled tables. We run the operating configuration without inflation, quoting Theorem 2 as the formal finite-n statement and the resample study as evidence that the un-inflated estimator's optimism is negligible at our calibration sizes (16k–22k scores per domain).

A.3 Proof of Theorem 3

On $\{\tau \geq t^*\} \cap \{\tau < \infty\}$, wealth must travel from $W_{t^*-1} < 1/\alpha$ (else the rule would have stopped before t^*) to $W_\tau \geq 1/\alpha$. Write $L_t = \log W_t$. For $t \geq t^*$, $\mathbb{E}[L_t - L_{t-1} \mid \mathcal{F}_{t-1}] \geq \log\gamma$ by assumption. Define, for $t \geq t^* - 1$, the stopped process $M_t = (L_{t\wedge\tau} - L_{t^*-1}) - ((t \wedge \tau) - (t^* - 1))\log\gamma$; it is a submartingale with $M_{t^*-1} = 0$, and optional stopping (delays have finite expectation whenever $\mathbb{P}(\tau < \infty) > 0$ and increments are bounded below, which holds since $e(p) \geq e(1) = \frac{1}{|\mathcal{K}|}\sum_\kappa \kappa > 0$) gives

$$\mathbb{E}[\Delta \mid \tau < \infty] \cdot \log\gamma \;\leq\; \mathbb{E}[L_\tau - L_{t^*-1} \mid \tau < \infty] \;\leq\; \log\tfrac{1}{\alpha} - \log W_{t^*-1} + B,$$

where B bounds the expected overshoot $\mathbb{E}[\log e(p_\tau)]$ at the crossing step (finite because a single e-value is bounded by $e(u_{\min})$, the calibrator at the smallest achievable p-value). Dividing by $\log\gamma$ gives the statement. The pre-onset term $-\log W_{t^*-1}$ is not bounded uniformly in t^*: under the faithful null each pre-onset bet has $\mathbb{E}[\log e(p_t) \mid \mathcal{F}_{t-1}] \leq \log\mathbb{E}[e(p_t) \mid \mathcal{F}_{t-1}] \leq 0$ by Jensen, with strict inequality for any non-degenerate bet, and wealth has no reflecting barrier, so the expected drawdown grows with t^*. It is $O(1)$ for bounded onset horizons only. Optional stopping applies because increments are bounded below ($e(p) \geq e(1) = \frac{1}{|\mathcal{K}|}\sum_\kappa \kappa > 0$) and above ($e(p) \leq e(u_{\min})$), and $\tau < \infty$ almost surely on the alternative whenever $\log\gamma > 0$. Table 2 reports the empirical drawdown by onset. □

Table 2: Pre-onset drawdown $-\log W_{t^*-1}$ (nats) of the fused operating configuration on hallucinated RAGTruth test episodes, by onset checkpoint t^* (0-based): median / 90th percentile. The drawdown grows with t^*; $\log(1/\alpha) = 2.3$ at $\alpha = 0.1$.

Onset t^*	Summary	QA	Data2txt
0	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
1	0.23 / 0.56	0.25 / 0.58	0.05 / 0.45
2	0.52 / 1.03	0.48 / 1.03	0.52 / 1.08
3–4	0.77 / 1.90	0.54 / 1.71	0.89 / 1.46
5–7	1.48 / 3.36	1.42 / 2.70	0.73 / 2.36
8+	0.56 / 0.56 ($n = 1$)	2.45 / 5.56	1.62 / 5.66

Empirically, the operating configuration's mean delays at $\alpha = 0.1$ are 32–73 tokens (Table 1); halving α to 0.05 adds roughly $\log 2/\log\gamma$ checkpoints, matching the logarithmic scaling. The realized post-onset growth rate $\log\gamma$ (median per-checkpoint log-wealth gain on detected episodes) is 1.6 / 1.3 / 0.9 nats on Summary / QA / Data2txt, so $\log(1/\alpha)/\log\gamma$ predicts 1.4 / 1.8 / 2.6 checkpoints against measured mean delays of 1.3 / 2.1 / 3.2.

A.4 Proof of Proposition 1 and Corollary 1

Proposition 1 is Theorem 2 of Wang & Ramdas (2022): the e-BH procedure applied to arbitrary e-values controls FDR at level α under arbitrary dependence; we only need that each E_i is an e-value under its own null $H_{0,i}$ ("span i is faithful"). Let t_i^{start} be the first checkpoint whose prefix contains span i's text and t_i^{end} the first checkpoint covering its end, and $E_i = \prod_{u=t_i^{\text{start}}}^{t_i^{\text{end}}} e(p_u) = W_{t_i^{\text{end}}}/W_{t_i^{\text{start}}-1}$. Under $H_{0,i}$ the checkpoints inside the span are faithful, so by Assumption 1 each factor has conditional expectation at most 1 given the past, and iterating the tower property from t_i^{end} down to t_i^{start} gives $\mathbb{E}[E_i \mid \mathcal{F}_{t_i^{\text{start}}-1}] \leq 1$, hence $\mathbb{E}[E_i] \leq 1$, whatever earlier spans did. For the fused rule we average the per-detector ratios, a convex combination of e-values. The prefix wealth $W_{t_i^{\text{end}}}$ used in an earlier draft is an e-value for the intersection null "no hallucination up to t_i^{end}" only: if an earlier span was hallucinated, $\mathbb{E}[W_{t_i^{\text{end}}}]$ can exceed 1 under $H_{0,i}$. Table 10 reports both. For Corollary 1: a false abort occurs with probability $\leq \alpha$ (Theorem 1) and wastes at most the episode's T tokens, giving the first claim by Wald-type accounting; the second claim substitutes Theorem 3's delay bound into tokens-not-generated $= (T - \tau)_+ \geq (T - t^*)_+ - \Delta \cdot \bar{L}$, with $\bar{L}$ the mean tokens per checkpoint. The waste bound uses $\mathbb{E}[T \cdot \mathbf{1}\{\tau \leq T\}] \leq T_{\max}\,\mathbb{P}(\tau \leq T) \leq \alpha T_{\max}$; the version with $\mathbb{E}[T]$ needs flag and length uncorrelated, which Figure 2 (right) shows they are not quite (on faithful test episodes the ratio $\mathbb{E}[T\mathbf{1}\{\tau \leq T\}]/(\alpha\,\mathbb{E}[T])$ is 0.40–0.80 at $\alpha = 0.1$). □

B FULL PER-DOMAIN RESULTS AND PREMATURE-FLAG ACCOUNTING

Table 1 reports the operating points. The supplemental archive's `tables/table_main.tex` includes the full grid (4 budgets $\times$ 9 rules $\times$ 3 domains, with premature rates and p90 delays), regenerable from cached scores. Premature rates at $\alpha = 0.1$ are $.33/.40/.55$ per domain for B1 and $.020/.037/.019$ for fused SEQGUARD. In the interrupt-and-regenerate application, a premature flag on a to-be-hallucinated episode is arguably harmless: the regeneration replaces a generation that would have gone wrong. Counting premature flags as successes would raise every method's power without changing the ordering. We report the conservative accounting in the main text. For fused SEQGUARD at $\alpha = 0.1$, flags fire before the generation ends in 72 / 85 / 79% of detections (Summary / QA / Data2txt) and before the annotated span ends in 22 / 51 / 23%; the FAR-matched fused Fisher threshold fires before the end in 72 / 89 / 89% and before the span end in 29 / 59 / 45% (`results/fused_baselines.json`). Counting detections only and converting to tokens, the median number of tokens streamed past the annotated span's end before a fused SEQGUARD flag is 18 / 1 / 35 (mean 27 / 18 / 51); rolling back to the last checkpoint with wealth below 1 discards a median of 58 / 77 / 76 tokens and removes the whole span in 72 / 70 / 51% of detections (`results/exposure.json`). Table 3 is the cross-benchmark shift table discussed in §5.1.

Table 3: Realized FAR under shift: calibrate on the left dataset, deploy on the right (NLI detector, no recalibration). Bold marks budget violations. B2 holds only in-domain and forward; deployed on RAGTruth from external calibration it is over budget at every level (1.4–1.6$\times$ at $\alpha \in \{0.1, 0.2\}$). The e-process stays under budget throughout.

	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.2$	
Calibrate $\rightarrow$ deploy	B2	e-proc.	B2	e-proc.	B2	e-proc.
RAGTruth $\rightarrow$ RAGTruth (in-domain)	.042	.024	.085	.039	.175	.066
RAGTruth $\rightarrow$ AggreFact	.020	.009	.055	.017	.101	.031
RAGTruth $\rightarrow$ TofuEval	.024	.006	.053	.035	.118	.047
AggreFact $\rightarrow$ RAGTruth	**.062**	.023	**.141**	.040	**.323**	.063
TofuEval $\rightarrow$ RAGTruth	**.062**	.014	**.141**	.029	**.303**	.056

C SPAN-TYPE AND ABLATION TABLES

Table 4 gives the annotation-driven span-type breakdown of §5.2. Table 5 gives the ablations of §5.3, plus the linear-vs-power-mixture calibrator comparison of §3.2. Table 6 is the Theorem 2 inflation cost; Table 7 is the full fused marginal vs. conditional grid.

D CALIBRATION-RESAMPLE STUDY

Table 8 reports the study behind Theorem 2's empirical companion: the headline configuration re-evaluated on 20 bootstrap resamples of the calibration set, per domain, at $\alpha \in \{0.05, 0.1\}$.

E TOKEN CLOCK AND THE BONFERRONI HORIZON EFFECT

The token clock bets at every generated token using a per-token detector (teacher-forced log-probabilities), fused with the sentence-clock NLI detector by wealth mix. There are two findings. **(1) Validity holds at token granularity.** Realized FAR is $\leq \alpha$ for every e-process variant over horizons of $\sim$150–500 tokens, even with marginal calibration. Within-sequence dependence does not mint wealth on these detectors. **(2) Bonferroni does not survive the horizon.** Token-level Bonferroni with the realized horizon T realizes FAR at or near budget on every domain, with outright violations on Summary and Data2txt at $\alpha = 0.1$ and on Data2txt at $\alpha = 0.2$ (Table 9). The union bound holds under any dependence; the violation comes from its input. The per-token bar α/T is an extreme quantile of one shared finite table, the table's error at that quantile is not small relative to the bar, and T is a random stopping time supplied after the fact. With a fixed causal horizon $T_{\max} = 500$ the same rule is under budget everywhere (.024 / .027 / .065 at $\alpha = 0.1$) and has power

Table 4: Power by annotation-defined class at $\alpha = 0.1$ (fused detector). The comparison threshold is FAR-matched on the fused Fisher stream (same two detectors as SEQGUARD). Per-domain rows back the Data2txt ratios quoted in §5.2; on Summary and QA the threshold is level or ahead. The older score-defined $5.4\times$ is in the text as a mechanism check only.

Class	n	SEQGUARD	FAR-matched threshold	ratio
Annotation partitions, $\alpha = 0.1$; threshold is FAR-matched on the fused Fisher stream				
multi-span (pooled)	353	.567	.436	1.3×
single-span (pooled)	590	.254	.219	1.2×
evident (pooled)	866	.383	.305	1.3×
subtle-only (pooled)	77	.234	.247	0.9×
Per domain and class ($\alpha = 0.1$; delay in tokens)				
Summary multi-span	33	.485 (53)	.455 (19)	1.1×
Summary single-span	171	.175 (22)	.234 (5)	0.7×
Summary evident	185	.238 (33)	.281 (9)	0.8×
Summary subtle-only	19	.105 (29)	.158 (11)	0.7×
QA multi-span	49	.510 (42)	.490 (29)	1.0×
QA single-span	111	.324 (48)	.369 (14)	0.9×
QA evident	131	.405 (47)	.427 (20)	0.9×
QA subtle-only	29	.276 (37)	.310 (14)	0.9×
Data2txt multi-span	271	.587 (72)	.424 (44)	1.4×
Data2txt single-span	308	.273 (74)	.156 (50)	1.8×
Data2txt evident	550	.427 (73)	.284 (47)	1.5×
Data2txt subtle-only	29	.276 (61)	.241 (30)	1.1×

Table 5: Ablations (realized FAR / power at $\alpha = 0.1$). Calibration-size rows use the NLI detector; a few hundred faithful episodes give validity and near-final power. Fusion rows compare wealth-mix with per-step e-value averaging. Calibration-mode rows show the conditional table's safety–power trade.

	Summary	QA	Data2txt
Calibration-set size (NLI only, $\alpha = 0.1$): FAR / power			
n =100	.029 / .206	.038 / .263	.047 / .081
n =250	.030 / .196	.038 / .256	.047 / .073
n =500	.036 / .225	.039 / .287	.047 / .062
n =2000	.039 / .240	.048 / .294	.050 / .066
n = all	.039 / .230	.049 / .275	.050 / .066
Detector fusion ($\alpha = 0.1$): FAR / power			
NLI only	.039 / .230	.049 / .275	.050 / .066
probe only	.022 / .093	.054 / .244	.072 / .444
fused, per-step e-mean	.022 / .201	.045 / .350	.037 / .306
fused, wealth mix	.034 / .225	.062 / .381	.072 / .420
calibration: one ckpt / episode / bucket (mean of 5 seeds)	.032 / .229	.062 / .372	.069 / .399
Calibrator (NLI only, marginal): power at $\alpha = 0.01$ / $\alpha = 0.1$			
linear bet, $\lambda = 0.5$	.000 / .044	.019 / .119	.000 / .059
power mixture (eq. 2)	.108 / .230	.175 / .275	.017 / .066
Calibration mode (NLI only, $\alpha = 0.1$): FAR / power			
marginal (position-bucketed)	.039 / .230	.049 / .275	.050 / .066
conditional (§3.4)	.010 / .152	.020 / .219	.028 / .033

at most .025; with $T_{\max} = 1{,}000$ its FAR is .016 / .013 / .047. Power currently favors the sentence clock because the available per-token detector is weak (Appendix F). The token clock is validated infrastructure for stronger per-token detectors.

Table 6: Fused SEQGUARD with and without Theorem 2 p-value inflation (FAR / power). Inflation costs $\leq .019$ power at $\alpha = 0.1$ and pulls Data2txt at $\alpha = 0.01$ from .016 to .012.

	$\alpha = 0.01$		$\alpha = 0.1$	
Domain	raw	inflated	raw	inflated
Summary	.007 / .103	.007 / .098	.034 / .225	.033 / .225
QA	.007 / .206	.007 / .206	.062 / .381	.060 / .362
Data2txt	.016 / .206	.012 / .195	.072 / .420	.065 / .404

Table 7: Fused wealth-mix under marginal, conditional (v1: position $\times$ fixed previous-score bins, marginal fallback in cells under 50 points), and certified conditional calibration (position $\times$ previous-score half split at the calibration median, no bet in thin cells; valid pathwise), FAR / power. The certified rule keeps 74–80% of marginal power at $\alpha = 0.1$; v1 kept 55–68%. The result is insensitive to the fallback threshold: requiring 20, 50, or 100 points per cell moves conditional FAR by at most $.01$ and power by at most $.02$ at $\alpha = 0.1$. Under shift between RAGTruth Summary and AggreFact/TofuEval (NLI only, both directions), the conditional rule's FAR at $\alpha = 0.1$ stays at .010–.023.

	$\alpha = 0.01$			$\alpha = 0.1$		
Domain	marginal	cond. v1	certified	marginal	cond. v1	certified
Summary	.007 / .103	.001 / .029	.004 / .074	.034 / .225	.013 / .152	.017 / .181
QA	.007 / .206	.004 / .131	.004 / .156	.062 / .381	.018 / .231	.035 / .281
Data2txt	.016 / .206	.009 / .095	.012 / .142	.072 / .420	.037 / .230	.056 / .326

F THE LOG-PROBABILITY NEGATIVE RESULT

Teacher-forced mean segment log-probability carries almost no hallucination-onset signal on RAGTruth. As the sole detector at $\alpha = 0.1$, its power is $.029/.044/.009$ (Summary/QA/Data2txt) from a Qwen2.5-1.5B proxy, and $.034/.019/.007$ from Llama-2-7b-chat, the generator family of a subset of RAGTruth responses. Per-token variants are weaker still ($\leq .013$). Fusing either into the NLI detector costs its bounded wealth-mix factor and buys nothing. Our interpretation: hallucinated continuations are fluent and high-probability under both proxy and generator-family models, so realized-token likelihood does not carry the signal. Hidden states do: the probe reads layer-16 representations of the same model whose log-probabilities are uninformative. Two caveats. RAGTruth responses are pre-generated by six models, so only a subset is scored by its true generator, and API-generated responses cannot be teacher-forced at all. Decode-time signals with full access to the generator's distribution, such as token-level entropy, remain untested here.

G SPAN-LEVEL E-BH AND SYSTEMS PAYOFF TABLES

Table 10 gives the span-level e-BH results and Table 11 the interrupt-and-regenerate accounting of §5.3.

H BREADTH: HALUEVAL WITH NO ADAPTATION

Table 12 reports zero-adaptation transfer from RAGTruth Summary calibration to the three HaluEval tasks. Both rules exceed their budgets on summarization: B2 by $2.0\times$ ($.100$ at $\alpha = 0.05$, $.195$ at $\alpha = 0.1$) and the e-process by 1.2–$1.5\times$ ($.075$, $.117$). QA scores are so well separated that both methods are ultra-conservative, and the e-process loses more power than the threshold ($.100$ vs. $.403$ at $\alpha = 0.1$). The reason is length, not score shift: HaluEval QA answers are one checkpoint long (median 1), so the e-process has a single bet. Hallucinated answers score low (median NLI $.07$, median $p = .04$), but one power-mixture payoff at such a p is a few units and reaches 10 on only 10% of hallucinated episodes, while a threshold fires on any one low score. Faithful answers pay $e < 1$ at 94% of checkpoints, which is why FAR is 0. Dialogue is a deliberate far-out-of-distribution stress test, and both methods break there: the tuned threshold by $5.5\times$, the e-process by $2.2\times$. Weighted

Table 8: Realized FAR of the headline configuration across 20 bootstrap resamples of the calibration set, per domain and budget. No resample violates its budget ($0/120$ overall).

	$\alpha = 0.05$		$\alpha = 0.1$	
Domain	FAR range	violations	FAR range	violations
Summary	[.020, .027]	0/20	[.037, .043]	0/20
QA	[.031, .038]	0/20	[.048, .050]	0/20
Data2txt	[.025, .037]	0/20	[.047, .053]	0/20

Table 9: Token-clock realized FAR at $\alpha = 0.1$ (test episodes; horizons $\sim$150–500 tokens). Bold marks budget violations. "Oracle T" gives Bonferroni the realized episode length; $T_{\max} = 500$ is the causal version.

Method	Summary	QA	Data2txt
token e-process (marginal)	.085	.039	.078
token e-process (conditional)	.035	.046	.059
two-clock wealth mix (marginal)	.063	.052	.087
token Bonferroni (oracle T)	**.103**	.097	**.131**
token Bonferroni, $T_{\max} = 500$	.024	.027	.065
token Bonferroni at $\alpha = 0.2$ (oracle T)	.182	.173	**.215**

or adaptive conformal calibration (Tibshirani et al., 2019; Gibbs & Candès, 2021) is the natural fix when the deployment distribution is known to differ.

I Onset-jitter sensitivity

Table 13 shifts the annotated onset checkpoints by $\pm j$ and re-measures power. A flag counts as a detection only if it fires at or after onset. Later onsets ($+j$) therefore reclassify early detections as premature and lower measured power; earlier onsets ($-j$) do the reverse. At $j = 1$, the symmetric mean moves by at most .04 in every domain. Validity is untouched, since FAR uses no onset labels.

J Dependence: slack bound and autocorrelation stress test

A slack bound, and why it is not useful here. If the marginal rule's bets satisfy $\mathbb{E}[e(p_t) \mid \mathcal{F}_{t-1}] \le 1 + \delta$ under H_0, then $W_t(1+\delta)^{-t}$ is a test supermartingale and $\mathbb{P}_{H_0}(\tau \le T_{\max}) \le \alpha(1+\delta)^{T_{\max}}$, with $\delta = 0$ recovering Theorem 1. We estimated δ on held-out faithful train episodes (table built on one half, payoffs measured on the other, five seeds) by conditioning the marginal p-values on position and previous-score bin in cells of at least 50 points, taking the largest cell mean minus one, with a bootstrap upper limit. Most cells are conservative (the point-weighted mean slack is $-.11$ to $-.12$), but the maximum is driven by a few cells with heavy-tailed payoffs: its upper limit is .63–.76 for the fused rule, so the bound is 199α–$2{,}706\alpha$ at the 95th-percentile length and vacuous. Truncating the calibrator at $p \ge 1/1000$ made the estimate worse, not better. We report this as a negative result: the bound is true, but with unbounded payoffs and finite cells the plug-in constant is too loose to certify anything (`results/slack_report.json`).

Stress test. To see where marginal calibration breaks, we replace each faithful test episode's p-value stream by a synthetic one with the same length and exactly uniform marginals but controlled dependence: $z_1 \sim \mathcal{N}(0,1)$, $z_t = \rho z_{t-1} + \sqrt{1-\rho^2}\,\varepsilon_t$, $p_t = \Phi(z_t)$, for $\rho \in \{0, 0.3, 0.6, 0.8, 0.9, 0.95\}$ and 20 seeds. Position-bucketed marginal calibration is then exactly right marginally, and only the dependence changes; positive ρ produces runs of low p-values, the wealth-minting case. The conditional rules are fit on synthetic train streams generated the same way. Figure 4 shows realized FAR at $\alpha = 0.1$. The single-detector marginal e-process stays at or under budget up to $\rho = 0.9$ on every domain (.068 / .069 / .100) and crosses only on Data2txt at $\rho = 0.95$ (.108); the conditional and no-bet rules hold at every ρ ($\le .044$). A block bootstrap that concatenates blocks of 2 or 3 real consecutive checkpoints from other episodes inflates FAR only to .020–.034. The real lag-1 Spearman correlation of consecutive p-values on faithful test episodes is 0.70 / 0.61 / 0.56 for the NLI detector

Table 10: Span-level FDR / power. "Ratio" rows use the within-span wealth ratio $E_i = W_{t_i^{\text{end}}}/W_{t_i^{\text{start}}-1}$ of Proposition 1 (NLI alone, or fused NLI + probe by averaging per-detector ratios); "prefix" rows use the prefix wealth $W_{t_i^{\text{end}}}$, an e-value for the intersection null only. e-BH controls FDR under arbitrary dependence; the uncorrected rule (reject if $E \geq 1/\alpha$) does not. Nulls are pseudo-spans drawn from faithful episodes with matched length. Almost every annotated span covers one checkpoint, so ratio e-values cannot compound, and on Data2txt they support no e-BH rejections.

		$\alpha = 0.05$		$\alpha = 0.1$	
Domain, e-value	spans	e-BH	uncorrected	e-BH	uncorrected
Summary prefix NLI	244	.000 / .020	.161 / .107	.000 / .020	.191 / .139
Summary ratio NLI	244	.000 / .004	.000 / .074	.000 / .004	.032 / .123
Summary ratio fused	244	.000 / .004	.000 / .033	.000 / .004	.000 / .090
QA prefix NLI	235	.000 / .047	.023 / .179	.000 / .081	.017 / .243
QA ratio NLI	235	.000 / .004	.037 / .111	.000 / .026	.022 / .187
QA ratio fused	235	.000 / .013	.038 / .106	.000 / .021	.054 / .149
Data2txt prefix NLI	1054	.000 / .000	.400 / .017	.000 / .000	.412 / .028
Data2txt ratio NLI	1054	.000 / .000	.000 / .001	.000 / .000	.067 / .013
Data2txt ratio fused	1054	.000 / .000	.000 / .002	.000 / .000	.040 / .023

Table 11: Interrupt-and-regenerate accounting (fused detector): net corpus tokens saved (saved on detected hallucinations minus wasted on false-alarm regenerations), as % of all generated tokens, and % of hallucination-episode tokens saved at $\alpha = 0.1$. Cost model: a flagged faithful episode wastes its full length T (it is discarded and regenerated); a flagged hallucinated episode saves $T - \text{tokens}(\tau)$; the replacement generation is not charged on either side. With C the mean tokens wasted per faithful episode and S the mean tokens saved per hallucinated one, net savings per episode at prevalence π are $\pi S - (1-\pi)C$, positive above $\pi^* = C/(S+C)$; the last column gives π^* at $\alpha = 0.1$. Observed prevalence is .23 / .18 / .64. The sign tracks the hallucination base rate, as Corollary 1 predicts.

Domain	net @ α=.01	net @ α=.05	net @ α=.1	halluc. tokens saved @ .1	π^* @ .1
Summary	-0.3%	-0.7%	-1.7%	6.3%	.391
QA	+0.9%	-0.7%	-1.6%	18.5%	.233
Data2txt	+2.8%	+4.6%	+6.1%	13.9%	.357

and 0.28 / 0.38 / 0.39 for the probe (`results/autocorr_stress.json`), below the break on Data2txt and in the under-budget region on Summary and QA.

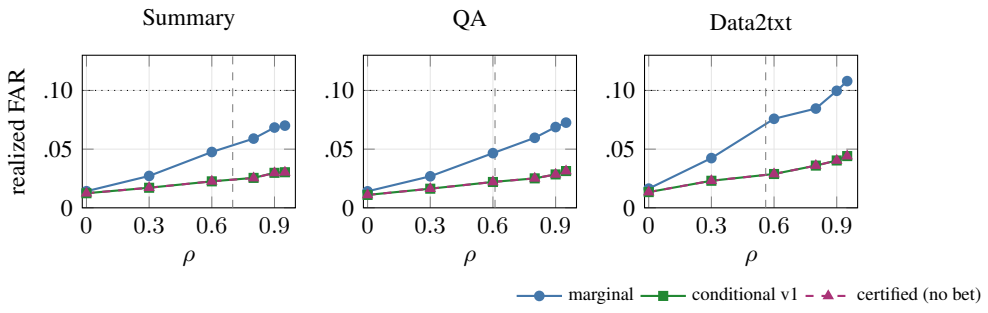

Figure 4: Realized FAR at $\alpha = 0.1$ on synthetic faithful streams with Gaussian-copula AR(1) dependence of strength ρ (exactly uniform marginals; 20 seeds; real episode lengths). Dotted line: the budget. Vertical dashed line: the real lag-1 correlation of the NLI p-value stream on that domain. Marginal calibration holds until $\rho = 0.9$ and crosses only on Data2txt at $\rho = 0.95$; the conditional rules hold everywhere.

Table 12: HaluEval (calibrated on RAGTruth Summary, no adaptation): realized FAR / power. Bold FAR marks budget violations. Dialogue is a deliberate stress test far outside the calibration distribution: both methods break, the tuned threshold by $5.5\times$ its budget, the e-process by $2.2\times$.

	$\alpha = 0.05$		$\alpha = 0.1$	
Task	B2 tuned	e-process	B2 tuned	e-process
Summarization	**.100** / .260	**.075** / .235	**.195** / .417	**.117** / .282
QA	.000 / .235	.000 / .033	.000 / .403	.000 / .100
Dialogue	**.280** / .530	**.110** / .350	**.547** / .848	**.217** / .502

Table 13: Measured power at $\alpha = 0.1$ (NLI detector) when annotated onsets are shifted by $\pm j$ checkpoints. Shifting onsets later ($+$) reclassifies early detections as premature, lowering measured power; shifting earlier ($-$) does the reverse. Realized FAR does not depend on onset labels and is unaffected.

	Summary		QA		Data2txt	
jitter	$+j$	$-j$	$+j$	$-j$	$+j$	$-j$
$j = 0$	.230	.230	.275	.275	.066	.066
$j = 1$	.157	.230	.181	.287	.062	.071
$j = 2$	.103	.245	.119	.294	.050	.074
$j = 3$	.059	.245	.094	.300	.036	.074

K HARNESS VALIDATION AGAINST PREFIXNLI, AND A REPRODUCTION NOTE

Our conclusions do not depend on matching PrefixNLI's reported benchmark numbers: every method in this paper consumes identical cached scores, so all comparisons are internal. We validated the harness anyway, and we document a discrepancy for transparency. We ran MiniTruePrefixes under the authors' released inference stack (vLLM, fp16, their exact manual prompt-token construction, greedy single-token decoding) on the full released RAGTruthPrefixes test split (210,678 instances). We obtain prefix-entailment F1 $= 37.3$, 95% CI $[36.8, 37.7]$ (unfaithful class; P $= .258$, R $= .671$), and 36.2 on the released dev split. The paper reports 47.6 ± 0.6. Our own transformers/bf16 harness agrees with the released stack to within 0.1 F1, so our scorer drives the model as the authors' code does. Class balance cannot explain the gap either: the released split is 7.8% unfaithful, and our sample matched it. Binned by prefix length, our numbers agree closely with the paper's Figure 4 on complete sentences (53.7 vs. ~ 55); the discrepancy concentrates on short prefixes. We have contacted the authors. An evaluation-filtering detail absent from the released artifacts is the likely cause. None of this affects the detector's use here: entailed prefixes score .997–.998, fabricated ones .047, and the main-text results measure exactly the streaming score separations our rule consumes.

L QUALITATIVE WEALTH PATHS

The supplemental archive and anonymized repository include per-episode wealth-path tables (`results/wealth_paths.md`: score, p-value, e-value, wealth per checkpoint) for detection, miss, false-alarm, and clean-pass cases in each domain; regenerate with `scripts/extract_qualitative.py -no_text`. Figure 1 (left) renders three. In the detection, no single checkpoint clears $1/\alpha$ at $\alpha = .01$, but compounding does. In the miss, two low-scoring checkpoints among high-scoring ones cannot sustain growth. In the clean pass, wealth decays monotonically: the guarantee, mechanically.