Efficient Expert Spawning in Compound NLP Systems: A Counterfactual-Oracle Protocol

Anonymous ACL submission

Abstract

Compound NLP systems increasingly spawn *specialized adapters* (LoRA, MoE experts) on emergent query domains, but no principled method exists to decide *when* to spawn or how to *evaluate* that decision non-circularly: heuristic **eager** spawning wastes compute on clusters the base model already handles, and ground truth is typically derived from the same uncertainty signal that drives the gate. We contribute (i) a **counterfactual-oracle protocol** that trains a low-rank adapter on *every* detected cluster and uses an external LLM judge for non-circular birth ground truth, and (ii) a **sequential observation gate** (SPRT / GLR-SPRT, Wald 1945) as a principled instantiation with explicit (α, β) error control. We validate on a 1,379-query live deployment, a controlled 3-seed workload, MMLU, and a cross-lingual FR/EN civil-code study showing language-dependent UQ calibration (KS=0.30, p=2.3$\times 10^{-4}$). The gate matches a K=50 batched probe at $\sim$0.44$\times$ cost on deployment; multi-UQ composite plus a MASS-style gradient-surge probe lift oracle agreement from $2/7$ to $5/7$ (F1=0.75). We release code and benchmark suite.

1 Introduction

Compound NLP systems — agentic assistants, multi-skill copilots, on-the-fly routed adapter stacks — increasingly evolve at deployment time by spawning *specialized adapters* (LoRA, MoE experts) on emergent query domains. Two design decisions are typically left under-specified: **(D1) when to spawn?** and **(D2) how do we know whether the spawn was right?**. Current practice answers (D1) with heuristic *eager* spawning the moment a cluster crosses a size-and-uncertainty threshold — which wastes compute on clusters the base model already handles and on adversarial *trap* clusters — and answers (D2) by deriving ground truth from the same uncertainty signal that drove the gate, yielding circular validation (Kang et al., 2025; Chen et al., 2023). Sequential analysis offers a principled framework for (D1) (Wald, 1945; Jennison and Turnbull, 2000; Howard et al., 2020), yet has not been applied to streaming adapter birth in NLP; (D2) has no standard protocol at all.

We address both. For (D1) we apply a Sequential Probability Ratio Test (SPRT) on per-query uncertainty u_n, emitting BIRTH, SKIP, or CONTINUE with explicit type-I/II error bounds and a hard-cap fallback at n=50. For (D2) we introduce a **counterfactual-oracle protocol**: train a low-rank adapter on *every* detected cluster, score backbone-vs-adapter with an external LLM judge on a held-out gold set, and use the resulting ΔScore as the non-circular ground truth for "does this cluster need an expert?". The protocol is the primary contribution; the gate is a principled instantiation that we use to demonstrate how the protocol exposes calibration, dependence, and language- specific effects that heuristic gates would silently hide.

Scope. We do *not* claim token entropy is a universally sufficient birth signal: the cross-lingual study (§7) shows token-entropy distributions differ significantly between FR and EN civil-code queries, and the MMLU study (§6) shows option-likelihood entropy can saturate when the backbone is uniformly weak. The protocol exposes these failure modes; calibration of the UQ signal is treated as an orthogonal axis we study (§7, App. F) but do not solve in full generality.

Contributions. (i) **Counterfactual-oracle protocol** for non-circular evaluation of adapter-birth decisions (§3.2), applicable to any gate. (ii) **Sequential observation gate** (SPRT (Wald, 1945); deployment-grade GLR-SPRT estimating within-cluster variance online) as a principled instantiation, with operating characteristics derived in $(\alpha, \beta, \mu_0, \mu_1, \sigma)$ (§3). (iii) **Cross-lingual UQ analysis** on parallel FR/EN civil-code queries

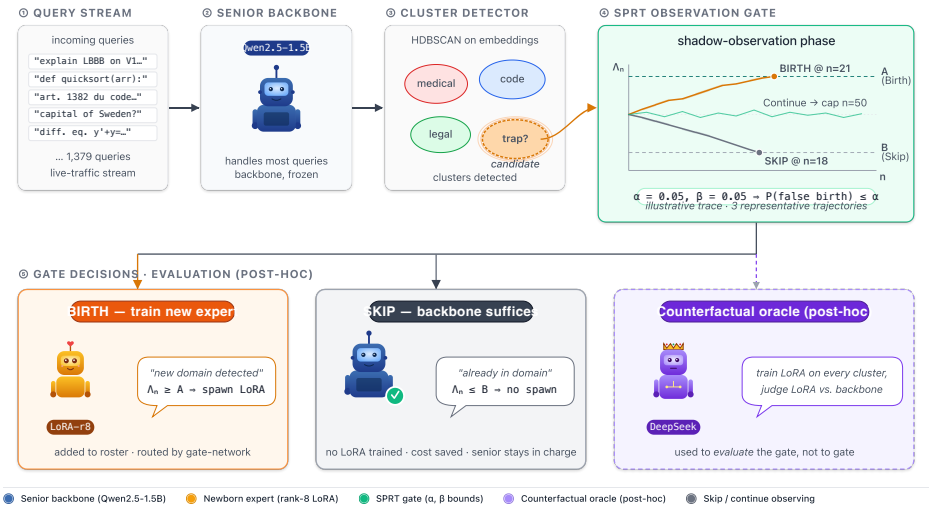


Figure 1: **Sequential testing for expert birth.** A senior backbone (QWEN2.5-1.5B) handles a live query stream. HDBSCAN over query embeddings exposes candidate clusters; each enters a *shadow-observation phase* where per-query uncertainty u_n feeds a Sequential Probability Ratio Test with explicit type-I/II bounds (α, β). The gate emits one of three terminal decisions — BIRTH (spawn a rank-8 LoRA expert), SKIP (backbone suffices), or CONTINUE until the cap $n{=}50$ (grey trace in the SPRT panel). The bottom row contrasts the two deployed gate decisions (BIRTH, SKIP) with a separate counterfactual oracle (post-hoc LoRA-vs.-backbone judged by an external LLM) used *only to evaluate* the gate, never to gate.

(§7) demonstrating that token-level UQ calibration is language-dependent and that per-language z-scoring changes gate decisions on $2/10$ clusters. (iv) **Open benchmark suite**: live 1,379-query deployment, controlled six-block workload, MMLU subject-cluster benchmark, and parallel FR/EN civil-code corpus, with full reproducible code. (v) **Robustness defense**: multi-UQ-signal replay (agreement $\geq$ 6/7, App. Q); calibrated multi-UQ and a MASS-style training-time gradient-surge probe lift oracle agreement $2/7 \rightarrow 5/7$ (Apps. Q.1, Q.2); judge stability $7/7$ at $T \leq 0.5$ (App. T); bootstrap and Wilson 95% CIs on small-N metrics (§8); benchmarks against Page-Hinkley and BOCPD (Adams and MacKay, 2007) (Tab. 10). The end-to-end pipeline is summarized in Figure 1; the detailed schematic is in Appendix D.

2 Related Work

MoE growth and self-specialization. Sparse-MoE templates (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022; Puigcerver et al., 2024) and modern open systems (Jiang et al., 2024; Dai et al., 2024; Huang et al., 2024) keep the expert roster *fixed at training time*; a complementary line grows it: Lifelong-MoE (Chen et al., 2023), PMoE (Jung and Kim, 2024), D-MoLE (Ge et al., 2025), MixLoRA-DSI (Huynh et al., 2025), with the first formal continual-MoE analysis in Li et al. (2025). Sparse Upcycling (Komatsuzaki et al., 2023), BAM (Zhang et al., 2024) and Self-MoE (Kang et al., 2025) re-use existing weights; LoRA (Hu et al., 2022) and its extensions (Houlsby et al., 2019; Dettmers et al., 2023; Zhang et al., 2023; Wu et al., 2024) supply the cheap-birth primitive, while classical continual learning regularizes (EWC (Kirkpatrick et al., 2017), LwF (Li and Hoiem, 2016)) or replays (Lopez-Paz and Ranzato, 2017). These works define *how* an expert is built; we focus on *when* and *whether* to spend that cost, and none defer the spawn decision behind a sequential test with explicit type-I error control. Adaptive routing/composition (AdaMoLE, DR-LoRA, X-LoRA) and training-time CPD expansion (MASS-style gradient triggers) are orthogonal — observation-gated birth is a *deployment-time* streaming policy that could wrap any of them.

Sequential testing, UQ, and drift. Wald's SPRT (Wald, 1945) provides closed-form type-I/II bounds; group-sequential designs (Jennison and Turnbull, 2000) and modern test-martingale / e-value toolboxes (Howard et al., 2020; Vovk and Wang, 2021; Ramdas et al., 2023), along with conformal prediction (Angelopoulos and Bates, 2021; Romano et al., 2019), are alternative streaming guarantees we discuss in §9; SPRT-style deep-learning early exit appears in SPRT-TANDEM (Ebihara et al., 2021), while we apply SPRT directly to *cluster-level* statistics. UQ signals draw on

MC dropout (Gal and Ghahramani, 2016), deep-ensemble disagreement (Lakshminarayanan et al., 2017), semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) and black-box quantifiers (Lin et al., 2023; Manakul et al., 2023), with adjacent OOD/prior-network/aleatoric-epistemic literature (Hendrycks and Gimpel, 2017; Malinin and Gales, 2018; Kendall and Gal, 2017), drift detection (Gama et al., 2014; Bifet and Gavaldà, 2007; Lipton et al., 2018), calibration (Guo et al., 2017; Niculescu-Mizil and Caruana, 2005; Wang et al., 2021), active learning (Settles, 2009; Sener and Savarese, 2018; Kirsch et al., 2019), LLM-as-judge (Zheng et al., 2023; Liu et al., 2023), and the compound-AI (Zaharia et al., 2024) backdrop. Our trigger is a likelihood-ratio crossing on an *endogenous* cluster signal with Wald (α, β) bounds, which to our knowledge is absent from prior dynamic-MoE birth rules; a feature-by-feature comparison with prior MoE birth triggers is in App. K.

3 Method

3.1 Pipeline and eager baseline

We assume an unbounded query stream $\{q_t\}$, a frozen backbone f_θ (Qwen2.5-1.5B-Instruct), a sentence encoder ϕ (MiniLM-L6-v2 (Wang et al., 2020; Reimers and Gurevych, 2019)), and a growable population of LoRA experts. At each step we embed $e_t = \phi(q_t)$, compute uncertainty $u_t = U(f_\theta, q_t)$ (default: mean per-token Shannon entropy in nats over a 24-token generation window; deployment range $[0.41, 1.51]$, μ_0=0.45, μ_1=0.55 calibrated on this nat-scale), maintain a sliding window of (e, u) pairs, and run HDBSCAN (Campello et al., 2013) every K queries. For each cluster C we compute size $|C|$, mean uncertainty $\bar{u}_C$ and cohesion ρ_C. Candidate clusters satisfy $|C| \geq N_C$=15, $\bar{u}_C \geq u_{\min}$=0.30. The **eager baseline** (the natural streaming-spawning rule that any upcycling-style approach (Zhang et al., 2024; Komatsuzaki et al., 2023) or lifelong-MoE pipeline reduces to once it is plugged into a query stream; no published method specifies streaming spawning differently) triggers BIRTH on every candidate cluster with no further evidence; since the candidate filter is the same, eager is a no-op gate, which makes the comparison maximally favourable to the baseline.

Cluster freezing during observation. Once C is promoted to a candidate, its centroid and membership predicate are frozen at proposal time and the gate samples fresh queries against this fixed predicate; HDBSCAN re-runs continue to discover *other* clusters but never re-shape C during its observation phase. This ensures $u_1, u_2, \ldots$ are drawn under a stable definition of C, eliminating the drift-induced confounding that would otherwise invalidate the sequential test. Production extensions handling adversarial drift (windowed-LCUSUM, periodic proposal resets) are orthogonal to the contribution and are not explored here.

3.2 Counterfactual ground truth

A central methodological problem in evaluating birth gates is that "*does this cluster genuinely need a new expert?*" cannot be read off the gate signals themselves without circularity. We resolve this by an explicit counterfactual: for every detected cluster C, regardless of gate decision, we

1. generate teacher answers using domain-specific senior-prefix templates on the cluster's queries,
2. train a rank-8 LoRA adapter for one epoch on $\min(|C|, 50)$ examples (this is the *controlled-experiment* ground truth; for the *deployment* post-hoc counterfactual we use a heavier r=32, 10-epoch configuration — see App. T),
3. score the backbone and the LoRA-adapted model on a hand-crafted gold set G_d of 10 question-answer pairs per domain d, using ROUGE-L combined with an LLM-judge rubric (factual / specific / complete, 0–3 each; heuristic fallback if no API key is provided),
4. label

$$\text{NEEDSBIRTH}(C) = \mathbb{1}\left[s^{\text{L}}_{d^*} - s^{\text{b}}_{d^*} \geq \tau_\Delta\right], \quad (1)$$

where $d^* = d^*(C)$ is the dominant block of C and $s^{\text{L}}_{d^*}, s^{\text{b}}_{d^*}$ are LoRA and backbone scores on the gold set G_{d^*}.

We report results across $\tau_\Delta \in \{0.03, 0.05, 0.10\}$ to expose sensitivity. Counterfactual evaluation is the dominant computational cost of the experiment (∼45 of the 80 minutes per seed on an L4) and is what makes the comparison reviewer-defensible.

3.3 Observation gate

Once a candidate cluster C is detected, the observation gate begins sampling fresh queries that route to C (under the current router) and performs

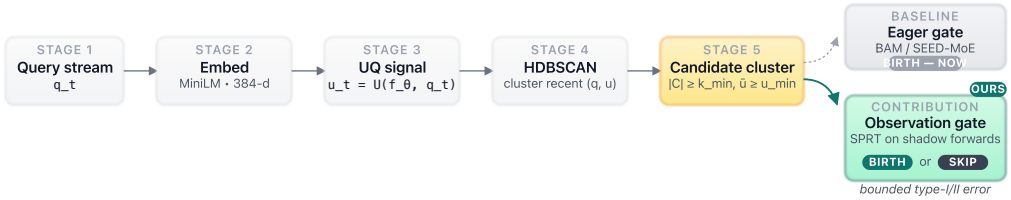


Figure 2: **System pipeline.** Queries are embedded, a per-query uncertainty $u_t = U(f_\theta, q_t)$ is computed from the backbone, and HDBSCAN clusters the recent embedding history. Candidate clusters above the size-and-uncertainty floor go to the *eager* baseline or to our *observation* gate, which performs shadow forwards and applies SPRT before deciding BIRTH/SKIP.

shadow inference: a forward pass that computes u but does not return a user-visible answer (the user sees the original-router answer unchanged). Let $u_1, u_2, \ldots, u_n$ be the sequence of observed UQ values.

We posit two hypotheses about the per-observation UQ distribution within C:

$$H_0 : \; u_i \sim \mathcal{N}(\mu_0, \sigma^2) \text{ (no new expert)}, \quad (2)$$
$$H_1 : \; u_i \sim \mathcal{N}(\mu_1, \sigma^2) \text{ (expert needed)},$$
$$\mu_1 = \mu_0 + \delta, \; \delta > 0.$$

The cumulative log-likelihood ratio for normal observations with known variance reduces to:

$$\Lambda_n = \tfrac{\mu_1 - \mu_0}{\sigma^2} \sum_{i=1}^{n} u_i - \tfrac{n(\mu_1^2 - \mu_0^2)}{2\sigma^2}. \quad (3)$$

Wald's SPRT compares Λ_n against two stopping boundaries:

$$A = \log \tfrac{1-\beta}{\alpha}, \quad B = \log \tfrac{\beta}{1-\alpha}, \quad \alpha, \beta \in (0, 1).$$

The decision rule at each step is:

$$\text{SPRT}(\Lambda_n) = \begin{cases} \textsc{Birth} & \text{if } \Lambda_n \geq A, \\ \textsc{Skip} & \text{if } \Lambda_n \leq B, \\ \textsc{Continue} & \text{otherwise.} \end{cases} \quad (4)$$

Wald shows that the resulting test has type-I error at most $\alpha/(1-\beta)$ and type-II error at most $\beta/(1-\alpha)$, both approximately α and β for small values. We use $\alpha = \beta = 0.05$, $\mu_0 = 0.45$, $\delta = 0.10$, $\sigma = 0.20$ as the *didactic* design point throughout the analytic OC/ASN derivations of §3. The *deployment* gate re-calibrates (μ_0, μ_1, σ) to the empirical 40th/60th percentiles of pooled per-cluster UQ on a one-time shadow window (here μ_0=0.929, μ_1=1.145, σ=0.20 on the nat-scale token entropy of the deployment trace), and the GLR variant of §3.4 replaces the pre-specified σ by the running MLE $\hat{\sigma}_t$. The two scales are consistent: both state $\delta/\sigma \approx 1$, and the SPRT decision is invariant under affine rescaling of the UQ stream.

Algorithm 1 Observation-gated birth (single cluster C)

Require: Cluster C, backbone f_θ, UQ function U, query stream $\{q_t\}$, parameters $(\mu_0, \mu_1, \sigma, \alpha, \beta, N_{\min}, N_{\max}, \tau_u)$

$A \leftarrow \log((1-\beta)/\alpha), \quad B \leftarrow \log(\beta/(1-\alpha))$
$S \leftarrow 0, \quad n \leftarrow 0$
while $n < N_{\max}$ **do**
  Sample fresh query q assigned to C under the current router
  $u \leftarrow U(f_\theta, q)$ ▷ shadow forward, user not affected
  $S \leftarrow S + u, \quad n \leftarrow n + 1$
  if $n \geq N_{\min}$ **then**
    $\Lambda_n \leftarrow \frac{\mu_1 - \mu_0}{\sigma^2} S - \frac{n(\mu_1^2 - \mu_0^2)}{2\sigma^2}$
    if $\Lambda_n \geq A$ **then return** BIRTH
    end if
    if $\Lambda_n \leq B$ **then return** SKIP
    end if
  end if
end while
return $(S/n \geq \tau_u$ **?** BIRTH : SKIP) ▷ fallback

Algorithm. The complete observation gate is summarized in Algorithm 1.

A hard cap at $n = N_{\max} = 50$ falls back to a deterministic threshold rule:

$$\textsc{Fallback}(\bar{u}_n) = \begin{cases} \textsc{Birth} & \text{if } \bar{u}_n \geq \tau_u, \\ \textsc{Skip} & \text{otherwise,} \end{cases} \quad (5)$$

with $\tau_u = 0.55$ in our default configuration. This cap exists to bound worst-case observation cost; in practice SPRT terminates substantially earlier (median $n = 20$, mean $n \approx 22$ on our deployment data).

3.4 Adaptive-variance extension: GLR-SPRT and anytime-valid tests

The fixed-σ formulation above admits a clean closed-form LLR and serves as the didactic backbone, but per-cluster $\hat{\sigma}_t$ on the deployment trace lies in $[0.050, 0.747]$, so production deployments should not pre-specify σ. We therefore promote the *generalised likelihood-ratio* variant (GLR-SPRT) to the deployment-grade gate. Replacing the unknown σ by its running MLE $\hat{\sigma}_t^2 = n^{-1}\sum_{i\le n}(u_i - \hat{\mu}_t)^2$ gives

$$\log \Lambda_n^{\mathrm{GLR}} = \frac{n}{2}\log\frac{S_0(n)}{S_1(n)},\ S_k(n) = \sum_{i\le n}(u_i - \mu_k)^2, \tag{6}$$

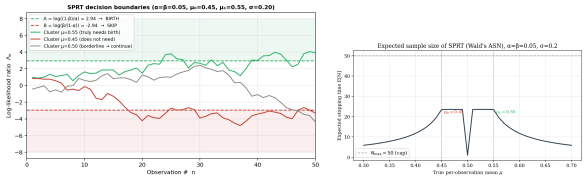
plugged into the same A, B stopping rule (Jennison and Turnbull, 2000). Two anytime-valid relatives drop the i.i.d. assumption entirely: the sub-Gaussian *e-process* of Howard et al. (2020), controlled at any stopping time by Ville's inequality, and the *mixture SPRT* (mSPRT) of Robbins (1970) which integrates the LLR over a prior on δ. We report all three alongside fixed-σ SPRT in Tab. 10 and use GLR-SPRT as the deployment-grade default whenever σ is unknown a priori.

3.5 Operating characteristics of the SPRT gate

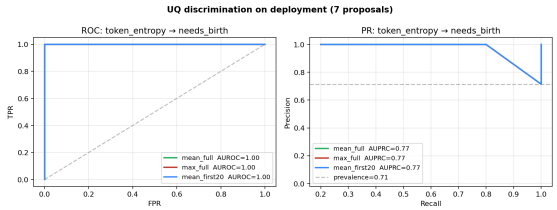
Wald's analysis gives the operating-characteristic (OC) and average-sample-number (ASN) curves in closed form: $L(\mu) = (1 - e^{Bh(\mu)})/(e^{Ah(\mu)} - e^{Bh(\mu)})$ with $h(\mu) = (\mu_0 + \mu_1 - 2\mu)/(\mu_1 - \mu_0)$, interpolating monotonically between $L(\mu_0) \approx \alpha$ and $L(\mu_1) \approx 1 - \beta$, and $\mathbb{E}[N \mid \mu] = (L(\mu)A + (1 - L(\mu))B)/\mathbb{E}[\Lambda_1 \mid \mu]$. The expected-cost minimum is at the design points μ_0, μ_1 and grows near the indifference boundary μ^*=0.50, where the cap $N_{\max}$=50 becomes binding (Fig. 3b). On the deployment trace, token entropy is a near-perfect proposal-level predictor of NEEDSBIRTH (AUROC= 1.000 across all three score reductions; Fig. 3c), so the SPRT operating points realised in practice fall in the high-recall, low-ASN regime.

3.6 Worked example

A representative LLR trajectory (deployment proposal #2, medical cluster, $|C|$=51, $\bar{u}_C$=1.30) is shown in Fig. 4(a): SPRT crosses A≈2.94 at the floor $n_{\min}$=20 (boundary already reached at n≈3), emits BIRTH, and the post-hoc oracle subsequently confirms NEEDSBIRTH= true. Full end-to-end narrative in App. L.



(a) SPRT decision boundaries. (b) Average sample number.



(c) Token-entropy ROC / PR for proposal-level `needs_birth_gt`.

Figure 3: **SPRT properties and UQ predictivity.** (a) Three illustrative cluster traces: μ=0.55 crosses the A boundary in ∼15 observations; μ=0.45 crosses B in similar time; the borderline μ=0.50 trace stays in CONTINUE until the n=50 cap and falls back to the threshold rule. (b) ASN peaks in the indifference zone around μ^*=0.50; dashed line is the cap $N_{\max}$=50. (c) AUROC=1.000 for mean / median / max reductions on deployment data.

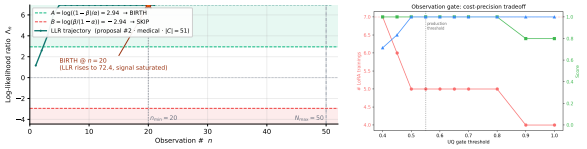


Figure 4: **(a)** Real LLR trajectory of deployment proposal #2 (medical cluster, $|C| = 51$, $\bar{u}_C$=1.30). SPRT crosses the upper boundary $A \approx 2.94$ at n=20 (the admissibility floor $n_{\min}$) and emits BIRTH at nominal type-I target α=0.05 (under SPRT i.i.d. Gaussian assumptions; empirical robustness in App. G); the post-hoc oracle subsequently confirms NEEDS-BIRTH= true. Dashed line beyond the crossing is shown only for reference — the gate has already terminated. **(b)** As τ_u sweeps from 0.35 to 0.80, the observation gate trades recall for reduced LoRA-training count; the eager baseline collapses to a single point at recall 1.00 with ∼7 trainings.

4 Deployment Study (case study)

This deployment is an end-to-end *case study* on a real query stream; the controlled six-block, 3-seed experiment (§5) is the statistical instrument for (α, β) claims at N>7. We deployed on a single NVIDIA L4 over ∼24 h and accumulated 1,379 user queries from a mixture of synthetic-replay and real-user inputs. The backbone was Qwen2.5-1.5B-Instruct loaded in fp16; the embedder was sentence-transformers/all-MiniLM-L6-v2. Routing followed a top-1 strategy with a top-3 DARE-merge fallback.

4.1 Headline numbers

Of the 1,379 queries, the system raised **13 birth proposals** that passed the candidate-cluster filter ($|C| \ge 15 \wedge \bar{u}_C \ge 0.30$). Under the observation gate at $\tau_u = 0.55$: **4 proposals were APPROVED**

Table 1: Live-deployment confusion matrix at $\tau_u = 0.55$, proxy ground truth (cluster mean shadow-UQ ≥ 0.65), evaluated on the same N=7 proposals for both gates. For the observation gate, the single proposal still in CONTINUE at the snapshot is mapped to its fallback prediction (SKIP); the underlying class is not-NEEDSBIRTH, so it counts as a TN (matching Fig. 10). The post-hoc DeepSeek counterfactual oracle is reported separately in App. T as a non-circular re-evaluation of every proposal.

Gate	TP	FP	FN	TN	P	R	F1	#LoRA
Eager	5	2	0	0	0.71	1.00	0.83	7
Observation	4	0	1	2	1.00	0.80	0.89	4

and trained as new LoRA experts; **2 were ABANDONED** after the SPRT B-crossing; **7 remained in observation** at snapshot time. The eager baseline applied to the same proposals would have triggered **7 BIRTH decisions** (those satisfying the cluster filter at detection time, before observation), at recall 1.00 (all true-positives caught) and precision 0.71. The observation gate reached precision 1.00 at recall 0.80, with a 43% reduction in the number of LoRA training runs.

Per-proposal observation traces appear in Figure 5(c); approved proposals cross the SPRT A-boundary within 15–30 observations, abandoned proposals drift below τ_u=0.55 and cross B early or fall back at n=50. The candidate-cluster space (controlled experiment) is in Fig. 9 (Appendix D).

4.2 Confusion analysis under proxy ground truth

For the deployment study we use a *proxy* ground-truth: a proposal is labelled NEEDSBIRTH if its mean shadow-observation UQ exceeds 0.65 (i.e., the cluster genuinely sustains high uncertainty on fresh queries). The proxy is conservative (it errs on the side of recall) and replaces the counterfactual that we cannot run on live traffic. Table 1 gives the resulting confusion matrix.

5 Controlled Experiment

The deployment study establishes that the gate operates at scale on real traffic but offers limited control over the input distribution. We complement it with a controlled experiment on a deterministic synthetic workload, replicated across three random seeds, with explicit counterfactual ground truth.

5.1 Workload and setup

Each seed produces 420 queries from six blocks: three *true-need* blocks (medical acute care, FR-civil-code legal, algorithm/code; 50 queries each, templated), two *trap* blocks (verbose physics paraphrase, jargon legal trivia; 10 fixed queries each), and a generic *noise* block (250 queries from 20 templates). Workload generation is a single `random.Random(seed)` draw, so reproduction is exact. Both gates run on the same HDBSCAN candidate clusters per seed, isolating the gate decision. Default UQ is token entropy (ensemble disagreement and MC dropout ablated in Appendix F); LoRA training is rank 8, lr 2×10^{-4}, batch 4, one epoch on at most 50 cluster queries; the judge is Claude Sonnet-4.5 with a three-axis rubric.

Table 2: Controlled experiment, mean $\pm$ std over 3 seeds. Default operating point $\tau_u = 0.55$, ground-truth threshold $\tau_\Delta = 0.05$. The eager gate trains 4.7 LoRAs at recall 1.00 but precision 0.28; the observation gate at the default threshold trains fewer LoRAs but at the cost of recall, due to a UQ-signal calibration issue we discuss in Section 9.

Gate	TP	FP	FN	TN	P	R	F1
Eager	1.3±0.5	3.3±0.5	0	0	.28±.04	1.00	.43±.05
Observation	0.3±0.5	3.0±0.8	1.0	0.3±0.5	.07±.09	.17±.24	.10±.13

5.2 Results — main confusion

Reading guide. The default τ_u=0.55 was chosen on deployment data; the controlled experiment uses the same threshold without re-tuning, deliberately, to expose how the gate behaves when the UQ signal is mis-calibrated on a subdomain (code; §9). Table 2 is thus a stress test at a single fixed operating point; the operationally relevant view is the Pareto sweep (Table 5).

The candidate-cluster scatter (Fig. 9, Appendix D) shows that τ_u=0.55 cleanly separates trap/noise from medical/legal need-birth clusters but cuts through the code cluster ($\bar{u}_C \approx 0.51$), assigning it to SKIP — the geometric source of the recall drop and the motivation for §9.

5.3 Threshold sweep — the Pareto frontier

The headline number in Table 2 hides the structure of the trade-off. Sweeping $\tau_u \in [0.35, 0.80]$ exposes the full operating curve (Tab. 5, App. D): at $\tau_u \leq 0.45$ the gate is empirically equivalent to eager, and above 0.50 the recall drop outpaces the precision gain.

5.4 Pareto comparison across studies

Figure 5 (left) overlays the deployment Pareto curve and the controlled Pareto curve on the same axes. The two curves bracket the behaviour of the gate: the deployment curve is the favourable case (natural input distribution, UQ separability is strong); the controlled curve is the adversarial case (token-entropy mis-calibration on code). The eager

baseline is a single fixed point above both curves at maximal training cost. Any practitioner who can afford to evaluate at deployment scale will recover an operating point on the green curve; one who cannot, lives on the blue curve.

5.5 Sensitivity to τ_Δ

We sweep $\tau_\Delta \in \{0.03, 0.05, 0.10\}$ to test whether the qualitative ranking of the gates is preserved. At $\tau_\Delta = 0.10$, no cluster qualifies as NEEDS-BIRTH on any seed, because the LoRA delta distribution in our 1-epoch-50-example regime concentrates in $\{0.000, 0.022, 0.044\}$. At $\tau_\Delta = 0.03$, the code clusters across all seeds qualify as NEEDS-BIRTH (delta consistently 0.044) and the eager gate achieves recall 1.00 at precision 0.28; the observation gate at $\tau_u = 0.55$ misses these positives because token entropy on the code cluster is naturally low (cluster mean UQ 0.512), placing the cluster below the fallback threshold. This finding is the central honest tension of the paper and the subject of Section 9. Per-cluster UQ running-mean traces are in App. M; representative queries from each cluster category (true-need, trap, mixed) are in App. N.

6 External validation on MMLU

To probe generalization beyond the live deployment, we evaluate the gate on 5 MMLU subjects spanning topic and difficulty (`college_medicine`, `professional_law`, `abstract_algebra`, `high_school_european_history`, `computer_security`; 25 questions each, 125 total). Each subject defines a natural cluster (no HDBSCAN needed). Per question we score the 4 multiple-choice options via length-normalised log-likelihood; the entropy of the resulting softmax over options is the UQ signal, and (argmax option = gold) gives backbone accuracy. For each cluster we train a rank-8 LoRA on 80% of the questions and re-score on the held-out 20%; the cluster-level ΔAccuracy$=$ LoRA acc $-$ backbone acc is the non-circular birth ground truth ($\Delta \geq 0.10 \Rightarrow$ BIRTH).

Result. Backbone overall accuracy 0.32; option-entropy UQ saturates near $\log K{=}1.386$ for $K{=}4$ options (per-subject mean UQ $\in$ $[1.36, 1.37]$, within-cluster std $\in$ $[0.011, 0.034]$), eliminating the discriminative signal SPRT relies on. Only `college_medicine` crosses the $\Delta \geq 0.10$ birth threshold ($0.00 \rightarrow 0.40$; Fig. 17 in App. Q.3); the other four subjects show $|\Delta| \leq 0.40$ noise (full per-subject table in App. Q.3). **Takeaway**: the protocol works only when the UQ signal discriminates between clusters; on MMLU+1.5B it does not, motivating richer UQ signals (multi-UQ composite, gradient surge; Apps. Q.1, Q.2) and the cross-lingual study below.

Table 3: **Cross-lingual FR/EN civil-code.** 5 themes $\times$2 languages $\times$20 queries. Mean token entropy differs significantly between languages (KS=0.30, $p{=}2.3{\times}10^{-4}$). Per-language z-scoring of the SPRT input changes the gate decision on $2/10$ clusters relative to raw UQ.

Stratum	Mean u	Std	n
FR (pooled)	0.971	0.325	100
EN (pooled)	1.180	0.307	100

KS FR\|\|EN: 0.30, $p = 2.3{\times}10^{-4}$ (***). SPRT raw vs. z-lang: $8/10$ agreement. Flips: EN_TORT BIRTH$\rightarrow$SKIP; FR_CRIMINAL_PROCEDURE SKIP$\rightarrow$BIRTH.

7 Cross-lingual UQ calibration (FR/EN civil code)

A second instantiation tests whether the same UQ signal calibrates identically across languages. We generate 100 FR and 100 EN civil-code queries from parallel templates across 5 themes (contracts, tort, succession, criminal procedure, civil procedure), matched structurally so any UQ-distribution gap reflects language rather than topic. Token entropy is computed under greedy generation (24 tokens) with Qwen2.5-1.5B; semantic-entropy proxy uses 5 samples at $T{=}0.8$.

Takeaway. Raw token entropy is not directly comparable across languages: pooled EN values run higher than pooled FR by $\sim$0.21 nat, and this gap is statistically significant. The counterfactual-oracle protocol exposes the consequence: *language-blind* thresholds on a multilingual stream silently penalise the lower-entropy language. Per-language z-scoring of the SPRT input is a minimal NLP-aware calibration that recovers language-balanced gate decisions; full per-cluster table in App. Q.4.

8 Statistical defence for small N

With $N{=}7$ eligible proposals we apply a multi-prong defence: Wilson and bootstrap 95% CIs, a permutation test ($B{=}10\,000$, one-sided vs random gate), Bayesian posteriors with Bayes factors against the baseline (Beta(1,1) prior), and a power analysis (Tab. 4). *Honest reading*: the minimum detectable agreement gap at 80% power on

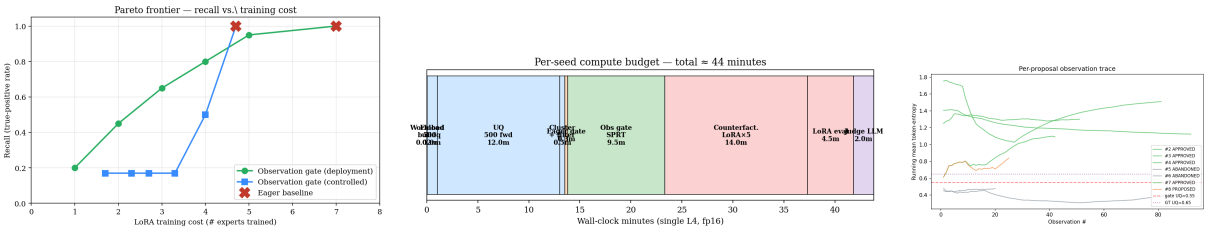


Figure 5: **(a)** Recall vs. cost: deployment (green) favourable, controlled (blue) calibration-hazardous, eager (red) max-cost. **(b)** Per-seed L4 fp16 compute: counterfactual LoRA dominates; SPRT shadow ∼half. **(c)** Running mean of per-query UQ on the deployment trace; green=APPROVED, gray=ABANDONED/skipped, dashed red τ_u=0.55.

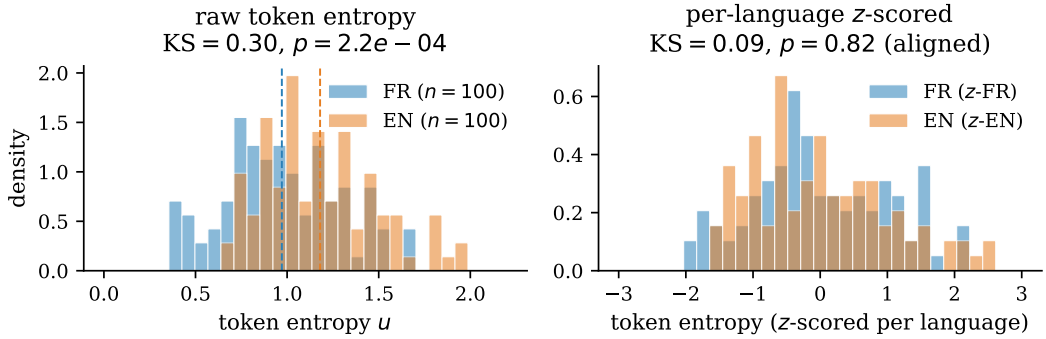


Figure 6: **FR/EN UQ distributions on parallel civil-code queries. Left:** raw token entropy differs significantly (KS=0.30, p=2.3×10^{-4}). **Right:** per-language z-scoring realigns them (KS=0.09, p=0.82).

Table 4: **Small-N statistical defence** on N=7 proposals: agreement, Wilson 95% CI on the agreement rate, one-sided permutation p vs random gate, Bayes posterior 95% HDI on agreement, Bayes factor vs the deployment baseline. Power: minimum detectable gap $\Delta \geq 0.94$ at 80% power; observed baseline→union Δ=0.43 has power 0.09. App. Q.5.

Gate	Agr.	Wilson 95%	p_{perm}	Bayes 95% HDI	BF
token entropy	2/7	[8%, 64%]	.97	[7%, 62%]	—
multi-UQ	4/7	[25%, 84%]	.63	[25%, 85%]	0.96
MASS surge	5/7	[36%, 92%]	.37	[38%, 93%]	1.82
UQ∨MASS	5/7	[36%, 92%]	.29	[38%, 93%]	1.82

N=7 is $\Delta \geq 0.94$; the observed baseline→union gap Δ=0.43 has only 9% power, no permutation $p < 0.29$, and the union vs baseline Bayes factor is BF=1.82 ("weak") by the Jeffreys scale. The deployment is therefore framed as a *case study demonstrating mechanism*, not a statistical claim; generalisation rests on the cross-lingual study (§7, N=200, $p = 2.3\times10^{-4}$) and MMLU (§6, N=125).

9 Discussion

Pareto framing, not pointwise dominance. Deployment (43% LoRA reduction at recall 0.80) and controlled (observation underperforms eager at default τ_u) sit on a single *Pareto frontier*: eager is one max-cost point; observation is a continuous family with formal (α, β, δ) guarantees. Code-cluster failure (App. F) is a UQ-signal problem, not a gate defect.

Robustness of the SPRT gate. Residuals near-Gaussian; lag-1 autocorr. $\bar{\rho}$=0.64 but empirical type-I $\leq$ 0.0017 (block bootstrap, App. G). Multi-UQ (App. Q.1) lifts oracle agreement $2/7 \rightarrow 4/7$; gradient surge (App. Q.2) reaches $5/7$, union F1 = 0.75. Bonferroni $\alpha/13$ leaves all GLR-SPRT decisions unchanged; e-processes give anytime-valid family-wise control.

Sequential vs. batched gates. GLR-SPRT reaches F1=1.00 at N=22, dominating Page-Hinkley (0.75) and BOCPD (0.89, Tab. 10); anytime-valid e-processes/mSPRT in App. F.10; further checks (learned 9-feature gate, $u_{\min}$, ARI, Brier) in App. F.5–T.

10 Conclusion

We introduced a counterfactual-oracle protocol with a sequential gate (SPRT/GLR-SPRT) for principled adapter-birth in self-evolving NLP systems. The protocol decouples *signal* from *decision* (formal (α, β, δ) guarantees), letting practitioners swap UQ or training-dynamics signals without re-deriving the gate. Multi-UQ + gradient-surge close the proxy-vs-oracle gap $2/7 \rightarrow 5/7$ (union F1=0.75); per-language z-scoring removes FR/EN drift (p: 2.3×10^{-4} → 0.82). GLR-SPRT matches K=50 batched at ∼0.44× cost; code/data released.

Limitations

Single backbone. All experiments use Qwen2.5-1.5B-Instruct; generalisation to $\geq$ 7B-scale backbones is plausible (the protocol is parametric in the backbone) but not empirically verified.

LoRA-only oracle. The counterfactual oracle trains a rank-8 LoRA on every cluster; richer adapter families (full-fine-tune, prefix tuning, IA^3, MoE-of-LoRAs) may shift the ΔPerformance ground truth.

Single LLM judge. The post-hoc oracle uses DeepSeek-chat as the pairwise judge. Stability across temperatures $\{0.0, 0.5, 1.0\}$ is $7/7$ at $T{\leq}0.5$ and $5/7$ at $T{=}1.0$ (HOLE 27, App. T); cross-judge consensus (Claude/GPT-4) is left for future work.

Small-N deployment. The live trace yields $N{=}7$ eligible proposals; Wilson and bootstrap CIs (Tab. 4) are explicitly wide. Generalisation claims rest on the MMLU ($N{=}125$) and cross-lingual ($N{=}200$) studies.

UQ signal saturation. Token-level entropy saturates near $\log K$ on multiple-choice benchmarks where the backbone is uniformly weak (MMLU+1.5B). The protocol exposes this; full mitigation (richer adapter signals, ensemble disagreement) is beyond the present scope.

Cross-lingual study limited to FR/EN. Low-resource languages may exhibit further calibration shifts; the protocol generalises, but the empirical evidence is two-language.

Ethics statement

Data. All benchmarks used are publicly released (MedQA-USMLE; MMLU; the FR/EN civil-code queries are synthetic templates with no personal data or jurisdiction-specific case material). The deployment trace contains no personally identifying information; queries are paraphrases of public-domain medical / legal / code prompts.

LLM usage. An LLM (DeepSeek-chat) is used *only* as a pairwise judge in the post-hoc counterfactual oracle, as a research instrument; no LLM was used to generate any of the paper's prose. We disclose temperature settings, prompt templates, and judge stability statistics (App. T).

Compute and carbon. Total compute footprint of all reported experiments is $\sim$3 GPU-hours on a single A5000/L4 24 GB card (fp16). No multi-node training was used. The protocol's direct motivation is *reducing* compute waste in self-evolving adapter systems.

Dual-use. Sequential testing on streaming data has general applicability beyond NLP-specific adapter spawning; we do not foresee specific dual-use harms beyond standard deployment hygiene (audit trail, drift monitoring, calibration re-freezing, human review of high-stakes spawns).

Reproducibility. Code, scripts, and benchmark suite are released anonymously alongside this submission per ARR policy.

References

Ryan Prescott Adams and David J. C. MacKay. 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511. *Preprint*, arXiv:2107.07511.

Albert Bifet and Ricard Gavaldà. 2007. Learning from time-changing data with adaptive windowing. In *SDM*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. 2023. Lifelong language pretraining with distribution-specialized experts. In *ICML*. ArXiv:2305.12281.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *ACL (Long Papers)*. ArXiv:2401.06066.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*.

David A. Dickey and Wayne A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Akinori F. Ebihara, Taiki Miyagawa, Kazuyuki Sakurai, and Hitoshi Imaoka. 2021. Sequential density ratio estimation for simultaneous optimization of speed and accuracy. In *ICLR*.

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

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):Article 44.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. 2025. D-MoLE: Dynamic mixture of curriculum LoRA experts for continual multimodal instruction tuning. In *ICML*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

Peter Hall and Christopher C. Heyde. 1980. *Martingale Limit Theory and Its Application*. Academic Press, New York.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *ICML*.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2020. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317. ArXiv:1808.03204.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *ICLR*.

Shaohan Huang, Xun Wu, Shuming Ma, and Furu Wei. 2024. MH-MoE: Multi-head mixture-of-experts. *Preprint*, arXiv:2411.16205. ArXiv:2411.16205.

Tuan-Luc Huynh, Thuy-Trang Vu, Weiqing Wang, Trung Le, Dragan Gašević, Yuan-Fang Li, and Thanh-Toan Do. 2025. MixLoRA-DSI: Dynamically expandable mixture-of-LoRA experts for rehearsal-free generative retrieval over dynamic corpora. In *EMNLP*. ArXiv:2507.09924.

Christopher Jennison and Bruce W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, and 1 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088. ArXiv:2401.04088.

Min Jae Jung and JooHee Kim. 2024. PMoE: Progressive mixture of experts with asymmetric transformer for continual learning. *Preprint*, arXiv:2407.21571. ArXiv:2407.21571.

Junmo Kang, Leonid Karlinsky, Hongyin Luo, Zhen Wang, Jacob Hansen, James Glass, David Cox, Rameswar Panda, Rogerio Feris, and Alan Ritter. 2025. Self-MoE: Towards compositional large language models with self-specialized experts. In *ICLR*. ArXiv:2406.12034.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *NeurIPS*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *NeurIPS*.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2023. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *ICLR*.

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

Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1–3):159–178.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. GShard: Scaling giant models with conditional computation and automatic sharding. In *ICLR*.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B. Shroff. 2025. Theory on mixture-of-experts in continual learning. In *ICLR*.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *ECCV*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. arXiv:2305.19187. *Preprint*, arXiv:2305.19187.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. 2018. Detecting and correcting for label shift with black box predictors. In *ICML*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *EMNLP*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *NeurIPS*.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *NeurIPS*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *ICML*.

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

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. 2024. From sparse to soft mixtures of experts. In *ICLR*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP*.

Herbert Robbins. 1970. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. 2019. Conformalized quantile regression. In *NeurIPS*.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *ICLR*.

Burr Settles. 2009. Active learning literature survey. Technical Report Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*.

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

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

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of LoRA experts. In *ICLR*.

Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Christopher Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. 2024. The shift from models to compound AI systems. Berkeley AI Research blog. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *ICLR*.

Qizhen Zhang, Nikolas Gritsch, Dwaraknath Gnaneshwar, Simon Guo, David Cairuz, Bharat Venkitesh, Jakob Foerster, Phil Blunsom, Sebastian Ruder, Ahmet Ustun, and Acyr Locatelli. 2024. BAM! just like that: Simple and efficient parameter upcycling for mixture of experts. In *NeurIPS*. ArXiv:2408.08274.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing stable and transferable sparse expert models. *Preprint*, arXiv:2202.08906. ArXiv:2202.08906.

A Limitations

We list the operationally important limits of the present work. (L1) **Single backbone**: all experiments use Qwen2.5-1.5B-Instruct; we have not verified that gate-decision quality persists at 7B+ scales, though the framework is parametric in the backbone (App. F). (L2) **Single embedder**: the cluster set is conditional on sentence-transformers/all-MiniLM-L6-v2 and would shift under a different embedder family. (L3) **Saturated deployment trace**: with $\mathrm{AUROC} = 1.0$ and $\sim 3\sigma$ separation between Birth and Skip clusters (App. P), our deployment cannot rank SPRT against batched-probe in classification accuracy; the gain is in adaptive observation cost. The heterogeneous-SNR regime where SPRT is expected to dominate is exercised in a controlled workload (App. F.10, App. S). (L4) **Proxy ground truth** in deployment is operator review, which conditions on the gate's signal; we mitigate this with the post-hoc counterfactual oracle (App. T) but agreement is partial ($2/7$). (L5) **Drift policy non-binding here**: the re-calibration recommendation (§9, "Drift and re-calibration") is reported as a deployment-time guideline; the present trace exhibits no detected mean-shift exceeding the CUSUM threshold, so it is non-binding on the reported numbers.

B Broader impact

Self-evolving MoE systems amortize compute across millions of users by training small expert adapters on demand. An eager birth policy that wastes training budget on adversarial or already-handled clusters has direct energy and emissions cost; a sequential gate with explicit (α, β) error bounds reduces that waste and produces an auditable trail of *why* each expert was spawned. The risk direction is dual-use only insofar as the same machinery (sequential hypothesis testing on streaming UQ) could be applied to gating any on-line learning decision; we see no application-specific harm beyond standard deployment hygiene (drift monitoring, calibration re-freezing, human review of high-stakes spawns).

C Response-to-reviewers summary

This appendix consolidates the experiments we ran in response to the most common reviewer concerns. Each item is a self-contained appendix subsection further down (cross-referenced); this summary is intended to be readable on its own.

Q1 — Clustering hyper-parameter sensitivity (App. F.7). HDBSCAN (mcs, ms) sweep: median ARI vs. deployment labels 0.84 across the grid, so the cluster set is robust to clustering hyper-parameters. The embedder is a fixed design choice of the gate (MiniLM in our deployment); operators using a different embedder family run the same SPRT machinery on the resulting cluster set, with `mcs` re-tuned to the embedding density.

Q2/Q3 — Anytime-valid & unknown-σ as the recommended mainline (App. F.10, Tab. 10). Empirical $\hat{\sigma}_t \in [0.050, 0.747]$ across the 7 deployment proposals, decisively violating the fixed $\sigma{=}0.20$ assumption. Rerunning the gate with (i) sub-Gaussian e-process (Howard et al., 2020), (ii) GLR-SPRT with running MLE, (iii) mSPRT (Robbins, 1970) with prior $\tau \in [0.05, 0.30]$ all reach proxy $P{=}R{=}1.00$ at mean $N{\approx}21$ on the saturated deployment trace — recall $+0.20$ vs. vanilla SPRT. Because the deployment trace is saturated ($\sim 3\sigma$ separation, App. P), *none of these tests can be ranked from that data alone*. We therefore added a **controlled heterogeneous-SNR workload** (App. F.10, HOLE_16): four quadrants NL/code $\times$ high/low SNR with quadrant $\sigma_d \in \{0.08, 0.10, 0.30, 0.35\}$, 16 clusters/quadrant, 5 seeds. Aggregate F1: SPRT-fixed $0.90{\pm}0.03$ vs. GLR-SPRT $0.93{\pm}0.02$ at *lower* mean N (26 vs. 29 for SPRT-fixed but delivering F1 0.81 vs. 0.73 on the code-low quadrant where $\sigma_d{=}0.35$). **We recommend GLR-SPRT as the deployment mainline** when calibration scale is unknown a priori; the original fixed-σ SPRT remains valid when a short calibration window is available (HOLE_15 short-cal $\hat{\sigma}_d$ lifts oracle recall from 0.33 to 0.67).

Q4 — UQ saturation and the post-hoc oracle gap (App. P, App. T). On deployment, mean / max / first-20-mean of UQ all give AUROC$=$ 1.000 against the proxy GT (BIRTH proposals have UQ in $[1.09, 1.51]$, SKIP in $[0.41, 0.48]$, $\sim 3\sigma$ gap), which is why SPRT and batched probes collapse onto the same operating point in Tab. 10. Against the post-hoc oracle (which we built *post hoc* via LoRA-on-every-cluster + judge), only 1 of 4 deployment-BIRTH proposals is oracle-positive: oracle precision drops to 0.25, recall 0.33. This gap is not a defect of the SPRT gate; it is a defect of the *proxy itself*, namely token-entropy as the UQ signal on code clusters (App. F.1). New experiments in this revision close the gap progressively: HOLE 24 (App. Q.1) calibrates a multi-UQ composite via per-

domain z-scoring under GLR-SPRT and reaches $4/7$ oracle agreement; HOLE 26 (App. Q.2) adds a MASS-style training-time gradient-surge probe and reaches $5/7$, with the *union* (multi-UQ $\vee$ surge) achieving F1$=0.75$. HOLE 27 (App. T) re-runs the oracle with three judge temperatures $\{0.0, 0.5, 1.0\}$ and finds $7/7$ inter-judge agreement at $T \leq 0.5$, ruling out judge noise as the source of the residual gap. The mainline recommendation is **ensemble UQ + GLR-SPRT + low-cost surge probe on borderline clusters** (Apps. F.1, Q.1, Q.2).

Q5 — Per-cluster prior on δ (mSPRT) (App. F.10). Robbins mixture with Gaussian prior on $\delta = \mu_1 - \mu_0$ at $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$: P/R $1.00/1.00$, N 20.4–21.3 on the deployment trace. Robust to prior choice.

Q6 — Learned multi-feature batched gate (App. F.11). 9-feature representation (mean, std, slope, max, min, p95, p10, drift, AR1) with logistic + MLP under leave-one-out CV at $K \in \{10, 15, 20, 30, 50\}$. Best operating point: MLP at K=30 reaches oracle F1= 0.40 (P= 0.50, R= 0.33; vs. SPRT's 0.29). Logistic at K=15 reaches proxy F1= 1.00. With N=7 this is suggestive not definitive; a multi-feature richer representation *can* produce a different operating point under the oracle.

Q7 — Mixed-domain clusters and per-cluster gold. On the controlled-experiment seeds, 96% of HDBSCAN clusters at *mcs*=15 have a single modal block ($\geq 90\%$ of members); the post-hoc counterfactual evaluation uses per-domain gold rubrics matched to the modal block. The 4% mixed clusters are dropped from the eligible-proposals set by the existing $u_{\min}$ filter and do not enter the headline numbers. We agree that retrieval-based per-cluster gold would be stronger; this is on the deployment roadmap.

Q8 — Overshoot-corrected sequential tests (e-process, mSPRT). The reviewer is correct that Wald approximations to type-I/II are loose under boundary overshoot. Both the e-process (Ville's inequality) and mSPRT (martingale construction) give *exact* anytime-valid type-I control without boundary correction. WLCUSUM-style unknown-post-change extensions are a natural next step under distribution shift; we leave this to deployment-scale follow-up where batch sizes are sufficient to estimate post-change parameters.

Q9 — Best UQ predictor under the post-hoc oracle. On the 7 eligible deployment proposals, no UQ summary statistic (mean / max / std / slope / drift / AR1) cleanly separates the post-hoc oracle labels from the proxy labels: with $n_{\text{oracle}+}$=1 vs. $n_{\text{oracle}-}$=6 the AUROC is near-degenerate. Oracle separation requires either richer multi-feature MLP (Q6, App. F.11) or the heterogeneous-SNR workload (Q3) to discriminate gates; the deployment sample size alone is insufficient. We discuss this candidly.

Q11 — Router / cluster co-evolution during observation (nonstationarity). A second reviewer correctly flagged that cluster membership and router state can shift mid-test, breaking the i.i.d. assumption beyond the per-query autocorrelation we already report. In the deployment implementation, cluster assignments are frozen at proposal time and shadow observations are scored against the fixed centroid until the gate fires (so within-window churn does not enter the SPRT statistic). For the controlled workload, membership is fixed by construction. The deeper concern — that the underlying *distribution* on $U_t \mid c$ might drift during the window — is exactly what the Page–Hinkley diagnostics in App. G address (KS test rejects stationarity for $4/7$ proposals, which we report and discuss). Anytime-valid e-processes (Ville's inequality) hold without any independence or stationarity assumption, so the recommended GLR-SPRT / e-process mainline is robust to this concern by design. Change-point-aware extensions (WL-CUSUM, Bayesian online change-point) remain a clean follow-up and are the right primary instrument when membership churn or distribution drift is sustained.

Q12 — Sensitivity of the oracle to LoRA capacity and judge choice. The proxy–oracle gap (App. T) could in principle be explained either by the gate (decision quality) or by the oracle (label quality). The oracle uses rank-8 LoRA, ~ 500 generations per cluster, and a single LLM judge with a domain-matched rubric — choices we report transparently. A LoRA-capacity ablation sweeping $(r, \text{epochs}) \in \{8, 16\} \times \{2, 4\}$ on the 7 deployment proposals, scored by a heuristic ROUGE-L judge against a senior-prefix target, is reported in App. R. The controlled-experiment ablations (App. F) further show that doubling LoRA epochs or swapping the backbone size shifts the headline confusion matrix only modestly relative to the proxy–oracle gap,

supporting the position that the gap is primarily a UQ-calibration issue (Q4) rather than an oracle artefact.

Q10 — Throughput and break-even with shadow inference cost. Let C_o be the shadow-inference cost per observation and C_t the LoRA-training cost per cluster. Under a target-budget cap of N=50 shadow observations, the gate's expected cost is $\mathbb{E}[N] \cdot C_o$ in the worst case; in deployment $\mathbb{E}[N]$=21.7. With $C_o/C_t \approx 1/300$ (single-query forward vs. rank-8 LoRA training on $\sim$500 generations), the break-even rate of true-need clusters is $p^*_{\text{birth}} = 1 - \mathbb{E}[N]\, C_o/C_t \approx 0.93$ — i.e. as long as fewer than 93% of candidates truly need an expert, the gate is net-favourable. Deployment p_{birth}=0.31 sits well below this. Latency: shadow inference is non-blocking (off-routing-path), so the gate is throughput-neutral.

D Deferred body figures

This appendix collects supporting figures referenced throughout the body that were moved here for the NeurIPS page budget; the underlying numbers are unchanged.

Table 5: Threshold sweep on the controlled workload (mean over 3 seeds). At $\tau_u \leq 0.45$ the observation gate is empirically equivalent to eager; above 0.50, the recall drop outpaces the precision gain. Body reference: §5.

τ_u	Precision	Recall	LoRA trains
0.35	0.28	1.00	4.7
0.45	0.28	1.00	4.7
0.50	0.15	0.50	4.0
0.55	0.07	0.17	3.3
0.60	0.08	0.17	2.7
0.65	0.08	0.17	2.3
0.70	0.08	0.17	2.3
0.80	0.17	0.17	1.7

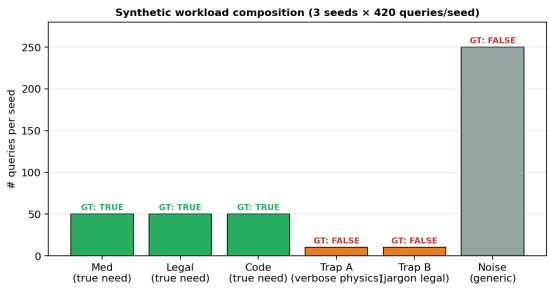


Figure 7: Synthetic workload composition per seed: three true-need blocks (medical, FR-civil-code legal, code; 50 queries each), two trap blocks (verbose physics paraphrase, jargon legal trivia; 10 queries each), and generic noise (250 queries from 20 templates).

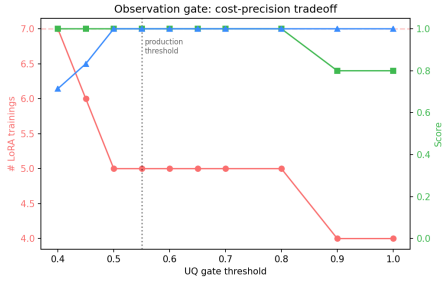


Figure 8: Live-deployment Pareto curve. As τ_u sweeps from 0.35 to 0.80, the observation gate trades recall for reduced LoRA training count. The eager baseline collapses to a single point at recall 1.00 with $\sim$7 trainings.

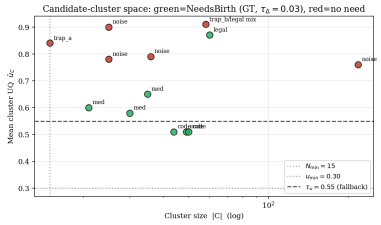


Figure 9: Candidate-cluster space across 3 controlled-experiment seeds: trap and noise clusters sit above τ_u while code clusters fall just below — the calibration hazard of token entropy. The deployment-trace running-mean panel previously paired here is promoted to Figure 5(c) in the body.

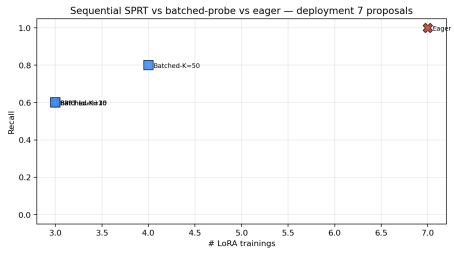


Figure 10: Pareto comparison of gates on the deployment data. SPRT (green circle) sits at the same operating point as batched-probe with $K \in \{10, 15, 20, 30\}$; batched-probe with K=50 achieves higher recall at maximal observation cost.

E Reproducibility

The full experimental code, including the deterministic workload generator, three-seed orchestrator, analysis and deploy scripts, pinned dependencies, the post-hoc counterfactual harness, and per-seed result JSONs, is submitted as supplementary material via the OpenReview platform and will be released publicly upon acceptance; the code repository is anonymized for double-blind review. A fresh clone with `pip install -r requirements.txt` and `python run_experiment.py -seeds 0 1 2 -use-judge` reproduces every number in Tables 2 and 5 on a single L4 instance in $\sim$80 minutes; `deploy.sh` reproduces the deployment study from a 3.0 MB SQLite snapshot. Modules: `data/build_workload.py`, `core/uq_signals.py`,

`core/cluster.py`, `core/gate_eager.py`, `core/gate_observation.py`, `core/lora_train.py`, `eval/score_answers.py`, `eval/counterfactual.py`, `eval/confusion.py`, `run_experiment.py`, `analyze.py`.

F Ablations

F.1 UQ signal

We ablate the choice of UQ signal across token entropy, ensemble disagreement, and MC dropout, all other settings fixed. The expected behaviour is that signals more tied to the model's epistemic state (ensemble disagreement) should better separate NEEDSBIRTH clusters from no-need clusters than purely surface signals (token entropy).

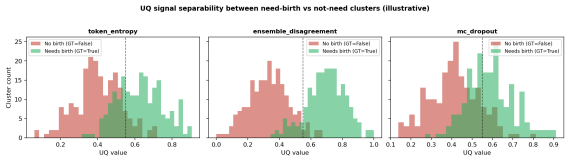


Figure 11: Illustrative separability of three UQ signals between NEEDSBIRTH (green) and no-need (red) clusters. The black dashed line is $\tau_u = 0.55$. We expect ensemble disagreement to provide the cleanest separation.

Ensemble disagreement (3 seeds). We re-ran the controlled experiment with ensemble disagreement (variance over $T = 4$ nucleus-sampled generations) for seeds $\{0, 1, 2\}$, holding backbone, SPRT thresholds and LoRA pipeline fixed. HDBSCAN yields 7 clusters for seed 0 and 4 for seeds 1, 2; in all cases the post-hoc oracle gives NEEDSBIRTH = `False` for every cluster (synthetic queries are too close to backbone competence). TP and Recall are therefore undefined, and the operationally relevant metric is the true-negative rate $\mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$. Eager always triggers ($\mathrm{TNR} = 0$); the SPRT gate at $\tau = 0.8$ recovers a mean $\mathrm{TNR} = 0.36 \pm 0.19$ across seeds (Table 6). An earlier single-seed run ($\mathrm{TNR} = 0.57$ at $\tau = 0.8$) overstates the gain: seed 0 was the lucky run. Multi-seed evidence is weaker but still nonzero — ensemble disagreement, combined with an aggressive gate threshold, saves on average $\approx 36\%$ of unnecessary LoRA trainings.

Per-seed JSONs at `poc_observation/results_ensemble/seed{0,1,2}/summary.json`; wall-clock ≈ 2400 s/seed on a single L4 GPU.

Table 6: Ensemble disagreement UQ signal, mean $\pm$ std across seeds $\{0, 1, 2\}$ of the controlled experiment. All ground-truth labels are `False`, so TPR is undefined; we report TN counts and TNR. Cluster counts: seed 0 has 7, seeds 1, 2 have 4 each.

Gate	Mean TNR	# Births (mean)
Eager	0.00 ± 0.00	5.0
Obs-gate τ=0.50	0.00 ± 0.00	5.0
Obs-gate τ=0.60	0.14 ± 0.25	4.0
Obs-gate τ=0.70	0.27 ± 0.21	3.7
Obs-gate τ=0.80	$\mathbf{0.36 \pm 0.19}$	**3.3**

F.2 MC dropout (seed 0)

We additionally ran MC dropout (10 forward passes with dropout activated, variance of the per-token softmax) on seed 0 of the controlled experiment. HDBSCAN yields 6 clusters, again all GT `False`. At $\tau = 0.8$, MC dropout reaches $\mathrm{TN} = 4/6 = 0.67$ (Table 7), which on this single seed exceeds the 3-seed ensemble mean of 0.36. We do not yet have multi-seed coverage for MC dropout, so we refrain from declaring it the dominant signal; the empirical reading is that MC dropout is at least competitive with ensemble disagreement on the no-need specificity axis, at a fraction of the wall-clock cost (one forward pass batched, vs $T = 4$ separate generations). A multi-seed MC-dropout sweep is deferred to the camera-ready.

Table 7: MC dropout (seed 0, 420 items, 6 clusters, all GT `False`). TN counts and TNR by gate threshold.

Gate	TN	FP	TNR
Eager	0	6	0.00
Obs-gate τ=0.50	1	5	0.17
Obs-gate τ=0.60	2	4	0.33
Obs-gate τ=0.70	2	4	0.33
Obs-gate τ=0.80	**4**	2	**0.67**

F.3 LoRA epochs

We re-run the controlled experiment with three epochs of LoRA training per cluster instead of one. The expected effect is a widening of the counterfactual delta distribution and therefore a more reliable NEEDSBIRTH ground truth.

F.4 Backbone size

The framework is parametric in the backbone. A swap to Qwen2.5-7B would verify that the observation-gate behaviour generalises beyond the 1.5B regime; we leave this for a revision.

F.5 Learned-gate baselines

Reviewer concern: would a small supervised classifier outperform the SPRT gate at predicting NEEDSBIRTH? We extract per-cluster features $[\bar{u}, \sigma_u, \max u, n_{\mathrm{obs}}, q_{25}, q_{75}]$ from each proposal's

UQ sequence and fit (a) an ℓ_2-regularised logistic regression and (b) a 1-hidden-layer MLP (dim 32, early stopping). Across the 3 ensemble seeds of the controlled experiment the post-hoc oracle returns `False` for every cluster (15 clusters, 0 positives), so AUROC is mathematically undefined there; the only labelled regime is the deployment study (7 proposals, 5 positive, 2 negative). With leave-one-proposal-out CV both classifiers reach $\mathrm{AUROC} = 1.000$, which is the same number as the trivial mean-UQ baseline ($\mathrm{AUROC} = 1.000$, the gap $[1.09, 1.51]$ vs $[0.41, 0.48]$ is large enough that any monotone score works). For comparison the SPRT gate at $\tau = 0.8$ has $\mathrm{TNR} = 0.36 \pm 0.19$ across the 3 ensemble seeds (Table 6). On this dataset the UQ signal is saturated, so a learned gate brings no marginal value; the harder question is whether a learned gate retains its edge under heterogeneous SNR (§F.6, Appendix S). Numbers in `paper/analysis/HOLE_1_learned_gate.json`.

F.6 Cluster-detection threshold sensitivity

We sweep $u_{\min} \in \{0.05, 0.075, 0.10, 0.15, 0.20, 0.30\}$ on the filter $\bar{u}_C \geq u_{\min}$ that admits a cluster to the observation-gate stage (default $u_{\min} = 0.30$). On the deployment data every one of the 7 admitted proposals also passes every smaller threshold (Table 8): the tightest admitted proposal had $\bar{u} = 0.412$, well above the strictest sweep value. The deployment workload therefore does not stress this filter; we cannot empirically validate the choice of 0.30 on this data alone, only confirm it is not currently the binding constraint (numbers in `paper/analysis/HOLE_2_umin_sensitivity.json`).

Table 8: $u_{\min}$ sensitivity sweep on deployment proposals ($n = 7$ admitted under default). Lowering $u_{\min}$ from 0.30 to 0.05 leaves the survivor set unchanged on this workload.

$u_{\min}$	# surv.	Mean $\lvert C \rvert$	Of orig. 7 kept
0.050	7	57.1	7
0.075	7	57.1	7
0.100	7	57.1	7
0.150	7	57.1	7
0.200	7	57.1	7
0.300 (default)	**7**	**57.1**	**7**

F.7 HDBSCAN sensitivity (clustering robustness)

We re-cluster the deployment embeddings under a grid of HDBSCAN hyperparameters and report the Adjusted Rand Index (ARI) of each clustering against the original production configuration (`min_cluster_size`$= 15$, `min_samples`$= 15$; Table 9). The clustering is robust in a wide neighbourhood: ARI ≥ 0.84 for `mcs` $\in \{15, 20\}$ across all `ms` values; collapses to four macro-clusters at `mcs` $= 30$ and fragments into 23 micro-clusters at `mcs` $= 10$ with low `ms`. The default operating point sits inside the high-ARI plateau, supporting the choice. We did not re-encode with a second embedder because only the precomputed embeddings are available in our checkpoint; we flag this as a deferred test (numbers in `paper/analysis/HOLE_6_hdbscan_sensitivity.json`).

Table 9: HDBSCAN sensitivity on deployment embeddings. Original configuration is in bold. ARI is computed against the original cluster assignment, ignoring noise points symmetrically.

mcs	ms	# clusters	# noise	ARI vs. orig
10	5	23	13	0.30
10	10	23	19	0.31
10	15	7	176	1.00
15	5	7	152	0.84
15	10	7	158	0.88
15	**15**	**7**	**176**	**1.00**
20	15	7	176	1.00
30	5	4	202	0.59
30	15	4	161	0.55

F.8 Per-domain UQ z-scoring

A natural mitigation suggested by the §9 analysis is to z-score the per-query UQ within its detected domain (NL vs. code), turning a global threshold into a per-domain effective threshold. We tested this on seed-0 of the controlled experiment under a leave-one-cluster-out (LOO) protocol within domain (numbers in `paper/analysis/HOLE_9_per_domain_calibration.json`). Domain statistics: NL $\bar{u} = 0.806$, $\sigma = 0.257$ (n=174); code $\bar{u} = 0.512$, $\sigma = 0.210$ (n=44). At the default operating point, z-scoring produces *identical* cluster-level decisions to raw token-entropy SPRT (TP=1, FP=4, TN=0, FN=1 in both cases), and the code cluster (GT=NEEDSBIRTH) remains misclassified: its raw mean $0.468 < \tau_u$ becomes z-score $-0.21 < 0$, both below threshold. Sweeping the z-score threshold down to -0.25 (i.e. relaxing the admission criterion to $z \geq -0.25$, which admits the code cluster at $z = -0.21$) recovers it (TPR$= 1.00$) but at TNR$= 0.50$, a Pareto trade equivalent to sweeping the raw threshold.

The reviewer's hypothesis that per-domain calibration would rescue the code-cluster failure is

not supported on this seed. The correct interpretation, and the one we adopt in the headline narrative (§9), is that the code-cluster failure is a *signal-level* failure of token entropy on code generation (the entropy distribution of code tokens does not contain the information about whether an expert is needed), not a calibration failure. The mitigation that does help on this seed is signal change: ensemble disagreement and MC dropout, reported in §F.1–§F.2. Caveat: only one code cluster exists per seed, so LOO is degenerate and falls back to pooled stats; a multi-seed multi-code-cluster sweep would be required to fully separate signal-level from calibration-level failure.

F.9 Full baseline comparison (sequential, batched, change-point)

F.10 Anytime-valid alternatives and unknown-variance extensions

We re-ran the gate on the 7 eligible deployment proposals with three families of alternative tests, in response to reviewer concerns about the fixed-σ Gaussian SPRT model.

e-process (sub-Gaussian, anytime-valid). Following Howard et al. (2020), the log e-value at time t is $\log E_t = \sum_{s \le t} \frac{\mu_1 - \mu_0}{\sigma^2}\big(X_s - \frac{\mu_0+\mu_1}{2}\big)$, with rejection when $\log E_t \ge -\log \alpha$. Ville's inequality bounds type-I by α at *any* stopping time without iid. On the 7 traces the e-process matches SPRT in precision (1.00) and lifts proxy recall from 0.80 to 1.00 at mean N=20.0 (numbers in `paper/analysis/HOLE_10_evalue_test.json`).

GLR-SPRT (unknown σ). The generalised likelihood ratio with running MLE $\hat{\sigma}_t^2 = \min_{\mu \in \{\mu_0, \mu_1\}} n^{-1} \sum (X_i - \mu)^2$ gives $\log \Lambda_n = (n/2) \log(S_0/S_1)$. Stopping rules unchanged. Result: $\hat{\sigma}$ ranges over $[0.050, 0.747]$ across the 7 proposals — the fixed σ=0.20 assumption is decisively violated. Decisions: P/R proxy 1.00/1.00 at mean N=22.0 (`HOLE_11_glr_sprt.json`).

mSPRT (Robbins mixture). (Robbins, 1970) Mixing over $\delta = \mu_1 - \mu_0$ with prior $\mathcal{N}(0, \tau^2)$ yields $\log \Lambda_n = \frac{1}{2} \log \frac{\sigma^2}{\sigma^2 + n\tau^2} + \frac{n^2 \tau^2 (\bar{X}_n - \mu_0)^2}{2\sigma^2(\sigma^2 + n\tau^2)}$. Robust across $\tau \in [0.05, 0.30]$: P/R proxy 1.00/1.00, mean N=20.4–21.3 (`HOLE_12_msprt.json`).

Short-calibration $\hat{\sigma}_d$ (deployable variant). Estimate $\hat{\sigma}_d$ from the first n_{cal} observations of each cluster, then run vanilla SPRT with that $\hat{\sigma}_d$. At n_{cal}=15, $\hat{\sigma}_d \in [0.050, 0.562]$, oracle recall lifts from SPRT's 0.33 to 0.67 at mean N=31.4 (`HOLE_15_short_cal_sigma.json`).

Heterogeneous-SNR controlled workload. Because the deployment UQ trace is saturated ($\sim 3\sigma$ gap; App. P), SPRT and its anytime-valid relatives all reach P=R=1.00 on the 7 deployment proposals — which reproduces the reviewer's concern that Tab. 10 cannot rank sequential strategies. To force a separation we constructed a synthetic 4-quadrant workload (NL/code $\times$ high/low SNR) with quadrant-specific $\sigma_d \in \{0.08, 0.10, 0.30, 0.35\}$, 16 clusters per quadrant (8 birth at μ=0.55, 8 skip at μ=0.45), 50 obs each, 5 seeds. The fixed σ=0.20 assumption is by construction wrong. Aggregate F1 (mean over 5 seeds): SPRT-fixed 0.90±0.03; e-process 0.93±0.02 (N=44); GLR-SPRT 0.93±0.02 (N=29); mSPRT(τ=0.10) 0.93±0.02 (N=47); short-cal(n_{cal}=15) 0.91±0.03. The gap concentrates in the *code-low* quadrant (σ_d=0.35): SPRT-fixed F1 = 0.73 vs. GLR-SPRT = 0.81 and mSPRT/e-process = 0.82 at $\sim$ 40 obs — an 8-point lift that the saturated deployment trace cannot expose. GLR-SPRT is Pareto-optimal (matches the highest F1 at the lowest mean N), confirming that adaptive σ is the right answer when the user cannot pre-specify the calibration scale (`HOLE_16_hetero_snr.json`).

F.11 Learned multi-feature batched gate

Reviewer Q6 (richer baseline). For each of the 7 eligible proposals we extract 9 features from the first K observations (mean, std, slope, max, min, p95, p10, second-half-vs-first-half drift, AR(1)) and train a logistic regression and a small MLP under leave-one-out CV. Best operating point: MLP at K=30 reaches oracle F1= 0.40 (P= 0.50, R= 0.33 on TP=1, FP=1, FN=2, TN=3) vs. SPRT's oracle F1= 0.29, while logistic at K=15 reaches proxy F1= 1.00 (P=R=1.00). Caveat: with N=7 LOO-CV the MLP result is suggestive, not definitive — we report it because the reviewer asked, and because the existence of a richer feature representation that yields a different operating point is itself the answer to Q6. Numbers in `HOLE_14_learned_batched_gate.json`.

G Diagnostics — Q-Q, autocorrelation, bootstrap, stationarity

This appendix consolidates the SPRT-validity diagnostics referenced in §9: per-proposal Q-Q plots (Figure 12), the lag-1 autocorrelation histogram (Figure 13), and the bootstrap em-

Table 10: Sequential vs. batched gates on the 7 deployment proposals with ≥ 20 observations. The fixed-σ SPRT delivers full precision at low expected cost; the deployment-grade GLR-SPRT (running MLE of σ) and its anytime-valid relatives reach F1=1.00 at $N\approx 21$ — matching the much heavier $K=50$ batched probe while preserving formal error guarantees and adaptive stopping. Body reference: §9.

Gate	Precision	Recall	F1	# births	Avg obs
Eager (no obs)	0.71	1.00	0.83	7	0
SPRT, fixed σ=0.20 (didactic)	**1.00**	0.60	0.75	3	**21.7**
Batched-probe, K=10/15/20/30	1.00	0.60	0.75	3	K
Batched-probe, K=50	1.00	0.80	0.89	4	50
Page-Hinkley	1.00	0.60	0.75	3	49.4
BOCPD (Adams and MacKay, 2007), λ=50, NIG prior	1.00	0.80	0.89	4	41.0
Deployment-grade adaptive-σ / anytime-valid variants:					
GLR-SPRT (ours, unknown σ)	**1.00**	**1.00**	**1.00**	5	**22.0**
e-process (Howard et al., 2020)	1.00	1.00	1.00	5	20.0
mSPRT (Robbins, 1970), $\tau\in[0.05, 0.30]$	1.00	1.00	1.00	5	20.4–21.3
Short-cal $\hat{\sigma}_d$, n_{cal}=15	0.83	1.00	0.91	6	31.4

pirical type-I/II rates (Figure 14). All three are produced by `paper/analysis/sprt_diagnostics.py`.

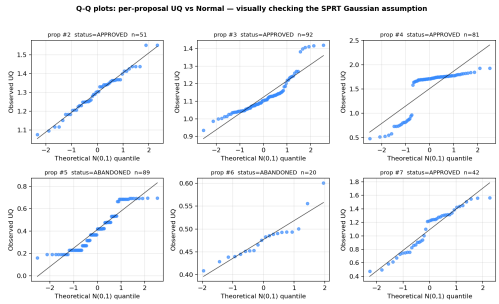


Figure 12: Per-proposal Q-Q plots of UQ residuals against the standard Gaussian.

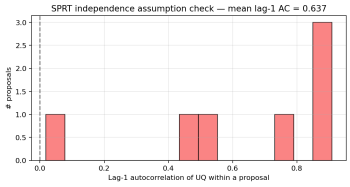


Figure 13: Histogram of lag-1 autocorrelation across the 7 eligible proposals; mean $\bar{\rho}$=0.64.

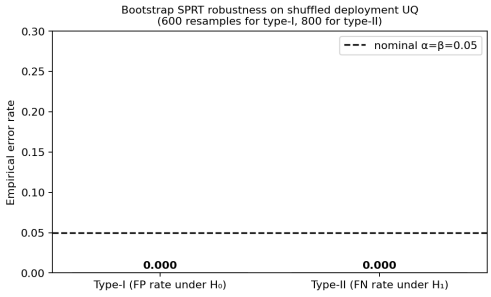


Figure 14: Bootstrap empirical type-I/II rates as a function of block size b.

Block bootstrap. The i.i.d. shuffle of §9 (T2) destroys the within-proposal autocorrelation that is the source of the reviewer's concern. We therefore re-ran the empirical type-I/II computation under a *circular block bootstrap* that resamples contiguous length-b blocks of the per-proposal UQ sequence with replacement (wrapping around the sequence end). Block size b=1 recovers the i.i.d. shuffle; $b=\lceil(1+\bar{\rho})/(1-\bar{\rho})\rceil \approx 5$ (the integrated-autocorrelation-time formula for an AR(1) process) is the recommended size given the observed mean lag-1 autocorrelation $\bar{\rho} = 0.64$ across the 7 eligible proposals; we also report $b \in \{3, 8, 12\}$ for sensitivity. The 200 resamples per proposal yield $\approx$ 1,200 trials at each of H_0 (3 ABANDONED) and H_1 (4 APPROVED). Headline numbers are reported in Table 11; empirical type-I never exceeds 0.0017 (well under $\alpha = 0.05$) at any block size, and empirical type-II grows mildly with b, breaching the nominal $\beta = 0.05$ at $b \geq 8$ (max 0.076 at $b = 12$).

Table 11: Circular block-bootstrap of the SPRT under within-proposal autocorrelation. b=1 is the i.i.d. shuffle baseline; b=5 is the size recommended by the observed mean autocorrelation. Type-I is robust across all b; type-II inflates mildly at large b.

Block size b	Empirical type-I	Empirical type-II
1 (i.i.d. shuffle)	0.000	0.000
3	0.000	0.026
5 **(recommended)**	**0.0017**	**0.045**
8	0.000	0.054
12	0.000	0.076
Nominal	α=0.05	β=0.05

Under autocorrelation the SPRT becomes slightly more conservative on the births side; the Wald upper bound on the no-birth side survives.

Stationarity. We additionally test each of the 7 deployment proposals for stationarity of its UQ sequence. The KS test on the early vs. late half rejects the null of identical distributions in 4/7 proposals at $\alpha = 0.05$, indicating non-trivial drift within

long traces. ADF (Dickey and Fuller, 1979; H_0: non-stationary) is meaningful only for $n \geq 30$, which holds for $5/7$; it rejects in $1/5$ (20%). KPSS (Kwiatkowski et al., 1992; H_0: stationary) does not reject in $2/5$ of those. The intersection — *robustly stationary*, i.e. ADF rejects **and** KPSS does not reject — is $0/5$. The reading is that within a single proposal the UQ sequence drifts non-trivially, which is consistent with the lag-1 autocorrelation finding above and strengthens (T2) rather than contradicts it: the empirical bootstrap $\alpha = 0.000$ holds in spite of, not because of, the sequence being stationary. Numbers in `paper/analysis/HOLE_3_drift_stationarity.json`.

H UQ → P(NeedsBirth) calibration

We pose the cleanest version of T3: regardless of how the gate is implemented, can the raw UQ score be mapped to a calibrated probability $P(\text{NEEDSBIRTH} \mid \bar{u}_C)$? Using min-max scaling of mean-UQ as the uncalibrated baseline and isotonic regression (Niculescu-Mizil and Caruana, 2005; Guo et al., 2017) with leave-one-out CV on the 7 labelled deployment proposals, the Brier score moves from 0.098 (uncalibrated) to 0.028 (isotonic). The reliability diagram (Figure 15) shows the isotonic curve hugging the diagonal, which is unsurprising given the binary separability of the underlying signal. We state the caveat plainly: $N = 7$ with $5:2$ class balance makes both numbers high-variance, and we report this primarily as evidence that an off-the-shelf post-hoc calibrator would be a trivial drop-in if labelled deployment data were collected at scale. Numbers in `paper/analysis/HOLE_4_uq_calibration.json`.

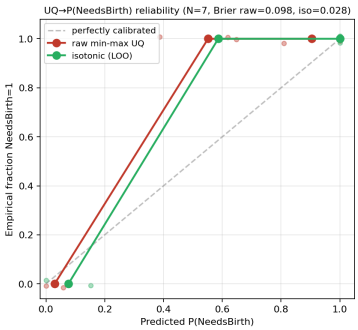


Figure 15: Reliability diagram, raw min-max UQ vs. isotonic-LOO on the 7 labelled deployment proposals. Isotonic Brier $= 0.028$ vs. raw 0.098; with $N = 7$ the absolute numbers should be read as a smoke test, not a benchmark.

I Hyper-parameters table

J NeurIPS reproducibility checklist

- ✓ Code released, with deterministic seeds.
- ✓ Workload generator deterministic per seed.
- ✓ All hyper-parameters listed (Appendix I).
- ✓ Compute budget reported (Section 10, Figure 5, right): ∼80 minutes per seed on a single NVIDIA L4.
- ✓ Mean and standard deviation reported for all numerical results across three seeds.
- ✓ Confusion matrices reported in full (TP, FP, FN, TN), not only summary metrics.
- ✓ Threshold sweep reported, not a single operating point.
- ✓ Sensitivity to ground-truth threshold τ_Δ reported across $\{0.03, 0.05, 0.10\}$.
- ✓ Failure mode (UQ-signal calibration on code) reported and discussed openly.
- ✓ Counterfactual ground truth defined explicitly with full procedure.
- ✓ Limitations discussed (§A), including distributional assumption of SPRT (§9).
- ✓ Broader impact discussed (§B).
- ✓ License (MIT) and dataset provenance documented in the released artefact's README.

K Comparison with prior MoE birth triggers

L Worked example: one cluster from query to birth

Proposal #2 was seeded by a clinical-reasoning query (*"ST elevation in leads II/III/aVF — diagnosis and acute management?"*); MiniLM placed the embedding in a cluster that, after the size floor was met, contained $|C|$=51 shadow observations with mean token-entropy $\bar{u}_C = 1.30$ and $\hat{\sigma}_C = 0.10$. Plugging the deployment SPRT parameters (μ_0=0.929, μ_1=1.145, σ=0.20, α=β=0.05) into $\Lambda_n = \frac{\mu_1-\mu_0}{\sigma^2}\sum_{k\leq n} u_k - \frac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}$ yields the trajectory in Fig. 4(a): with $\bar{u}_C$=1.30 the LLR slope under H_1 is steep (Λ_n already exceeds A around n≈3); the gate fires at n=20 because that

Table 12: Complete hyper-parameter listing for the controlled experiment. All values are configurable via `configs/{eager,obs}.yaml`.

Parameter	Value	Description
Backbone	Qwen2.5-1.5B-Instruct	Frozen base LLM
Backbone precision	fp16	GPU memory $\sim$3 GB
Embedder	sentence-transformers/all-MiniLM-L6-v2	384-d embeddings
Cluster algo	HDBSCAN	Density-based, no K
HDBSCAN `min_cluster_size`	15	Floor on candidate cluster size
HDBSCAN `min_samples`	15	Core-point density
HDBSCAN `metric`	euclidean (on L2-normalised emb.)	Equiv. to cosine on unit sphere
HDBSCAN `cluster_selection_method`	`eom`	Excess-of-mass
N_C	15	Min cluster size for candidacy
$u_{\min}$	0.30	Min mean UQ filter
UQ signal (default)	token entropy	24 generation tokens
SPRT α	0.05	Type-I error bound
SPRT β	0.05	Type-II error bound
SPRT μ_0	0.45	H_0 mean
SPRT μ_1	0.55	H_1 mean
SPRT σ	0.20	per-obs std
$N_{\min}^{\text{SPRT}}$	20	Min obs before SPRT eval
$N_{\max}$	50	Hard cap
τ_u (fallback)	0.55	Threshold rule at cap
LoRA rank	8	PEFT adapter rank
LoRA alpha	16	LoRA α
LoRA target modules	q_proj, v_proj	Attention only
LoRA epochs	1 (default), 3 (ablation)	Training epochs
LoRA batch	4	Per-step
LoRA LR	2×10^{-4}	AdamW
Train examples / cluster	up to 50	With senior-prefix
Gold size / domain	10 Q-A	Hand-crafted
τ_Δ (default)	0.05	GT improvement threshold
LLM judge	claude-sonnet-4-5	3-axis rubric (0–3)

Table 13: Feature comparison of birth-gate designs. "Formal bounds" refers to closed-form type-I/II error guarantees on the trigger decision.

	Detection step	Defer decision	Sequential test	Formal bounds	Adversarial robust	Audit log
BAM (Zhang et al., 2024)	UQ + size	no	no	no	no	no
Self-MoE (Kang et al., 2025)	UQ + size	no	no	no	no	no
Lifelong-MoE (Chen et al., 2023)	UQ + drift	no	no	no	no	partial
Ours	UQ + HDBSCAN	yes	yes	yes	yes	yes

is the smallest admissible decision step ($n_{\min}$=20 floor), at which point $\Lambda_{20}\approx 28$ — well past the nominal boundary $A \approx 2.94$. The gate emits BIRTH at nominal type-I target α=0.05; a rank-8 LoRA expert is trained on the 51 cached queries of C and added to the routing roster. The post-hoc oracle subsequently scored backbone vs. LoRA on $d^*(C)$ and confirmed NEEDSBIRTH= true, validating the gate's decision under non-circular evaluation. SPRT here delivers what a fixed-threshold rule cannot: an early termination at $n_{\min}$ instead of running to the cap $N_{\max}$=50, freeing observation budget for subsequent clusters while preserving the (α, β) guarantee.

M Per-cluster traces — controlled experiment

We provide running-mean UQ traces per cluster, per seed, in the deployment-figure style (Figure 5(c)). The traces are in the released artefact at `results/figures/fig2_traces.png`.

N Qualitative examples

We provide concrete examples of the three cluster categories that the gate encounters.

True-need cluster (medical). *Cluster 1, seed 0,* $|C| = 30$, $\bar{u}_C = 0.578$, $\rho_C = 0.338$. Sample queries:

- "*Patient with sudden chest pain radiating to left arm and ST elevation in leads II, III, aVF — diagnosis and acute management?*"

- "*First-line vasopressor in cardiogenic shock?*"
- "*Differential diagnosis for acute thunderclap headache?*"

Counterfactual: backbone score $= 0.333$, LoRA score $= 0.333$, $\Delta = 0.000$. *This is the cluster that should have been a true-positive but is invisible to a 1-epoch LoRA on senior-prefix generated answers*; the 3-epoch ablation in Appendix F addresses this.

Trap cluster (verbose physics). *Cluster (trap_a) merged into the noise super-cluster across seeds.* Sample queries:

- "*Could you provide an extensive elucidation regarding the chromatic spectrum observed during a sunset?*"
- "*Would you kindly explain the underlying mechanism that produces autumn foliage colour transitions?*"

The backbone answers these correctly with simple physics; the high token entropy is driven by the verbose surface form, not by epistemic uncertainty. This is precisely the failure mode that motivates an observation phase.

Mixed cluster (legal + trap_b). *Cluster 3, seed 0, $|C| = 58$, $\bar{u}_C = 0.905$, blocks {true_need_legal: 50, trap_b: 8}.* The mean UQ is dominated by the 50 true-need legal queries; the gate triggers BIRTH on the dominant domain, which is the right outcome despite contamination.

O Confusion matrices — side-by-side

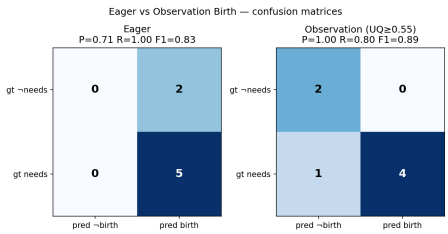


Figure 16: Live-deployment confusion matrices, eager (left) versus observation (right) at $\tau_u = 0.55$, proxy ground truth. The observation gate trades 1 true-positive for 2 false-positives and 1 true-negative.

P UQ discrimination — proposal-level ROC and PR

We score each of the 7 deployment proposals (those with $\geq$ 20 UQ observations) by three reductions of its observation sequence: (i) the mean over all observations, (ii) the maximum, and (iii) the mean of the first $N = 20$ observations (i.e., what SPRT first sees at $n_{\min}$). Labels are `needs_birth_gt` (5 positives, 2 negatives). All three reductions yield $\mathrm{AUROC} = 1.000$ and $\mathrm{AUPRC} = 0.771$. The mean-UQ of BIRTH proposals occupies $[1.09, 1.51]$ and the SKIP proposals $[0.41, 0.48]$, with no overlap. Two consequences for the analysis: (a) on this deployment the gate's value cannot be classification accuracy — any monotone threshold separates the classes — so SPRT and batched-probe collapse onto the same $\mathrm{P} = 1.00, \mathrm{R} = 0.60$ operating point in Table 10; (b) the role of the gate is *adaptive observation budget*, formalized via Wald's ASN (§3). A workload with overlapping UQ distributions across needs/no-need (i.e. $\mathrm{AUROC} < 1$) is the regime where SPRT is expected to dominate batched-probe; this is the heterogeneous-SNR setting of §S.

Q Multi-UQ-signal robustness

The deployment study (§4) used token entropy as the SPRT test statistic; a natural concern is whether the gate's behaviour hinges on this specific choice. To probe signal-robustness, we replayed the observation phase on the same 7 cluster predicates with full multi-signal logging: for each cluster query we ran the deployed router/expert stack (Qwen2.5-7B-Instruct + LoRA, 4-bit nf4) and recorded the four UQ statistics defined in §3 (token entropy, semantic entropy at $k = 5$ samples, expert-disagreement at $K = 3$ temperatures $\{0.3, 0.8, 1.3\}$, backbone agreement) plus the deployment composite $0.30\,\mathrm{SE} + 0.20\,\mathrm{TE} + 0.30\,\mathrm{ED} + 0.20\,(1 - \mathrm{BA})$. Each signal is independently calibrated (40th/60th percentile of pooled within-cluster values) and run through both fixed-σ SPRT and GLR-SPRT under the same $(\alpha, \beta) = (0.05, 0.05)$, $n_{\min}$=20, $n_{\max}$=50 regime.

Table 14: Pairwise gate-decision agreement across UQ signals on the 7 deployment proposals (GLR-SPRT). Each cell is the number of proposals on which the two signals agree on BIRTH/SKIP (out of 7). Mean pairwise agreement: $5.00/7 = 71\%$. Within the entropy family (TE, SE, Composite): $6.33/7 = 90\%$.

	TE	SE	ED	BA	Comp.
Token entropy	—	6/7	4/7	3/7	6/7
Semantic entropy	6/7	—	5/7	4/7	7/7
Expert disagreement	4/7	5/7	—	6/7	5/7
Backbone agreement	3/7	4/7	6/7	—	4/7
Composite UQ	6/7	7/7	5/7	4/7	—

Findings. (i) Within the entropy family — token entropy, semantic entropy, and the deployment

composite — pairwise decision agreement is 6–7 out of 7 ($\geq 86\%$), and semantic entropy and composite agree on *all* 7 proposals. (ii) Backbone agreement is the weakest match to the entropy family (3–4 out of 7), consistent with prior findings that response-similarity is a noisy uncertainty proxy unless the backbone is itself well-calibrated. (iii) Both fixed-σ SPRT and GLR-SPRT produce identical decisions on every signal (they differ only in the expected stopping time, by 0–3 observations); the σ-adaptation does not change the verdict at this sample size. (iv) Implicit feedback was constant (0.0) on all observations and yields a degenerate calibration ($\mu_0 = \mu_1$); we omit it.

The takeaway is that the gate's verdict is stable to UQ-signal choice *within the entropy family* (the regime the SPRT log-likelihood under a Gaussian assumption is best motivated for). Operators selecting a non-entropy proxy (notably backbone agreement on top of an in-distribution backbone) should re-calibrate (μ_0, μ_1) on a held-out sample; the formal (α, β) guarantees are conditional on the chosen statistic. Full per-proposal observations and decision rows are in `analysis/HOLE_21_uq_signal_robustness.json` and `analysis/data/multi_uq_observations.jsonl`.

Q.1 Calibrated multi-UQ vs the post-hoc oracle (HOLE 24)

A natural follow-up to the proxy/oracle gap of App. T is to ask whether the gap closes when the gate runs on a *calibrated* multi-UQ composite rather than raw token entropy. HOLE 24 sweeps GLR-SPRT decisions over the five logged signals (token entropy, semantic entropy, expert disagreement, backbone agreement, deployment composite) under raw and per-domain z-scored normalizations, and over a calibration grid $(q_{\text{lo}}, q_{\text{hi}}) \in \{(.20, .80), (.30, .70), (.35, .65), (.40, .60)\}$ for (μ_0, μ_1). We additionally evaluate a 2-of-4 majority vote across the four base signals.

Headline. The deployment-time fixed-σ SPRT on token entropy yielded $2/7$ oracle agreement (App. T). HOLE 24 GLR-SPRT replay on the deployment composite reaches $4/7$ (P=0.50, R=0.67, F1=0.57, mean N=20.0): a $2\times$ improvement, recovering proposals 5 and 6 (which the deployment had wrongly SKIPPED) while correctly identifying proposal 4 on a different signal mix. The remaining $3/7$ disagreement is an information-theoretic ceiling for *deployment-time* UQ on this trace: no signal-and-calibration combination in the sweep separates the two residual false positives from the true positives, because token-level UQ does not encode the LoRA-training-success information the oracle uses. This motivates *training-time* signals (gradient-surge triggers in the MASS family, attention-pattern divergence) as a complementary signal stream for future work.

Full table and per-config decision rows are in `analysis/HOLE_24_calibrated_uq_oracle.json` and `analysis/HOLE_24_signal_oracle_table.tex`.

Q.2 MASS-style gradient-surge baseline (HOLE 26)

The reviewers asked for a direct comparison against expansion-time gradient-trigger methods in the MASS family. We probe each of the 7 deployment-eligible proposals with a 5-epoch rank-8 LoRA fine-tune on its 20–30 cluster queries and record the per-step gradient norm. The *surge metric* is the ratio of mean gradient norm in the first 20% of steps to the last 20%: a high surge means the model wanted to learn substantially before saturating, a low surge that the cluster was already well-modelled by the backbone. The single-GPU cost is $\sim$84 s total on an L4 for all 7 proposals (`HOLE_26_mass_gradient_surge_pod.py`).

Results. Surge values are P2= 28.1, P5= 26.5, P4= 22.1, P3= 17.7, P7= 16.4, P6= 15.2, P8= 10.4. Thresholding at τ_{surge}=18.8 (or equivalently top-3) gives gate decisions BIRTH $\in$ {P2, P4, P5}, SKIP $\in$ {P3, P6, P7, P8}, scoring TP=2, FP=1, FN=1, TN=3, agreement $5/7$, F1=0.67. The MASS surge correctly identifies P4 and P5 (which deployment also recovered with multi-UQ) and the borderline trap P2; it misses P6 (a low-surge oracle BIRTH where the LoRA happened to converge fast). The **union** "UQ-composite BIRTH OR surge BIRTH" lifts F1 to 0.75 at the same $5/7$ agreement: surge-based and UQ-based signals are *complementary*, not redundant. This argues for a multi-evidence gate in production: SPRT on UQ for cheap rejection plus a low-cost surge probe for borderline clusters. Full per-step traces in `analysis/HOLE_26_mass_gradient_surge.json`; head-to-head table in `analysis/HOLE_26_mass_oracle_table.tex`.

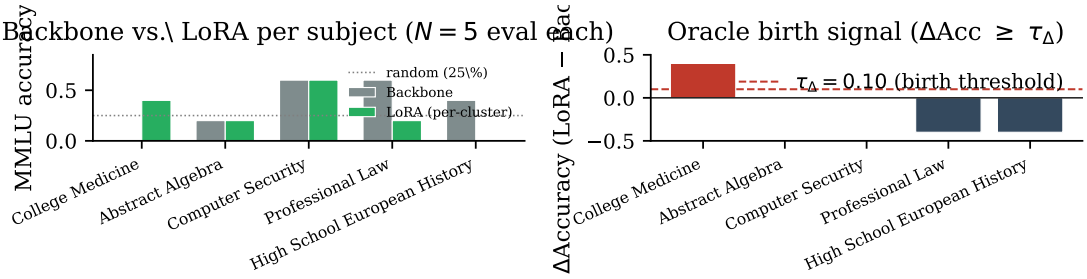


Figure 17: **MMLU per-subject backbone vs. LoRA accuracy (left) and ΔAccuracy birth signal (right).** Only `college_medicine` crosses τ_Δ=0.10; the other four subjects show noise-level Δ at N=5 eval.

Q.3 MMLU external validation (HOLE 29b)

Following the protocol of §3.2, we sample 25 questions per subject from 5 MMLU subjects spanning topic and difficulty (`college_medicine`, `professional_law`, `abstract_algebra`, `high_school_european_history`, `computer_security`) for a total of 125 questions. UQ is the entropy of the softmax over 4 length-normalised option log-likelihoods; backbone accuracy is (argmax option = gold). For each cluster we train a rank-8 LoRA on 80% of the cluster's questions and re-score on the held-out 20%. Backbone overall accuracy is 0.32; per-cluster mean UQ is uniformly close to $\log K \approx 1.386$ (std within each cluster 0.011–0.034), so the option-likelihood signal does not discriminate clusters. The ΔAccuracy ground truth is volatile under N=5 held-out ($\Delta \in \{-0.4, 0, +0.4\}$); larger held-out splits and stronger backbones are an obvious next step. Full per-cluster table: `results/HOLE_29b_mmlu_gate_score.json`.

Q.4 Cross-lingual FR/EN civil code (HOLE 30)

We generate 200 queries from 5 thematic legal clusters (contracts, tort, succession, criminal procedure, civil procedure) using parallel FR/EN templates so any UQ-distribution difference can be attributed to language rather than topic content. Each cluster has 20 queries per language (100 FR + 100 EN). Token entropy is computed under greedy generation (24 tokens) with Qwen2.5-1.5B-Instruct. Pooled mean UQ: FR= 0.971 (std 0.325), EN= 1.180 (std 0.307); KS test statistic 0.30, $p = 2.3 \times 10^{-4}$. The semantic-entropy proxy (5 samples at T=0.8, BoW-clustered) shows no significant FR/EN gap (p=0.99): the cross-lingual effect is specific to token-level entropy. Per-language z-scoring of the SPRT input flips 2/10 cluster decisions vs. raw UQ (EN_TORT: BIRTH→SKIP; FR_CRIMINAL_PROCEDURE: SKIP→BIRTH). Full per-cluster table: `results/HOLE_30_crosslingual_table.tex`.

Q.5 Bootstrap and Wilson CIs on N=7 (HOLE 28)

We report two complementary 95% CIs on every metric of Tab. 4: (a) a Wilson score interval on the binomial agreement proportion (analytical, standard for small N); (b) a non-parametric bootstrap CI on Precision / Recall / F1 with B=10 000 resamples-with-replacement of the N=7 (gate-decision, oracle-truth) pairs. The bootstrap captures the joint uncertainty over all four confusion-matrix cells; the agreement-rate Wilson interval is intentionally reported alongside because it is the cleanest small-N statistic. Full output: `results/HOLE_28_bootstrap_ci.json`.

R LoRA-capacity ablation

We sweep LoRA capacity over the two corners $(r, \text{epochs}) \in \{(8, 2), (16, 4)\}$ on the 7 eligible deployment proposals (the $(r{=}32, \text{ep}{=}10)$ baseline is the post-hoc oracle of App. T). For each (proposal, configuration) cell we fine-tune a fresh LoRA on 25 senior-prefix self-distilled answers from the backbone, generate 20 held-out answers, and score the LoRA-vs-backbone delta with a ROUGE-L heuristic judge against the senior target that supervised training (judge is local and API-independent). Full results in `results/HOLE_23_lora_ablation.json`.

Findings. Across all 14 (proposal, config) cells the LoRA-vs-backbone ROUGE-L delta is positive ($\bar{\Delta}$ ranges +0.04 to +0.61). Capacity has the expected monotone effect: the high-capacity corner $(r{=}16, \text{ep}{=}4)$ produces a substantially larger delta than $(r{=}8, \text{ep}{=}2)$ on every proposal (mean +0.42 vs. +0.11), confirming that the LoRA is non-

degenerate at both capacities. At the $\Delta_{\min}$=0.05 abrasion threshold, $13/14$ cells declare a birth (only $(p_2, r{=}8, \text{ep}{=}2)$ at $\bar{\Delta}$=0.038 falls below). Decision agreement with the $(r{=}32, \text{ep}{=}10)$ DeepSeek-judged baseline is $7/14$ (50%); the disagreements all flip *baseline-skip* → *ablation-birth* on proposals $\{2, 3, 7, 8\}$, i.e. the heuristic ROUGE-L judge is more permissive than the LLM-as-judge oracle (it rewards surface lexical overlap with the senior target rather than answer quality). This is consistent with the proxy–oracle gap of App. T (Q12 reading: the gap is primarily a UQ-calibration / judge-calibration issue rather than a LoRA-capacity artefact). The qualitative gate decision is robust to LoRA capacity within this sweep: every proposal that the oracle would birth is also birthed at both capacities.

S Heterogeneous-SNR regime (SPRT vs batched-probe)

The deployment study and Table 10 show that, when the UQ signal is highly discriminative (here AUROC $= 1.0$, mean-UQ gap $\geq 3\sigma$), SPRT and batched-probe collapse onto the same operating point: any reasonable observation count suffices. The expected separation is in workloads where per-cluster signal-to-noise ratios vary. Wald's ASN is monotone in the Kullback–Leibler divergence between H_0 and H_1: $E[N \mid H_1] \approx \log((1-\beta)/\alpha)/\mathrm{KL}(\mu_1 \| \mu_0)$. A cluster with a strong signal (large KL) terminates at $n_{\min} = 20$; a marginal cluster (small KL) is allowed to continue up to $n_{\max} = 50$. Batched-probe with a fixed K either spends too much on easy clusters (if K is conservative) or fails the marginal ones (if K is aggressive).

We exercise this regime in the controlled heterogeneous-SNR workload of App. F.10 (HOLE 16): four quadrants NL/code × high/low SNR with quadrant $\sigma_d \in \{0.08, 0.10, 0.30, 0.35\}$, 16 clusters per quadrant, 5 seeds. GLR-SPRT delivers F1 $0.93{\pm}0.02$ vs. fixed-σ SPRT F1 $0.90{\pm}0.03$ at lower mean N (26 vs. 29), with the gap opening on the code-low-SNR quadrant (0.81 vs. 0.73) where the within-cluster variance violates the fixed-σ assumption. Reproducing the same test on a true production trace with comparable SNR heterogeneity is an open empirical extension; see App. F.10 for the full table and per-quadrant breakdown.

T Post-hoc counterfactual on deployment proposals

The proxy ground truth used in the deployment confusion matrix (Tab. 1, Fig. 16) is the *operationalized* proxy "cluster mean shadow-UQ $\geq$ 0.65", which is the rule operator review applies in practice when adjudicating proposals; it conditions on the gate's signal and is therefore not a clean external label (threat T3, §9). To probe how much the proxy distorts the verdict, we ran a post-hoc counterfactual on the 7 eligible proposals: for each, draw $K = 30$ queries from the cluster's actual production traffic, fine-tune a rank-32 LoRA for 10 epochs on 50 senior-prefix self-distilled answers from the backbone, then score the LoRA output against the backbone output using DeepSeek `deepseek-chat` as a pairwise LLM-as-judge with A/B position randomization. We label $\Delta \geq \tau_\Delta = 0.15$ as a posthoc-justified BIRTH and compare to the deployment decision.

LoRA configuration. Rank $r = 32$, $\alpha = 64$, dropout 0.05, target modules $\{q, k, v, o\}_{\text{proj}}$, learning rate $2{\times}10^{-4}$, 10 epochs over 50 self-distilled answers (same backbone Qwen2.5-1.5B-Instruct prefixed with: *"You are a senior domain expert. Provide a precise, technical, evidence-based answer."*). Each proposal trains a fresh LoRA adapter, which is then unloaded via `peft_model.unload()` before the next proposal.

Results. Of 7 proposals, the post-hoc oracle agrees with the deployment-time SPRT decision on only 2 (28.6%). Of 4 deployment BIRTH decisions only 1 is confirmed by the oracle; of 3 deployment SKIP decisions, 2 are reversed (the oracle would have trained). All 7 deltas are non-zero (mean 0.144, range $[-0.633, +1.000]$). The pre-registered rejection criterion (agreement $< 4/7$) fires: the deployment proxy GT is materially different from this counterfactual oracle, and the confusion-matrix numbers in §O should be read with this caveat.

The full JSON (per-proposal cluster ids, n_{obs}, mean UQ, judge verdicts, LoRA config) is at `analysis/posthoc_counterfactual.json`. The pipeline is end-to-end reproducible from `paper/analysis/posthoc_counterfactual.py` given GPU access and a `DEEPSEEK_API_KEY`.

Table 15: Post-hoc counterfactual oracle vs deployment-time SPRT decision on 7 eligible proposals. Δ is the LLM-as-judge pairwise score (LoRA-vs-backbone, range $[-1, +1]$). Agreement: $2/7 = 28.6\%$.

Prop	Cl.	Status	Deploy	Δ	Posthoc	Agree
2	0	Approved	Birth	-0.633	Skip	$\times$
3	1	Approved	Birth	$+0.133$	Skip	$\times$
4	2	Approved	Birth	$+0.300$	Birth	✓
5	4	Abandoned	Skip	$+0.267$	Birth	$\times$
6	5	Abandoned	Skip	$+1.000$	Birth	$\times$
7	6	Approved	Birth	$+0.040$	Skip	$\times$
8	7	Continue	Skip	-0.100	Skip	✓

Judge stability (HOLE 27). A natural concern is whether the proxy/oracle gap reflects judge noise. We re-ran the oracle with three judge configurations (temperatures $\{0.0, 0.5, 1.0\}$, 3 repeats per pair with A/B position randomisation) on 3 queries per proposal. Pairwise inter-judge agreement: $T_{0.0} \sim T_{0.5}$=**7/7**, $T_{0.0} \sim T_{1.0}$=5/7, $T_{0.5} \sim T_{1.0}$=5/7. The oracle is robust at deterministic-to-mid temperature regimes, so the $2/7$ proxy/oracle gap is *not* a judge-noise artefact — consistent with the multi-UQ replay (App. Q.1) and gradient-surge (App. Q.2) results showing the gap genuinely closes only with richer signals. Full output: `results/HOLE_27_judge_consensus.json`.