CERTIFICATE SEMANTICS DETERMINES QUERY GEOMETRY IN LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) evaluations commonly aggregate responses across prompt formats, option orders, and other representation variants. Such aggregation can obscure whether behavior is consistent within every registered representation group. We formulate representation-conditioned certification as a group-level ternary decision: for a fixed item, determine whether all group-conditioned response probabilities lie above a threshold, all lie below it, or groups occur on both sides. Certificate semantics induces distinct first-order information geometry. Robustness is universal, requiring dense resolution of potential blockers; heterogeneity is existential, requiring only a positive and a negative witness. By specializing established fixed-confidence pure-exploration geometry, we derive closed-form characteristic times and introduce ACTIVEORBIT, a sequential allocator that tracks the weakest blocker or balances opposite witnesses. We establish item-level anytime correctness and selector-specific first-order behavior. A mechanism intervention matches the predicted regime asymmetry: witness-selective allocation yields exactly 0% query reduction in all six robust scenarios and positive reductions in all six heterogeneous scenarios (mean 72.77%). Across six LLM model–task cells spanning BoolQ, PIQA, and RTE and three open-weight model families, ACTIVEORBIT reduces query cost by 12.03%–36.61% relative to round robin, with every paired 95% confidence interval excluding zero. These results establish a practical principle: query allocation should be designed around the logical structure of the claim being certified.

1 INTRODUCTION

Large language model (LLM) evaluations can be sensitive to prompt formatting, option order, and in-context presentation (Sclar et al., 2024; Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Zhao et al., 2021; Lu et al., 2022; Perez et al., 2021; Min et al., 2022; Webson & Pavlick, 2022). Evaluations may therefore repeat an item across representation variants intended to preserve the task. A common approach aggregates across variants—for example, by averaging, voting, or certifying a dominant response under a sampling distribution. Aggregation answers an average-case question. Claims of consistency across *every* registered representation group instead depend on the vector of group-conditioned response probabilities, not only on a pooled response rate.

This distinction is empirically substantial. Across 1,000 items spanning two reasoning families and two model families, the aggregate certificate rate is 85.7%, while option-wise and parity-wise robustness rates are 54.2% and 60.6%, respectively; no item satisfies the joint-robust criterion. On BoolQ (Clark et al., 2019) validation items, aggregate certification reaches 71.5% while joint robustness is again absent. Robustness and heterogeneity are separate positive certificates, leaving items without sufficient evidence unresolved; the results nevertheless establish that aggregate consistency and group-wise robustness are distinct statistical claims.

Retaining the representation groups exposes a second question: *how should queries be allocated to certify the desired claim?* In our grouped binary-threshold setting, certificate semantics determines the relevant evidence. Robust certification is universal: every group can remain a blocker. Heterogeneity is existential: one reliably positive and one reliably negative group suffice as witnesses. Specializing established pure-exploration geometry to this ternary certificate yields closed-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

form characteristic times with dense robust support and sparse heterogeneous support (Garivier & Kaufmann, 2016; Degenne & Koolen, 2019; Qiao & Tewari, 2025).

We operationalize this structure with ACTIVEORBIT, a sequential allocator that queries the weakest blocker when robustness is plausible and balances opposite witnesses under apparent heterogeneity while retaining anytime-valid stopping. Under the stated grouped Bernoulli assumptions, we establish item-level correctness and selector-specific first-order behavior.

Experiments validate both the mechanism and its practical consequences. With identical stopping semantics, removing witness selectivity yields 0% reduction in all six robust scenarios and positive reduction in all six heterogeneous scenarios. Across six model–task cells, ACTIVEORBIT reduces query cost relative to round robin by 12.03%–36.61% and remains competitive with stronger adaptive allocation rules.

The resulting principle is simple: *the claim being certified determines which representation groups remain information-theoretically relevant*. Evaluation protocols should therefore choose where to spend queries jointly with what they aim to establish. Our contributions are:

- We **formulate representation-conditioned LLM certification as a group-level ternary decision problem** and show empirically that aggregate certification need not provide an all-group robustness certificate.
- We **derive closed-form characteristic times for the grouped certificate**, exposing dense blocker support for robust claims and sparse opposite-witness support for heterogeneous claims as a structured specialization of pure-exploration geometry.
- We **introduce ACTIVEORBIT**, a sequential allocator that tracks blockers or opposite witnesses while retaining anytime-valid stopping, and establish its selector-specific first-order behavior.
- We **validate the predicted regime asymmetry and its practical query consequences**. A controlled intervention isolates the witness-selectivity effect, while six LLM model–task cells show 12.03%–36.61% query-cost reductions versus round robin and competitive performance against stronger adaptive allocation rules.

2 REPRESENTATION-CONDITIONED LLM CERTIFICATION

We study a language model over a finite, specified set of representation groups. For a fixed item, the target is whether the group-conditioned response distributions all support the same direction or disagree across groups.

2.1 RESPONSE MODEL AND CERTIFICATE REGIONS

Let $H \geq 2$ denote the registered groups. A group indexes a specified observation condition, such as an option ordering or layout, intended to preserve the task label. Repeated observations may arise from fresh surface-form variation, stochastic decoding under a fixed group prompt, or both. The guarantees apply to this grouping, with semantic equivalence of its prompts treated as part of the evaluation design.

When group h is queried, the fixed model produces a raw output Y under a fixed generation protocol, and the specified observation rule maps an accepted output to $Z_{h,s} \in \{0, 1\}$. Repeated accepted observations from group h have mean $p_h = \Pr(Z = 1 \mid h)$, with $p = (p_1, \ldots, p_H) \in [0, 1]^H$. Whenever group h is selected, we assume its next accepted observation is conditionally Bernoulli(p_h) given the previous history; independent i.i.d. group streams are a sufficient stronger condition. Within one item-level evaluation, the model, tokenizer, renderer, parser, decoding configuration, and representation groups are fixed.

Using threshold $1/2$, we define

$$\mathcal{R}_+ = \left\{ p : \min_h p_h > \frac{1}{2} \right\},\ \mathcal{R}_- = \left\{ p : \max_h p_h < \frac{1}{2} \right\},\ \mathcal{R}_\pm = \left\{ p : \min_h p_h < \frac{1}{2} < \max_h p_h \right\}. \tag{1}$$

We use threshold $1/2$ because the registered binary decision asks which response direction is more probable; the same construction extends directly to a fixed known threshold $c \in (0, 1)$ by replacing $1/2$ with c. The first two regions are *robust positive* and *robust negative*: every registered group lies on the same side of the threshold. The third is *heterogeneous*: one group lies above and another below. Aggregate-mode estimation targets a different quantity; items outside the three certified regions remain unresolved.

These open regions are positive certificate targets rather than an exhaustive partition of $[0, 1]^H$. Coordinates exactly at $1/2$ may prevent finite-time robust certification, so the first-order results in Sections 4–5 impose an explicit off-boundary condition.

2.2 Sequential Certification

Let $A_t \in [H]$ be the group queried at global step t and $N_h(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = h\}$ the number of accepted binary updates from group h. The allocation rule may depend on the observed history, provided it is predictable.

Following the time-uniform confidence-sequence framework (Howard et al., 2021), for each group we maintain $C_{h,n} = [L_{h,n}, U_{h,n}]$ with error level δ_h and $\sum_h \delta_h \leq \delta$. Our implementation uses a Bernoulli beta-mixture confidence sequence with Jeffreys mixing and $\delta_h = \delta/H$. At time t, the common stopping rule declares

$$\text{robust positive} \quad \text{if} \quad \min_h L_{h,N_h(t)} > \frac{1}{2}, \qquad \text{robust negative} \quad \text{if} \quad \max_h U_{h,N_h(t)} < \frac{1}{2}, \tag{2}$$

$$\text{heterogeneous} \quad \text{if} \quad \max_h L_{h,N_h(t)} > \frac{1}{2} \ \text{ and } \ \min_h U_{h,N_h(t)} < \frac{1}{2}. \tag{3}$$

If none holds before the available budget is exhausted, the item is returned as *unresolved*.

The stopping rule is independent of the allocation strategy: round robin, ActiveOrbit, and the adaptive baselines may query different groups, but any positive declaration has the same statistical meaning. Section 4 explains the information requirements induced by these conditions, and Section 5 specifies the ActiveOrbit allocation rule.

2.3 Parser Semantics and Certificate Scope

The observation map is part of the statistical problem. Under a total binary map, every raw output is mapped to $\{0, 1\}$ and p_h is the unconditional registered mean. Under parse-or-abstain, $g(Y) \in \{0, 1, \perp\}$ and abstention ends the item before a Bernoulli update; conditional on continuation, the target is

$$p_h = \Pr(g(Y) = 1 \mid g(Y) \neq \perp,\, h)\,, \tag{4}$$

assuming positive parse probability. Parser choice can therefore change the estimand as well as the effective sample size.

The certificates characterize the registered response distribution rather than factual correctness or semantic equivalence, and their robustness scope is the specified set of groups. The error level δ applies to one item; benchmark-wide family-wise control requires a separate multiplicity design.

3 Aggregate Consistency and Group-Wise Robustness Are Distinct

The group-conditioned target of Section 2 is necessary whenever the intended claim concerns every representation group. Figure 1 shows that aggregate consistency and group-wise robustness differ empirically.

Across the 1,000 synthetic items, aggregate certification is substantially more frequent than robustness within either registered representation axis, and no item satisfies the joint-robust criterion. The gap includes unresolved cases because all-group certification demands sufficient evidence for every group. Certified joint heterogeneity accounts for 1.4% of this study. Together, these results separate aggregate certification from group-wise robustness without conflating unresolved items with heterogeneous behavior.

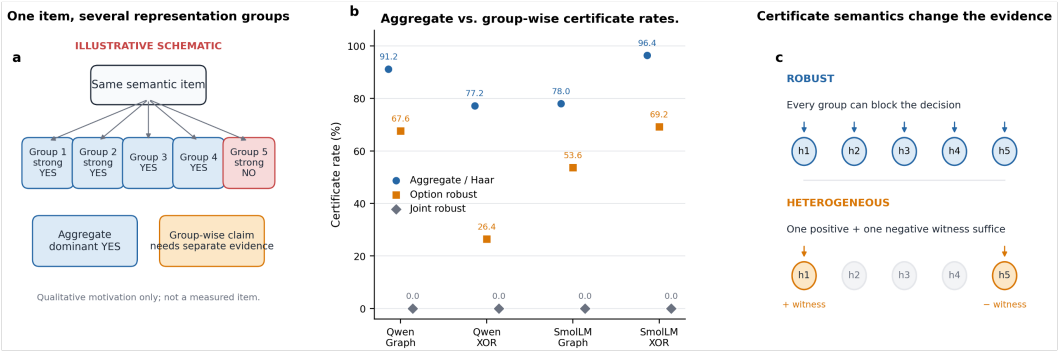

Figure 1: **Aggregate consistency and group-wise robustness are distinct. (a)** Schematic contrast; among the 1,000 studied items, none was both aggregate-certified and certified heterogeneous. **(b)** Aggregate certification is 85.7%, versus 54.2% option-wise and 60.6% parity-wise robustness, with no joint-robust certificates; non-robust items may remain unresolved. **(c)** Robustness is universal over groups, while heterogeneity needs only opposite witnesses.

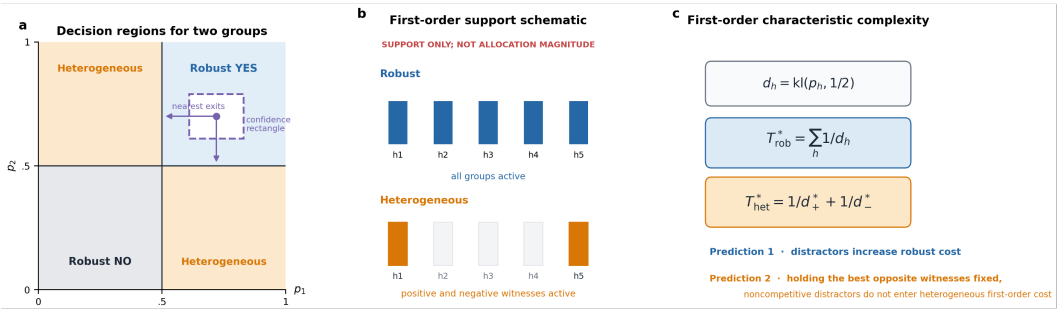

Figure 2: **Certificate semantics determines first-order query geometry. (a)** Robust regions place all groups on one side of $1/2$, whereas heterogeneous instances contain groups on both sides. **(b)** Robust certification has dense first-order support, while heterogeneous certification can use one maximum-information witness from each side. **(c)** The corresponding characteristic times quantify this distinction at first order.

BoolQ provides direct evidence that the distinction can reflect opposite-side group behavior. Across 200 BoolQ (Clark et al., 2019) validation items, aggregate certification is 71.5% and joint robustness is 0%, while 28.5% **receive a certified joint-heterogeneity certificate**. For those items, the data certify groups on opposite sides of the threshold. The complete hierarchy is reported in Appendix C.

4 Certificate Semantics Determines Query Geometry

Section 3 establishes aggregate certification and group-wise robustness as different statistical targets. We now derive how the logical form of the certificate determines where information must be collected. A robust declaration can be invalidated by any single group, whereas heterogeneity needs only one reliable witness on each side of the threshold. Figure 2 summarizes the resulting dense-blocker versus sparse-witness geometry.

4.1 Characteristic Information Rate

Consider fixed finite H and an off-boundary instance, $p_h \neq 1/2$ for all h, and define $d_h := \mathrm{kl}\big(p_h, \frac{1}{2}\big)$. Let $a(p) \in \{+, -, \pm\}$ denote the region and $\mathrm{Alt}(p) := \{q : a(q) \neq a(p)\}$. Following the information-theoretic fixed-confidence pure-exploration formulation (Kaufmann et al., 2016; Gariv-

ier & Kaufmann, 2016; Degenne & Koolen, 2019), for an asymptotic allocation $w \in \Delta_H$, define

$$\Gamma^\star(p) = \max_{w \in \Delta_H} \inf_{q \in \mathrm{Alt}(p)} \sum_{h=1}^{H} w_h \mathrm{kl}(p_h, q_h), \qquad T^\star(p) = \frac{1}{\Gamma^\star(p)}. \tag{5}$$

The closure accounts for the open decision regions. $T^\star$ is the first-order information complexity; empirical sections separately evaluate finite-budget stopping cost.

Proposition 1 (Certificate geometry)**.** *Let $d_h = \mathrm{kl}(p_h, 1/2)$ for fixed finite H and an off-boundary instance.*

For $p \in \mathcal{R}_+ \cup \mathcal{R}_-$,

$$T^\star_{\mathrm{rob}} = \sum_h d_h^{-1}, \qquad w^\star_h = \frac{d_h^{-1}}{\sum_\ell d_\ell^{-1}}, \qquad \Gamma^\star_{\mathrm{rob}} = (T^\star_{\mathrm{rob}})^{-1}.$$

For $p \in \mathcal{R}_\pm$, let $d^\star_+ = \max_{p_h > 1/2} d_h$ and $d^\star_- = \max_{p_h < 1/2} d_h$. Then

$$T^\star_{\mathrm{het}} = (d^\star_+)^{-1} + (d^\star_-)^{-1}, \qquad \Gamma^\star_{\mathrm{het}} = (T^\star_{\mathrm{het}})^{-1}.$$

The robust optimum has positive mass on every group; a heterogeneous optimum can be supported on one maximum-information group from each side, with nonuniqueness under ties.

The derivation is in Appendix A.1. For robust instances, changing any one coordinate can invalidate the claim, so maximizing $\min_h w_h d_h$ requires dense allocation $w^\star_h \propto d_h^{-1}$. For heterogeneous instances, an incorrect declaration must make all groups lie on one side, allowing first-order information to concentrate on the most informative group from each side.

This geometry predicts a regime-dependent advantage whose magnitude depends on the instance. For $H = 2$, the two witnesses are the entire problem and no nonwitness group can be avoided; more generally, a near-boundary witness can dominate the characteristic time.

The geometric core is closely related to one-dimensional convex-hull membership (CHM), which distinguishes means spanning a threshold from means lying entirely on one side and admits an asymptotically optimal sampling algorithm (Qiao & Tewari, 2025). Proposition 1 specializes this geometry to our ternary certificate, retaining robust-positive and robust-negative outcomes separately. Proposition 2 supplies the algorithmic bridge by analyzing the concrete blocker–witness selector used by ACTIVEORBIT.

5 ACTIVEORBIT: TRACKING BLOCKERS AND WITNESSES

Proposition 1 gives the oracle geometry. ACTIVEORBIT implements its structure online: when robustness is plausible, it queries the weakest *blocker*; when heterogeneity is plausible, it concentrates on opposite *witnesses*. It changes allocation while retaining the certificate semantics of Section 2.2.

5.1 ALLOCATION RULE

Let $S_h(t)$ and $N_h(t)$ denote positive and total accepted observations. Define

$$\widetilde{p}_h(t) = \frac{S_h(t) + 1/2}{N_h(t) + 1}, \qquad E_h(t) = N_h(t)\, \mathrm{kl}(\Pi_c(\widetilde{p}_h(t)), 1/2)\,.$$

These quantities guide allocation; certification uses the common confidence-sequence stopping rule.

After forced exploration to $q(t) = \max\{2, \lfloor \sqrt{t/H} \rfloor\}$, ordinary allocation chooses

$$h_t \in \begin{cases} \arg\min_h E_h(t), & \text{if all } \widetilde{p}_h(t) \text{ lie on one side of } 1/2, \\ \arg\min_{h \in \{h_t^+, h_t^-\}} E_h(t), & \text{otherwise,} \end{cases}$$

where $h_t^+ = \arg\max_h \widetilde{p}_h(t)$ and $h_t^- = \arg\min_h \widetilde{p}_h(t)$. Thus same-side evidence tracks the weakest blocker, whereas mixed evidence balances opposite candidate witnesses.

5.2 SELECTOR-SPECIFIC FIRST-ORDER BEHAVIOR

Let $b_\delta = \log(H/\delta)$ and $d_h = \mathrm{kl}(p_h, 1/2)$. For the implemented score, let $\epsilon_c = 10^{-12}$ and assume $p_h \in [\epsilon_c, 1-\epsilon_c] \setminus \{1/2\}$ for all h. We take fixed H and $\delta \to 0$; heterogeneous instances additionally require a unique maximum-information witness on each side.

Proposition 2 (Selector-specific blocker–witness behavior)**.** *Under these conditions and the conditionally fresh Bernoulli model of Section 2.1,*

$$\frac{\tau_{\mathrm{AO}}}{b_\delta} \xrightarrow{p} \begin{cases} \sum_{h=1}^{H} d_h^{-1}, & p \in \mathcal{R}_+ \cup \mathcal{R}_-, \\ (d_+^\star)^{-1} + (d_-^\star)^{-1}, & p \in \mathcal{R}_\pm, \end{cases} \qquad \frac{\tau_{\mathrm{all}}}{b_\delta} \xrightarrow{p} \sum_{h=1}^{H} d_h^{-1} \quad \textit{in both regimes.} \tag{6}$$

The all-group tracker *uses the same stopping rule as* ACTIVEORBIT*, but continues sampling the least-evidenced unresolved group instead of switching to opposite witnesses.*

Thus both selectors have the same dense first-order cost on robust instances, whereas heterogeneous ACTIVEORBIT pays only for the best positive and negative witnesses. Removing witness selectivity therefore has zero first-order penalty on robust instances and an instance-dependent, nonnegative heterogeneous penalty that can vanish, for example, when $H = 2$. Section 6.2 tests this asymmetry.

The implementation-aligned proof is in Appendix A.2. Proposition 2 characterizes first-order normalized stopping times in probability; finite-sample behavior is evaluated empirically.

5.3 ANYTIME-VALID CERTIFICATION

Theorem 1 (Anytime correctness)**.** *If each group has a time-uniform confidence sequence with error level δ_h satisfying $\sum_h \delta_h \le \delta$, then under any predictable allocation policy the probability of ever issuing an incorrect non-unresolved certificate for one registered item is at most δ.*

Predictable sampling preserves the groupwise time-uniform guarantees by optional skipping; a union bound and the stopping logic give the result (Appendix A.3). Under parse-or-abstain, the guarantee concerns the conditional parseable-label mean. It is an item-level false-declaration guarantee; raw-query cost, unresolved probability, benchmark-wide family-wise error, semantic validity, and task correctness are separate quantities.

6 DOES CERTIFICATE GEOMETRY PREDICT LLM QUERY COST?

Proposition 2 predicts a regime asymmetry: removing witness-selective allocation has no first-order penalty on robust instances, but can increase cost on heterogeneous instances when nonwitness groups carry substantial inverse-KL difficulty. We first test this prediction by intervention, then evaluate the same geometry on LLM response streams and against stronger adaptive allocation rules.

6.1 PROTOCOL AND METRICS

We evaluate response streams from BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), and GLUE RTE (Wang et al., 2019) using Qwen3-4B (Yang et al., 2025), SmolLM3-3B (Hugging Face, 2025), and Mistral-7B Jiang et al., 2023. The evaluated cells contain 100 BoolQ items per model, 368 label-balanced PIQA and the full 277-item RTE validation split for Qwen, and 64 PIQA items each for SmolLM and Mistral; budgets are 256 for BoolQ and 192 otherwise.

Our metric is *budget-censored query cost*: unresolved items receive the full budget. The theoretical $T^\star$ instead counts Bernoulli updates. For comparator B, $\mathrm{QR}(B) = 100(C_B - C_{\mathrm{AO}})/C_B$; certificate rate is reported separately, and task-stream intervals are paired over items.

Mechanism and baseline studies use fixed scenarios and common confidence-sequence stopping rules; trial counts and budgets appear in the corresponding captions and Appendix.

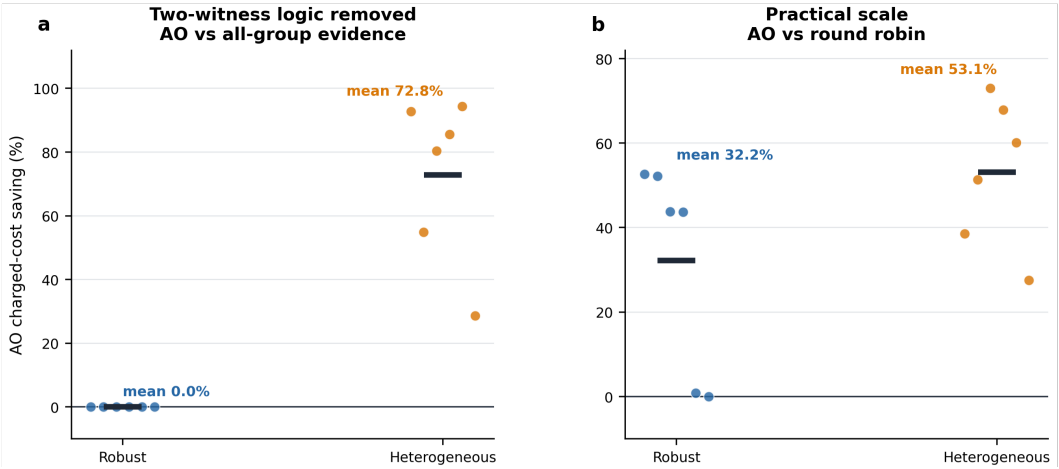

Figure 3: **Controlled test of the blocker–witness prediction.** **(a)** With identical stopping semantics, removing witness selectivity yields 0% reduction in all six robust scenarios and positive reduction in all six heterogeneous scenarios (mean 72.77%); the sign pattern matches the predicted regime effect. **(b)** Gains over round robin occur in both regimes, demonstrating the practical benefit of adaptive allocation.

Table 1: **Main results on frozen LLM response streams.** Query reduction compares ACTIVEORBIT (AO) with round robin (RR), charging unresolved items the full budget. All six paired 95% CIs exclude zero; AO never lowers certificate rate, and common certificates agree.

Model	Task	N	Budget	Queries AO / RR	Query reduction [95% CI]	Cert. AO / RR
Qwen3-4B	BoolQ	100	256	32.14 / 36.57	12.11% [6.09, 19.87]	99.0% / 98.0%
SmolLM3-3B	BoolQ	100	256	180.12 / 214.52	16.04% [9.50, 23.11]	34.0% / 20.0%
Qwen3-4B	PIQA	368	192	37.35 / 51.60	27.63% [23.94, 31.23]	95.7% / 91.8%
Qwen3-4B	RTE	277	192	34.17 / 38.85	12.03% [9.03, 15.20]	97.8% / 96.4%
SmolLM3-3B	PIQA	64	192	52.17 / 82.30	36.61% [27.94, 45.35]	90.6% / 79.7%
Mistral-7B	PIQA	64	192	26.84 / 39.52	32.07% [23.90, 40.27]	96.9% / 96.9%

6.2 CONTROLLED TEST OF THE BLOCKER–WITNESS PREDICTION

We isolate witness selectivity by comparing ACTIVEORBIT with an all-group tracker using the same confidence-sequence stopping rule but continuing to sample the least-evidenced unresolved group.

Figure 3(a) confirms the predicted sign pattern: all six robust scenarios give 0% query reduction, whereas all six heterogeneous scenarios are positive, with mean 72.77%. The reported magnitudes are finite-budget outcomes rather than asymptotic predictions; in one 12-group heterogeneous scenario, the first-order value is 98.12% and the observed reduction is 28.68% (Appendix E.1). Panel (b) supplies practical scale, and Appendix E.2 reports the smaller forced-exploration effects.

6.3 EVALUATION ON LLM RESPONSE STREAMS

Table 1 evaluates whether the allocation effect persists on LLM response streams. Across all six model–task cells, ACTIVEORBIT has lower query cost than round robin and every paired 95% confidence interval excludes zero; reductions range from 12.03% to 36.61%. The certificate rate is never lower, and policies agree on the logical region for every commonly certified item. The parse-or-abstain SmolLM/BoolQ cell remains a high-abstention stress case, illustrating why query cost and certificate rate are reported separately.

Within commonly certified heterogeneous items, AO-vs-RR reductions are 35.92%–43.92% across five task cells and are generally larger than for robust items (Appendix G.2). Because this split conditions on the realized certificate, it provides descriptive rather than causal evidence.

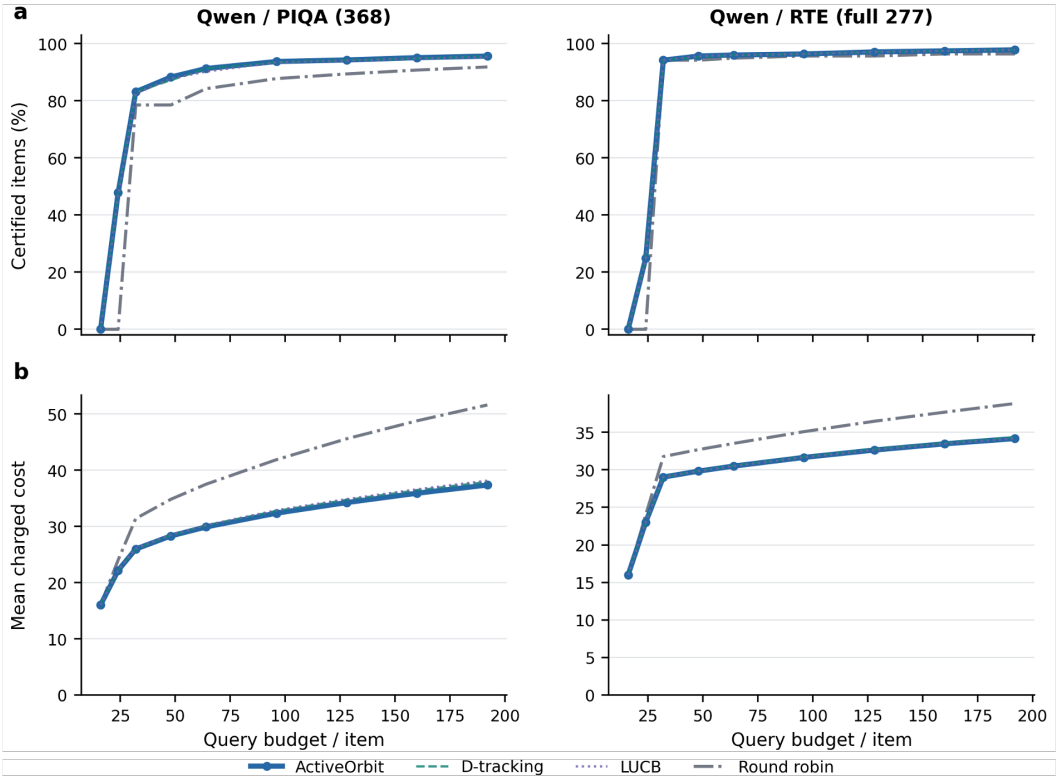

Figure 4: **Finite-budget behavior of adaptive versus uniform allocation.** On Qwen3-4B PIQA-368 and RTE-277 across budgets 16–192, ACTIVEORBIT, D-tracking, and LUCB show similar certificate-rate and query-cost trajectories, while round robin generally requires more queries. The full four-cell diagnostic is reported in the appendix.

Thus, across BoolQ, PIQA, and RTE, ACTIVEORBIT reduces query cost relative to round robin for all three model families. Sample-size extension and independent-seed results appear in Appendix D.

Figure 4 complements the endpoint results by tracing certificate rate and query cost across budgets on the two largest task cells. The adaptive policies behave more similarly to one another than to round robin, demonstrating the finite-budget value of geometry-aware allocation and locating ACTIVEORBIT's contribution in its certificate-specific blocker–witness rule and analysis.

6.4 COMPARISON WITH ADAPTIVE ALLOCATION BASELINES

To distinguish certificate-specific gains from generic threshold-aware allocation, we compare against APT (Locatelli et al., 2016), D-tracking (Garivier & Kaufmann, 2016), LUCB (Kalyanakrishnan et al., 2012), and round robin. We further compare with problem-adapted variants of APGAI (Jourdan et al., 2026) and P-Tracking (Tabata et al., 2023). Table 2 therefore compares progressively more geometry-aware policies.

The ordering of these gaps is informative. ACTIVEORBIT separates most strongly from uniform or less geometry-aware allocation, while its differences from D-tracking, LUCB, and the adapted APGAI/P-Tracking rules are substantially smaller. This ordering supports the central claim that exploiting certificate geometry drives the main allocation benefit.

Additional comparisons with adapted APGAI and P-Tracking further narrow the gap while remaining positive overall. These comparisons isolate allocation under a common confidence-sequence stopping rule. Appendix F.3 reports complementary cached-stream replays: direct-APGAI results are mixed, and cache exhaustion leaves some candidate termination incomplete, so these replays characterize available-cache behavior rather than a universal natural-task ranking.

Table 2: **Comparison with adaptive allocation baselines.** Top: AO query reduction on 18 locked scenarios, split by robust and heterogeneous regimes; gaps narrow for stronger geometry-aware baselines. Bottom: post-freeze comparisons with adapted APGAI and P-Tracking under the same stopping rule; these are allocation-layer comparisons, not end-to-end claims about algorithms.

Comparator	Overall	Robust	Heterogeneous	AO strictly lower
APT	32.79%	7.30%	58.27%	18/18
D-tracking	10.64%	5.43%	15.86%	18/18
LUCB	7.32%	0.00%	14.64%	5/18
Round robin	57.42%	58.33%	56.51%	18/18
Closest adapted baselines (post-freeze)				
Dual-APGAI-FE	5.12% [4.87, 5.36]			17/18
Exploration-matched Dual-APGAI	8.93% [8.65, 9.21]			18/18
Ternary P-Tracking	15.76% [15.48, 16.03]			18/18

7 Related Work and Positioning

Representation sensitivity and certification. LLM evaluation varies with prompt format, option order, demonstrations, and evaluation interfaces (Sclar et al., 2024; Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Zhao et al., 2021; Lu et al., 2022; Hua et al., 2025). Multi-prompt evaluation instead summarizes performance across prompt populations (Polo et al., 2024). Repeated generations support self-consistency and uncertainty estimation (Wang et al., 2023; Manakul et al., 2023; Kuhn et al., 2023; Xiong et al., 2024; Farquhar et al., 2024), while recent work provides formal statistical guarantees for LLM outputs (Cordero-Encinar & Duncan, 2025; Ota et al., 2026; Chaudhary et al., 2025; Feng et al., 2026; Quach et al., 2024). Our target retains specified group structure and identifies the logical region occupied by the vector of group-conditioned response probabilities.

Efficient evaluation. Evaluation cost has been reduced through benchmark subsets or mixtures (Maia Polo et al., 2024; Ni et al., 2024; Li & Xiong, 2025), adaptive evaluation (Zhou et al., 2025; Mohankumar & Khapra, 2022), and lower-cost judging (Wei et al., 2025; Dekoninck et al., 2025). These methods primarily allocate effort across examples, models, or judgments; we allocate repeated queries within an item across representation groups.

Pure exploration and positioning. Our allocation problem builds on fixed-confidence pure exploration and time-uniform inference (Kaufmann et al., 2016; Garivier & Kaufmann, 2016; Degenne & Koolen, 2019; Howard et al., 2021; Kaufmann & Koolen, 2021), with related threshold and adaptive allocation methods including APT, P-Tracking, and APGAI (Locatelli et al., 2016; Tabata et al., 2023; Jourdan et al., 2026). The closest geometric connection is one-dimensional convex-hull membership (CHM), which likewise distinguishes means spanning a threshold from means lying on one side and admits the asymptotically optimal Thompson-CHM algorithm (Qiao & Tewari, 2025). Our contribution specializes established CHM/pure-exploration structure to a ternary certificate and connects it to the blocker–witness selector analyzed in Proposition 2. Thompson-CHM couples sampling and stopping in its published binary CHM formulation, whereas our experiments fix ternary confidence-sequence stopping semantics and vary allocation. Evaluating an adapted Thompson sampling component remains future work.

8 Discussion and Conclusion

Our central conclusion is that certificate semantics determines where evidence must be collected: robustness requires resolving blockers, whereas heterogeneity can be established by opposite witnesses. The mechanism study and six LLM model–task cells support this regime asymmetry and demonstrate practical finite-budget savings. The results establish this principle for registered response distributions, fixed finite groups, and conditionally fresh Bernoulli observations, with parser and decoding choices included in the observation model. The empirical evidence spans BoolQ, PIQA, and RTE with three open-weight 3B–7B model families, and the theory gives first-order behavior for off-boundary instances. Thus, query allocation should be designed around the claim being certified; extensions to richer outputs and logical regions remain open.

AI USE STATEMENT

We used generative AI tools to assist with literature retrieval and reference organization, manuscript editing, organization of experimental results, and research and mathematical reasoning. All AI-assisted text, citations, code, results, and proofs were reviewed and verified by the authors, who made all final methodological and scientific decisions. We take responsibility for the complete content of this work.

ETHICS STATEMENT

This work studies statistical properties of language-model responses and does not involve human participants or the collection of personal data. The experiments use existing benchmark tasks and pretrained language models. The certificates studied here characterize response distributions under registered representation groups; they should not be interpreted as guarantees of factual correctness, fairness, or safety.

REPRODUCIBILITY STATEMENT

We provide an anonymized supplementary archive containing the implementation, experiment configurations, representation-group definitions, random seeds, response/result artifacts used in the reported analyses, and scripts for reproducing the tables and figures. The supplement also includes environment information and integrity manifests for the experimental artifacts. Experimental settings and evaluation metrics are specified in the main paper and Appendix, and complete proofs of the theoretical results are provided in the Appendix.

REFERENCES

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. Hugging Face technical release, 2025. URL https://huggingface.co/blog/smollm3.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. Certifying counterfactual bias in llms. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1c6decac1477fbcc2cbf12d314ce0133-Abstract-Conference.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, pp. 2924–2936, 2019. doi: 10.18653/v1/N19-1300.

Paula Cordero-Encinar and Andrew B. Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in llms. *arXiv preprint arXiv:2510.17472*, 2025. URL https://arxiv.org/abs/2510.17472.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/60cb558c40e4f18479664069d9642d5a-Abstract.html.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. Polyrating: A cost-effective and bias-aware rating system for LLM evaluation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d06537b4b38ccf008a54559d2c56fa23-Abstract-Conference.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of llms with imperfect judges. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/08a67eb74bfb9ded9e949dd52973e997-Abstract-Conference.html.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19889–19899. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1006.

Hugging Face. Smollm3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025. Accessed 2026.

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

Marc Jourdan, Andrée Delahaye-Duriez, and Clémence Réda. An anytime algorithm for good arm identification. *Journal of Machine Learning Research*, 27(19):1–90, 2026. URL https://www.jmlr.org/papers/v27/24-0680.html.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 655–662, 2012.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL https://www.jmlr.org/papers/v22/18-798.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://www.jmlr.org/papers/v17/kaufman16a.html.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tWS-S_aRDRe.

Guohong Li and Deyi Xiong. Towards optimal evaluation efficiency for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14176–14183. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.716.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1690–1698, 2016. URL https://proceedings.mlr.press/v48/locatelli16.html.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Akash Kumar Mohankumar and Mitesh Khapra. Active evaluation: Efficient NLG evaluation with few pairwise comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8761–8781. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.600.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98180–98212, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b1f34d7b4a03a3d80be8e72eb430dd81-Abstract-Conference.html.

Hirofumi Ota, Naoto Iwase, Yuki Ichihara, Junpei Komiyama, and Masaaki Imaizumi. Cite: Anytime-valid statistical inference in llm self-consistency. *arXiv preprint arXiv:2605.05873*, 2026. URL https://arxiv.org/abs/2605.05873.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5c04925674920eb58467fb52ce4ef728-Abstract.html.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of llms. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0707.

Gang Qiao and Ambuj Tewari. An asymptotically optimal algorithm for the convex hull membership problem. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=r8eAwBMtlN.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Koji Tabata, Junpei Komiyama, Atsuyoshi Nakamura, and Tamiki Komatsuzaki. Posterior tracking algorithm for classification bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10994–11022, 2023. URL https://proceedings.mlr.press/v206/tabata23a.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.167.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. RocketEval: Efficient automated LLM evaluation via grading checklist. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/937defc32e8ad2daba66a0e434177ae9-Abstract-Conference.html.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/54dd9e0cff6d9214e20d97eb2a3bae49-Abstract-Conference.html.

Jin Peng Zhou, Christian K. Belardi, Ruihan Wu, Travis Zhang, Carla P. Gomes, Wen Sun, and Kilian Q. Weinberger. On speeding up language model evaluation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a38cfba4c717e5a37b200294615e546e-Abstract-Conference.html.

A Theory Proofs and Scope

This appendix provides the full derivations and technical scope of the results stated in Sections 4 and 5. Unless stated otherwise, we consider a fixed finite number of groups H, threshold $\tau = 1/2$, and conditionally fresh Bernoulli observations as defined in Section 2.1. The first-order selector analysis concerns accepted binary updates rather than raw model queries; parser abstention and finite-stream censoring are treated separately in Appendix A.4.

A.1 Proof of Proposition 1

We first evaluate the characteristic information rate in Eq. equation 5. Recall

$$d_h = \mathrm{kl}\left(p_h, \frac{1}{2}\right), \tag{7}$$

and let

$$\Gamma^\star(p) = \max_{w \in \Delta_H} \inf_{q \in \mathrm{Alt}(p)} \sum_{h=1}^{H} w_h \,\mathrm{kl}(p_h, q_h). \tag{8}$$

Closest opposite-side alternative. For Bernoulli parameters $p, q \in (0,1)$,

$$\frac{\partial}{\partial q}\mathrm{kl}(p,q) = \frac{q-p}{q(1-q)}. \tag{9}$$

Hence, if $p < 1/2$, the divergence is increasing for $q \geq 1/2$ and

$$\inf_{q \geq 1/2} \mathrm{kl}(p,q) = \mathrm{kl}\left(p, \frac{1}{2}\right). \tag{10}$$

The symmetric statement holds for $p > 1/2$ and $q \leq 1/2$. Because the decision regions are strict, the threshold need not belong to the wrong open region itself; the same value is obtained by taking the infimum over its closure.

Robust case. Consider first $p \in \mathcal{R}_+$. Leaving the robust-positive region requires moving at least one coordinate to or below the threshold. For a fixed allocation w, the least expensive alternative therefore changes only the cheapest coordinate:

$$\inf_{q \in \mathrm{Alt}(p)} \sum_{h=1}^{H} w_h \mathrm{kl}(p_h, q_h) = \min_{h \in [H]} w_h d_h. \tag{11}$$

The same expression follows for $p \in \mathcal{R}_-$ by symmetry. Thus

$$\Gamma^\star_{\mathrm{rob}}(p) = \max_{w \in \Delta_H} \min_h w_h d_h. \tag{12}$$

Let

$$c = \min_h w_h d_h. \tag{13}$$

Then $w_h \geq c/d_h$ for every h, so

$$1 = \sum_{h=1}^{H} w_h \geq c \sum_{h=1}^{H} d_h^{-1}. \tag{14}$$

Therefore

$$c \leq \left(\sum_{h=1}^{H} d_h^{-1}\right)^{-1}. \tag{15}$$

Equality is achieved by

$$w_h^\star = \frac{d_h^{-1}}{\sum_{\ell=1}^{H} d_\ell^{-1}}, \tag{16}$$

for which every product $w_h^\star d_h$ is equal. Since every $d_h > 0$ on an off-boundary instance, these weights are unique and strictly positive. Consequently,

$$\Gamma^\star_{\mathrm{rob}}(p) = \left(\sum_{h=1}^{H} d_h^{-1}\right)^{-1}, \qquad T^\star_{\mathrm{rob}}(p) = \sum_{h=1}^{H} d_h^{-1}. \tag{17}$$

Heterogeneous case. Now suppose $p \in \mathcal{R}_{\pm}$ and define

$$P = \left\{ h : p_h > \frac{1}{2} \right\}, \qquad N = \left\{ h : p_h < \frac{1}{2} \right\}. \tag{18}$$

To turn p into a robust-negative instance, every coordinate in P must be moved to the threshold or below. Coordinates already in N need not move. Thus the minimum information required to reach the closure of $\mathcal{R}_-$ is

$$\sum_{h \in P} w_h d_h. \tag{19}$$

Similarly, the minimum information required to reach the closure of $\mathcal{R}_+$ is

$$\sum_{h \in N} w_h d_h. \tag{20}$$

Therefore

$$\Gamma^\star_{\mathrm{het}}(p) = \max_{w \in \Delta_H} \min \left\{ \sum_{h \in P} w_h d_h, \sum_{h \in N} w_h d_h \right\}. \tag{21}$$

Let

$$\alpha = \sum_{h \in P} w_h \tag{22}$$

be the total allocation mass assigned to the positive side. Writing

$$d^\star_+ = \max_{h \in P} d_h, \qquad d^\star_- = \max_{h \in N} d_h, \tag{23}$$

we have

$$\sum_{h \in P} w_h d_h \le \alpha d^\star_+, \tag{24}$$

$$\sum_{h \in N} w_h d_h \le (1-\alpha) d^\star_-. \tag{25}$$

Both inequalities can be attained by assigning all mass on each side to a maximum-information group. Hence

$$\Gamma^\star_{\mathrm{het}}(p) = \max_{\alpha \in [0,1]} \min \left\{ \alpha d^\star_+, (1-\alpha) d^\star_- \right\}. \tag{26}$$

The optimum equalizes the two terms:

$$\alpha^\star d^\star_+ = (1-\alpha^\star) d^\star_-, \tag{27}$$

so

$$\alpha^\star = \frac{d^\star_-}{d^\star_+ + d^\star_-}. \tag{28}$$

Substituting gives

$$\Gamma^\star_{\mathrm{het}}(p) = \frac{d^\star_+ d^\star_-}{d^\star_+ + d^\star_-}, \tag{29}$$

and therefore

$$T^\star_{\mathrm{het}}(p) = (d^\star_+)^{-1} + (d^\star_-)^{-1}. \tag{30}$$

If the maximizer of d_h is unique on each side, the corresponding two-witness support is unique. With ties, the characteristic value is unchanged, but there may be multiple optimal allocations; one optimum can always be chosen with support on a single maximum-information group from each side. This proves Proposition 1.

Positioning. The max–min quantity in Eq. equation 8 is the standard pure-exploration alternative-set geometry. Our result gives its closed-form specialization to the three representation-conditioned regions studied here, exposing the dense-blocker and sparse-witness allocation structures used by ACTIVEORBIT.

A.2 Implementation-Aligned Proof of Proposition 2

We now prove the selector-specific statement used to connect the oracle geometry to the actual ActiveOrbit and `all_group_kl` allocation rules.

Conventions and assumptions. Write

$$b := b_\delta = \log(H/\delta). \tag{31}$$

We take $\delta \to 0$ with fixed finite H. For exact correspondence with the implementation, let

$$\epsilon_c = 10^{-12}, \tag{32}$$

and assume

$$p_h \in [\epsilon_c, 1-\epsilon_c] \setminus \left\{\frac{1}{2}\right\} \qquad \forall h. \tag{33}$$

For heterogeneous instances, we additionally assume a unique maximum-information witness on each side.

Global time t denotes the number of *completed* binary updates. The selector evaluates the exploration floor $q(t)$ before issuing query $t+1$. All statements involving correct-side resolution are made on the simultaneous-validity and sign-stabilization events introduced below. The asymptotic argument uses ideal real arithmetic; implementation-level floating-point effects are outside the theoretical model.

A.2.1 Jeffreys-mixture and empirical KL evidence

For a fixed group h, let

$$S_{h,n} = \sum_{s=1}^{n} Z_{h,s}, \qquad \widehat{p}_{h,n} = \frac{S_{h,n}}{n}, \tag{34}$$

for $n \geq 1$. Define the raw empirical KL statistic

$$K_{h,n} = n\,\mathrm{kl}\left(\widehat{p}_{h,n}, \frac{1}{2}\right). \tag{35}$$

At the boundary $1/2$, the Jeffreys beta-mixture likelihood ratio is

$$M_{h,n} = \frac{2^n B\big(S_{h,n}+\frac{1}{2}, n-S_{h,n}+\frac{1}{2}\big)}{B\big(\frac{1}{2},\frac{1}{2}\big)}. \tag{36}$$

The confidence sequence excludes $1/2$ once the corresponding mixture evidence crosses the boundary $e^b = H/\delta$ on the appropriate empirical side.

Auxiliary Lemma 1 (Jeffreys-mixture approximation). There exist constants $C_0, C_1 < \infty$, independent of h and n, such that

$$|\log M_{h,n} - K_{h,n}| \leq C_0 + C_1 \log(n+1). \tag{37}$$

Proof. Using Eq. equation 36 and the beta–gamma identity, $\log M_{h,n}$ is the Jeffreys-mixture log likelihood relative to the Bernoulli$(1/2)$ null. Applying Stirling bounds to the three gamma terms in the beta function gives the Bernoulli maximized log-likelihood ratio

$$n\,\mathrm{kl}\left(\widehat{p}_{h,n}, \frac{1}{2}\right) \tag{38}$$

plus a Jeffreys-mixture regret of order $\log(n+1)$. The same bound also covers $S_{h,n} \in \{0, n\}$ because the beta function remains finite at the half-integer arguments. This yields Eq. equation 37. □

The actual selector does not use $K_{h,n}$. It uses the Jeffreys-smoothed mean

$$\widetilde{p}_{h,n} = \frac{S_{h,n} + 1/2}{n+1} \tag{39}$$

and the clipping map

$$\Pi_c(x) = \min\left\{1 - \epsilon_c, \max\{\epsilon_c, x\}\right\}. \tag{40}$$

Its evidence score is

$$E_{h,n} = n\,\mathrm{kl}\left(\Pi_c(\widetilde{p}_{h,n}), \frac{1}{2}\right). \tag{41}$$

Let

$$m_b = b^{1/4}. \tag{42}$$

Auxiliary Lemma 2 (uniform first-order equivalence). For every fixed $K < \infty$,

$$\max_{h\in[H]} \sup_{m_b \le n \le Kb} \frac{|E_{h,n} - nd_h|}{b} \xrightarrow{p} 0. \tag{43}$$

The same statement holds with $E_{h,n}$ replaced by $K_{h,n}$.

Proof. Since H is fixed and each group provides conditionally fresh Bernoulli(p_h) observations, the tail strong law implies

$$\max_{h} \sup_{n \ge m_b} |\widehat{p}_{h,n} - p_h| \xrightarrow{p} 0. \tag{44}$$

Moreover,

$$|\widetilde{p}_{h,n} - \widehat{p}_{h,n}| \le \frac{1}{n+1}. \tag{45}$$

Because every p_h belongs to $[\epsilon_c, 1-\epsilon_c]$ and Π_c is non-expansive relative to this interval,

$$\Pi_c(\widetilde{p}_{h,n}) \xrightarrow{p} p_h \tag{46}$$

uniformly over $n \ge m_b$. The function

$$g(x) = \mathrm{kl}\left(x, \frac{1}{2}\right) \tag{47}$$

is continuously differentiable on any compact subset of $(0,1)$ containing these tail values. Hence

$$\sup_{m_b \le n \le Kb} |g(\Pi_c(\widetilde{p}_{h,n})) - g(p_h)| = o_p(1). \tag{48}$$

Multiplying by $n \le Kb$ proves Eq. equation 43. The argument for $K_{h,n}$ is identical. □

A.2.2 First and persistent confidence-sequence crossings

Because the confidence sequences are time-uniform but not necessarily nested, a group may cross the threshold and later reopen after an additional observation. We therefore distinguish first crossing from permanent resolution.

Let $R_{h,n}(b)$ denote the event that after n local observations, group h is certified on its *true* side of $1/2$. Define

$$T_h(b) = \inf\left\{n \ge 1 : R_{h,n}(b)\right\}, \tag{49}$$

$$U_h(b) = \inf\left\{n \ge 1 : R_{h,m}(b) \text{ for every } m \ge n\right\}. \tag{50}$$

Auxiliary Lemma 3 (crossing scale). For every group h,

$$\frac{T_h(b)}{b} \xrightarrow{p} \frac{1}{d_h}, \qquad \frac{U_h(b)}{b} \xrightarrow{p} \frac{1}{d_h}. \tag{51}$$

Proof. Fix $\rho \in (0,1)$. By Lemma 1,

$$\log M_{h,n} = K_{h,n} + O(\log n). \tag{52}$$

By the strong law,

$$\frac{K_{h,n}}{n} \longrightarrow d_h \tag{53}$$

and the empirical sign eventually equals the true sign.

Uniformly for

$$m_b \le n \le (1-\rho)\frac{b}{d_h}, \tag{54}$$

we therefore have, with probability tending to one,

$$\log M_{h,n} < b. \tag{55}$$

For $n < m_b$, the deterministic bound

$$K_{h,n} \le n \log 2 = o(b) \tag{56}$$

together with Lemma 1 also prevents a crossing for sufficiently large b. Hence

$$\Pr\left\{T_h(b) < (1-\rho)\frac{b}{d_h}\right\} \to 0. \tag{57}$$

Conversely, the tail strong law implies that, with probability tending to one, for every sufficiently large

$$n \ge (1+\rho)\frac{b}{d_h}, \tag{58}$$

the empirical sign is correct and

$$\log M_{h,n} > b. \tag{59}$$

Thus the group is correctly resolved at all subsequent local times, and

$$\Pr\left\{U_h(b) > (1+\rho)\frac{b}{d_h}\right\} \to 0. \tag{60}$$

Since $T_h(b) \le U_h(b)$ and ρ is arbitrary, Eq. equation 51 follows. □

A.2.3 Discrete forced-exploration floor

The implemented exploration floor is the integer staircase

$$q(t) = \max\left\{2, \left\lfloor \sqrt{t/H} \right\rfloor\right\}. \tag{61}$$

A continuous derivative approximation is inappropriate at its jump points, so we use a discrete level argument.

Let $F_h(t)$ be the number of times group h has been selected by the forced-exploration branch by global time t, and let

$$F(t) = \sum_{h=1}^{H} F_h(t). \tag{62}$$

Auxiliary Lemma 4 (discrete floor bound). After the finite initialization phase,

$$N_h(t) \geq q(t) - 1 \qquad \forall h, t. \tag{63}$$

Moreover,

$$F_h(t) \leq q(t) + C_0 \tag{64}$$

for an initialization-dependent constant C_0, and hence

$$F(t) = O(\sqrt{Ht} + H). \tag{65}$$

Proof. For integers $k \geq 2$, the floor reaches level k when

$$t = Hk^2. \tag{66}$$

Assume inductively that all groups have been brought to at least level k before the next floor increase. When the floor rises from k to $k+1$, every group is therefore deficient by at most one sample:

$$N_h \geq k = q - 1. \tag{67}$$

Whenever a deficient group exists, the forced branch has priority and increments one such group. After it is incremented once, that group reaches $k+1$ and cannot be deficient again until the floor changes. Thus at most H forced selections are required to repair all deficits created by a level increase.

The next increase occurs after

$$H\big((k+2)^2 - (k+1)^2\big) = H(2k+3) > H \tag{68}$$

global updates. Hence all deficits are repaired before the next increase. This proves Eq. equation 63 by induction.

Each group can be selected by the forced branch at most once per floor-level increase, up to a finite initialization overhead. Therefore

$$F_h(t) \leq q(t) + C_0. \tag{69}$$

Summing over H groups and using

$$q(t) \leq 2 + \sqrt{t/H} \tag{70}$$

gives Eq. equation 65. □

In particular, whenever a stopping time $\tau_b = O_p(b)$,

$$F(\tau_b) = o_p(b) \tag{71}$$

for fixed H.

A.2.4 Stabilization and good events

We next isolate an initial phase after which empirical signs and, when needed, witness identities are stable.

Let

$$m_b = b^{1/4}. \tag{72}$$

Choose a deterministic global time s_b large enough that the floor guarantee in Lemma 4 implies

$$N_h(s_b) \geq m_b \qquad \forall h. \tag{73}$$

One may take $s_b = O(Hm_b^2)$, so

$$s_b = O(b^{1/2}) = o(b). \tag{74}$$

No confidence-sequence crossing can occur before s_b for all sufficiently large b: indeed, for every $n \leq s_b$,

$$K_{h,n} \leq n \log 2 = o(b), \tag{75}$$

and Lemma 1 adds only $O(\log n)$.

Since every group has at least $m_b \to \infty$ samples by time s_b, the tail strong law implies that, with probability tending to one, all empirical means have the correct sign for every local time after m_b. In a heterogeneous instance with unique best witnesses, finiteness of H implies a positive population gap between the best and second-best witness on each side. The same tail convergence therefore implies that the empirical witness identities are also permanently correct after s_b with probability tending to one.

We also require a coarse linear stopping bound.

Auxiliary Lemma 5 (linear stopping scale). For both ACTIVEORBIT and `all_group_kl`,

$$\tau_b = O_p(b). \tag{76}$$

Proof. By Lemma 3, all relevant persistent crossing times $U_h(b)$ are $O_p(b)$. After stabilization, any group receiving $\Theta(b)$ ordinary updates has evidence proportional to its local count by Lemma 2. Before stopping, at least one certificate requirement remains unresolved, so the evidence-balancing selectors cannot continue to assign ordinary updates at a rate exceeding a constant multiple of the $O(b)$ persistent-crossing scales. For robust ACTIVEORBIT, ordinary updates balance the least evidence over the blocker set. For heterogeneous ACTIVEORBIT, they balance only the two stabilized witnesses. For `all_group_kl`, ordinary updates are restricted to unresolved groups and minimize their actual evidence E_h.

Thus the number of ordinary post-stabilization updates before all required persistent crossings is $O_p(b)$. By Lemma 4 the number of forced updates by time t is $O(\sqrt{Ht} + H)$. Consequently, on an event whose probability tends to one,

$$\tau_b \le Cb + C'\sqrt{\tau_b} + o(b) \tag{77}$$

for finite constants C, C'. Solving this quadratic inequality yields $\tau_b = O_p(b)$. □

Fix now an arbitrary $\eta > 0$. By Lemma 5, choose $K = K(\eta) < \infty$ such that

$$\limsup_{b\to\infty} \Pr\{\tau_b > Kb\} \le \eta. \tag{78}$$

Let $\mathcal{G}_b(K)$ denote the intersection of the following events:

1. $\tau_b \le Kb$;
2. no confidence-sequence crossing occurs before s_b;
3. all empirical signs are correct after s_b;
4. in the heterogeneous unique-witness case, the empirical witness identities are correct after s_b;
5. all groupwise confidence sequences simultaneously cover their true means;
6. the uniform evidence approximation equation 43 holds on $m_b \le n \le Kb$;
7. the first and persistent crossings satisfy the windows implied by Lemma 3.

All events except the first have probability $1-o(1)$, while simultaneous confidence-sequence validity has probability at least $1 - \delta = 1 - o(1)$. Therefore

$$\Pr\{\mathcal{G}_b(K)\} \ge 1 - \eta - o(1). \tag{79}$$

All remaining first-order arguments are made on $\mathcal{G}_b(K)$. At the end, η is allowed to decrease to zero.

A.2.5 ROBUST ACTIVEORBIT

Assume p is robust and let $T = \tau_{\mathrm{AO}}$. On $\mathcal{G}_b(K)$, no incorrect heterogeneous declaration is possible. At stopping, every group is therefore correctly resolved on the common side.

For every group,

$$N_h(T) \ge T_h(b) = \frac{b}{d_h} + o_p(b), \tag{80}$$

so

$$T = \sum_h N_h(T) \ge b \sum_h d_h^{-1} + o_p(b). \tag{81}$$

For the matching upper bound, after stabilization the algorithm remains in blocker mode and ordinarily selects a minimum-evidence blocker. Before stopping there is at least one unresolved group g. For such a group, the boundary mixture has not permanently crossed, and on the truncated good event

$$E_g \le b + o(b). \tag{82}$$

An ordinarily selected blocker h has no larger evidence, so

$$E_h \le b + o(b). \tag{83}$$

By Lemma 2,

$$N_h \le \frac{b}{d_h} + o(b) \tag{84}$$

for ordinary allocations. Forced selections contribute only $o_p(b)$ by Eq. equation 71. Consequently,

$$T \le b \sum_h d_h^{-1} + o_p(b). \tag{85}$$

Combining Eqs. equation 81 and equation 85 yields

$$\frac{\tau_{\mathrm{AO}}}{b} \xrightarrow{p} \sum_h d_h^{-1}. \tag{86}$$

A.2.6 HETEROGENEOUS ACTIVEORBIT

Assume now that p is heterogeneous. Let $i^\star$ and $j^\star$ denote the unique maximum-information positive and negative witnesses, respectively:

$$d_{i^\star} = d_+^\star, \qquad d_{j^\star} = d_-^\star. \tag{87}$$

On $\mathcal{G}_b(K)$, after time s_b the empirical signs and witness identities are fixed. Hence every ordinary ACTIVEORBIT query after stabilization is assigned to either $i^\star$ or $j^\star$. All other groups receive only their $o_p(b)$ stabilization and forced-exploration updates.

Because no group crosses before s_b, and because a nonwitness receives only $o_p(b)$ additional observations, Lemma 3 implies that no nonwitness can accumulate the $\Theta(b)$ observations needed for a correct-side crossing before stopping. Thus the eventual heterogeneous certificate must be completed by $i^\star$ and $j^\star$. Therefore

$$N_{i^\star}(T) \ge T_{i^\star}(b) = \frac{b}{d_+^\star} + o_p(b), \tag{88}$$

$$N_{j^\star}(T) \ge T_{j^\star}(b) = \frac{b}{d_-^\star} + o_p(b). \tag{89}$$

This gives

$$T \ge b\left[(d_+^\star)^{-1} + (d_-^\star)^{-1}\right] + o_p(b). \tag{90}$$

For the upper bound, ordinary witness selection balances the two actual evidence scores. Before heterogeneous stopping, at least one of the two witnesses remains unresolved, and hence has evidence at most $b + o(b)$ on the good event. The selected witness has no larger evidence. Lemma 2 therefore bounds the first-order ordinary counts by $b/d_+^\star$ and $b/d_-^\star$, while all nonwitness and forced allocations are $o_p(b)$. Equivalently, once both witnesses have reached their persistent crossing times $U_{i^\star}(b)$ and $U_{j^\star}(b)$, the heterogeneous stopping condition must hold. Hence

$$T \le b\left[(d_+^\star)^{-1} + (d_-^\star)^{-1}\right] + o_p(b). \tag{91}$$

Combining the bounds,

$$\frac{\tau_{\mathrm{AO}}}{b} \xrightarrow{p} (d_+^\star)^{-1} + (d_-^\star)^{-1}. \tag{92}$$

A.2.7 ALL-GROUP KL SELECTOR

We finally analyze the implemented `all_group_kl` ablation. It uses the same stopping conditions as ACTIVEORBIT but, on each ordinary step, selects an unresolved group with minimum actual evidence E_h.

Robust instances. If p is robust, every group must be correctly resolved at stopping. The same first-crossing lower bound as above gives

$$\tau_{\text{all}} \geq b \sum_h d_h^{-1} + o_p(b). \tag{93}$$

Conversely, once every group reaches its persistent correct-side crossing $U_h(b)$, the robust condition holds permanently. Because ordinary selection is restricted to unresolved groups and forced updates are lower order,

$$\tau_{\text{all}} \leq \sum_h U_h(b) + o_p(b) = b \sum_h d_h^{-1} + o_p(b). \tag{94}$$

Thus

$$\frac{\tau_{\text{all}}}{b} \xrightarrow{p} \sum_h d_h^{-1}. \tag{95}$$

Heterogeneous instances: upper bound. The same dense expression is an upper bound even when the true instance is heterogeneous. If every group has reached its persistent correct-side crossing, then at least one positive and one negative group are permanently resolved, so the heterogeneous condition must already have triggered. Therefore

$$\tau_{\text{all}} \leq \sum_h U_h(b) + o_p(b) = b \sum_h d_h^{-1} + o_p(b). \tag{96}$$

Heterogeneous instances: lower bound. The nontrivial part is showing that the all-group selector pays this dense first-order cost even though its stopping rule itself needs only an opposite pair.

Let

$$T = \tau_{\text{all}}, \tag{97}$$

and define the *inclusive* correct-crossing set

$$\mathcal{A}_T = \{h : \text{group } h \text{ has had a correct-side crossing by time } T\}. \tag{98}$$

The definition includes the update at time T.

For every $h \in \mathcal{A}_T$,

$$N_h(T) \geq T_h(b) = \frac{b}{d_h} + o_p(b). \tag{99}$$

Let q be the group whose update at time T first completes the heterogeneous stopping condition. Let

$$s_q \leq T \tag{100}$$

be the last time at which q was selected by the *ordinary* `all_group_kl` branch. The equality $s_q = T$ is allowed.

We first show that

$$s_q > s_b \tag{101}$$

with probability tending to one. Since q is correctly resolved at T,

$$N_q(T) \geq T_q(b) = \Theta_p(b). \tag{102}$$

By contrast,

$$N_q(s_b) \leq s_b = o(b), \tag{103}$$

and the total number of forced selections after s_b is $o_p(b)$. If q had no ordinary selection after s_b, we would therefore have

$$N_q(T) = o_p(b), \tag{104}$$

a contradiction. Hence Eq. equation 101 holds.

Because q is resolved at time T, Lemmas 1–2 imply

$$E_q(T) \geq b - o_p(b). \tag{105}$$

After its final ordinary selection s_q, group q can be selected only by the forced branch. Therefore

$$N_q(T) - N_q(s_q^-) = o_p(b), \tag{106}$$

and the uniform evidence approximation yields

$$E_q(s_q^-) \geq b - o_p(b). \tag{107}$$

At s_q, the actual implementation selects the minimum-evidence unresolved group:

$$q \in \arg\min_{k \in \mathcal{U}(s_q^-)} E_k(s_q^-), \tag{108}$$

where $\mathcal{U}(t)$ denotes the unresolved groups at time t.

Now consider any $h \notin \mathcal{A}_T$. By definition it never has a correct-side crossing by time T. On the simultaneous-validity event it cannot have a wrong-side certificate either. Since $s_q < T$ or $s_q = T$ and $s_q > s_b$, group h is therefore unresolved at s_q^-:

$$h \in \mathcal{U}(s_q^-). \tag{109}$$

Using Eq. equation 108,

$$E_h(s_q^-) \geq E_q(s_q^-) \geq b - o_p(b). \tag{110}$$

Since $E_h \leq n \log 2$ up to lower-order clipping effects, Eq. equation 110 already implies $N_h(s_q^-) = \Theta(b)$. We may therefore invoke Lemma 2:

$$E_h(s_q^-) = N_h(s_q^-) d_h + o_p(b). \tag{111}$$

Consequently,

$$N_h(T) \geq N_h(s_q^-) \geq \frac{b}{d_h} - o_p(b) \qquad \forall h \notin \mathcal{A}_T. \tag{112}$$

Combining Eqs. equation 99 and equation 112, every group satisfies

$$N_h(T) \geq \frac{b}{d_h} - o_p(b). \tag{113}$$

Since H is fixed,

$$T = \sum_{h=1}^{H} N_h(T) \geq b \sum_{h=1}^{H} d_h^{-1} - o_p(b). \tag{114}$$

Together with Eq. equation 96,

$$\frac{\tau_{\text{all}}}{b} \xrightarrow{p} \sum_{h=1}^{H} d_h^{-1}. \tag{115}$$

The good-event argument above gives each lower bound with probability at least $1 - \eta - o(1)$ for arbitrary $\eta > 0$. Taking $b \to \infty$ and then $\eta \downarrow 0$ yields convergence in probability. This completes the proof of Proposition 2.

A.2.8 COROLLARY: FIRST-ORDER QUERY REDUCTION

The robust first-order constants of ACTIVEORBIT and `all_group_kl` are identical, so

$$S_\infty^{\text{rob}} = 0. \tag{116}$$

For a heterogeneous instance,

$$S_\infty^{\text{het}}(p) = 1 - \frac{(d_+^\star)^{-1} + (d_-^\star)^{-1}}{\sum_{h=1}^{H} d_h^{-1}}. \tag{117}$$

Because the denominator contains the two witness terms plus nonnegative additional inverse-KL terms,

$$S_\infty^{\text{het}}(p) \geq 0. \tag{118}$$

The reduction can approach one when substantial inverse-KL difficulty lies in nonwitness groups, but there is no positive instance-independent lower bound. In particular, it is exactly zero for $H = 2$.

A.3 Proof of Theorem 1

For group h, define the simultaneous-coverage event

$$\mathcal{C}_h = \{p_h \in C_{h,n} \text{ for every } n \geq 1\}. \tag{119}$$

By time-uniform validity,

$$\Pr(\mathcal{C}_h^c) \leq \delta_h. \tag{120}$$

The group queried at global time t is predictable with respect to the past. Thus observing a particular group at data-dependent global times amounts to optional skipping of that group's underlying observation stream. The groupwise time-uniform guarantee is therefore preserved under the adaptive allocation rule.

Let

$$\mathcal{C} = \bigcap_{h=1}^{H} \mathcal{C}_h. \tag{121}$$

By a union bound,

$$\Pr(\mathcal{C}^c) \leq \sum_{h=1}^{H} \delta_h \leq \delta. \tag{122}$$

It remains only to check the stopping logic on $\mathcal{C}$.

If robust positive is declared, then

$$L_{h,N_h(t)} > \frac{1}{2} \qquad \forall h. \tag{123}$$

Since $p_h \in C_{h,N_h(t)}$ on $\mathcal{C}$,

$$p_h \geq L_{h,N_h(t)} > \frac{1}{2} \qquad \forall h, \tag{124}$$

so $p \in \mathcal{R}_+$.

Similarly, a robust-negative declaration implies

$$p_h \leq U_{h,N_h(t)} < \frac{1}{2} \qquad \forall h, \tag{125}$$

and hence $p \in \mathcal{R}_-$.

Finally, if heterogeneous is declared, there exist i, j such that

$$L_{i,N_i(t)} > \frac{1}{2}, \qquad U_{j,N_j(t)} < \frac{1}{2}. \tag{126}$$

Coverage implies

$$p_i > \frac{1}{2}, \qquad p_j < \frac{1}{2}, \tag{127}$$

so $p \in \mathcal{R}_\pm$.

Therefore an incorrect non-unresolved declaration is impossible on $\mathcal{C}$. Using Eq. equation 122,

$$\Pr\{\text{ever output an incorrect non-unresolved region}\} \leq \delta. \tag{128}$$

This proves Theorem 1.

The theorem provides item-level control of false region declarations. Unresolved probability, expected stopping time, benchmark-level family-wise error, semantic validity of the representation transformations, and task-label correctness are separate quantities outside this guarantee.

A.4 Parser-Specific Corollaries

The theorem above concerns the binary observations supplied to the confidence sequences. The meaning of their Bernoulli mean depends on the specified parser protocol.

Total binary map. Suppose every raw model output Y is mapped by a fixed function

$$g_0(Y) \in \{0, 1\}, \tag{129}$$

including outputs that would otherwise be considered parse failures. Then

$$p_h = \Pr\{g_0(Y) = 1 \mid h\} \tag{130}$$

is the unconditional mapped-output mean. Under the conditional fresh-draw assumption, Theorem 1 applies directly to these p_h.

Kill on abstention. Now suppose

$$g(Y) \in \{0, 1, \perp\}, \tag{131}$$

and the policy terminates unresolved at its first $\perp$ before applying a Bernoulli update. Assume each group has positive parse probability and define

$$\pi_h = \Pr\{g(Y) = 1 \,|\, g(Y) \neq \perp,\, h\}\,. \tag{132}$$

Conditional on a draw being parseable, the accepted binary observation has mean π_h. Therefore any positive region declaration made before parser killing inherits the same confidence-sequence guarantee with p_h replaced by π_h.

The two protocols certify different estimands: encoding an unparseable output as zero defines an unconditional mapped-output mean, whereas conditioning on parseability defines a conditional mean. This corollary transfers the confidence-sequence guarantee to the latter estimand. Raw model-query cost additionally reflects queries consumed by abstentions, policy termination, and finite-stream censoring, effects outside the ideal infinite Bernoulli-update model of Proposition 2.

A.5 Theory Scope, Edge Cases, and Counterexamples

We collect several cases that delimit the interpretation of the first-order results.

Zero-gain heterogeneous instances. For

$$H = 2, \qquad p = (0.8, 0.2), \tag{133}$$

the two groups are already the required positive and negative witnesses. Hence

$$T^{\star}_{\text{rob-dense}} = d_1^{-1} + d_2^{-1} = T^{\star}_{\text{het}}, \tag{134}$$

and the first-order query reduction is exactly zero.

The gain can be arbitrarily small even for $H > 2$. Consider, for example,

$$p = (0.99, 0.9, 0.4999). \tag{135}$$

The negative witness is extremely close to the threshold and therefore dominates the inverse-KL complexity. The additional positive nonwitness contributes negligible first-order cost relative to that term, making $S^{\text{het}}_{\infty}(p)$ arbitrarily close to zero. Consequently, heterogeneous instances admit no positive instance-independent gain bound.

Tied witnesses. For an instance such as

$$p = (0.8, 0.8, 0.2, 0.2), \tag{136}$$

the characteristic value in Eq. equation 30 remains valid, but there are multiple maximum-information witnesses on both sides. An oracle optimum can still be chosen with two-group support, but neither that support nor the empirical selector's limiting identity is unique. This is why Proposition 2 assumes unique best witnesses for its heterogeneous allocation-convergence statement.

Exact-boundary coordinates. Consider

$$p = (0.8, 0.2, 0.5). \tag{137}$$

The first two groups already provide a valid heterogeneous certificate, so the logical heterogeneous stopping condition can succeed. However, the third coordinate has

$$d_3 = \mathrm{kl}(0.5, 0.5) = 0, \tag{138}$$

and the dense inverse-KL expression is not finite. The all-group selector may also spend substantial effort on this boundary coordinate. Accordingly, the selector-specific first-order theorem is explicitly restricted to off-boundary instances.

Finite-budget interpretation of first-order complexity. The `h12_weak_positive` component-ablation scenario provides a concrete example. The asymptotic dense-versus-witness formula implies a relative reduction of approximately

$$98.12\%, \tag{139}$$

whereas the observed finite-budget query reduction at budget 2,048 is only

$$28.68\%. \tag{140}$$

In that experiment the full method certifies 65.3% of trials, while the all-group ablation certifies only 0.1%. Proposition 2 characterizes $\delta \to 0$ first-order stopping-time constants, while this finite-budget experiment also reflects mixture-boundary corrections, initialization, forced exploration, censoring, and differences in finite-budget power. The theory therefore predicts the *direction and regime structure* of the mechanism ablation; these additional terms determine its finite-budget magnitude.

Mode of convergence. Proposition 2 establishes convergence in probability,

$$\frac{\tau_\delta}{b_\delta} \xrightarrow{p} c(p). \tag{141}$$

An expected-stopping-time limit of the form

$$\frac{\mathbb{E}[\tau_\delta]}{b_\delta} \to c(p), \tag{142}$$

requires additional tail or uniform-integrability control and is outside the proposition.

Query cost versus Bernoulli-update complexity. The theoretical characteristic times count accepted Bernoulli updates and assign equal cost to each update. Token count, latency, energy, and group-dependent costs remain separate resource metrics under this unit-cost abstraction. The experimental metric counts raw model queries and additionally reflects parser killing and finite-stream exhaustion, so it evaluates a different layer of the observation process from $T^\star$.

B Experimental and Implementation Details

This appendix specifies the experimental interfaces that determine what is queried, parsed, counted, and certified. The statistical object in our analysis is the binary observation produced by a fixed model, representation renderer, decoding protocol, and parser. The model layer uses fixed external 4-bit MLX inference weights.

B.1 Models, Tasks, and Evaluation Cells

Table 3 summarizes the principal evaluation cells used in the main paper. The main natural-task results use BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), and the GLUE RTE split (Wang et al., 2019) with Qwen3-4B (Yang et al., 2025), SmolLM3-3B (Bakouch et al., 2025), and Mistral-7B (Jiang et al., 2023). The synthetic hierarchy study additionally uses graph-reachability and XOR-SAT task families with Qwen3-4B and SmolLM3-3B.

The main paper uses *model family names* for readability; the corresponding configuration files record the exact external identifiers and revisions used for reproduction.

Generation seeds. Generation is deterministically keyed by the experiment configurations. For example, the synthetic hierarchy records seeds `5101/5102/5201/5202`, while the stochastic BoolQ evaluation records cell-specific seeds `10701/10901`. The PIQA/RTE item streams and their second-seed replicas likewise retain their generation seeds in the item files and experiment manifests. The configuration and manifest records provide the authoritative model–task-to-seed mapping for every cell.

Table 3: **Principal model-output evaluation cells.** N denotes the number of evaluation items and "Budget" the specified per-item query budget used for the corresponding main endpoint. The PIQA-368 cell is a prespecified label-balanced subset rather than the full validation set; RTE-277 is the complete validation split. Exact external model identifiers, revisions, generation seeds, and configuration hashes are recorded in the experiment configurations and manifests.

Evaluation setting	Model	Task / split	N	Budget	Observation protocol
Synthetic hierarchy	Qwen3-4B / SmolLM3-3B	Graph / XOR	250/cell	32 draws	start-anchored A/B label parser
Stochastic BoolQ evaluation	Qwen3-4B	BoolQ validation	100	256	strict parse-or-abstain
Stochastic BoolQ evaluation	SmolLM3-3B	BoolQ validation	100	256	strict parse-or-abstain
Final natural-task evaluation	Qwen3-4B	PIQA, label-balanced subset	368	192	canonical label parse-or-abstain
Final natural-task evaluation	Qwen3-4B	RTE, full validation	277	192	canonical label parse-or-abstain
PIQA evaluation	SmolLM3-3B	PIQA	64	192	canonical label parse-or-abstain
PIQA evaluation	Mistral-7B	PIQA	64	192	canonical label parse-or-abstain

Table 4: **Representation groups used in the experiments.** Groups are fixed before observing model responses. They are intended to preserve the task label while changing surface representation; semantic preservation is an experimental-design assumption outside the certification theorem. Exact renderers and group enumerations are provided in the released code.

Task	Representation axes	Binary interpretation
Graph	representation permutations; option order	task-specific binary answer
XOR	representation permutations; option order	task-specific binary answer
BoolQ	option order; layout	canonical answer direction
PIQA	fixed option-order groups	canonical `FIRST` vs. `SECOND`
RTE	fixed option-order groups	canonical `FIRST` vs. `SECOND`

B.2 Representation Groups

A representation group is a prespecified subset of prompts sharing a representation condition. The groups define the statistical target itself: the guarantees in this paper are conditional on this finite partition, with semantic equivalence of the associated transformations treated as an experimental-design assumption.

Table 4 summarizes these constructions. The synthetic graph and XOR tasks use representation permutations together with option-order changes. BoolQ uses option-order and layout groups. PIQA and RTE use fixed group-specific prompts and define the binary outcome through a canonical `FIRST`/`SECOND` answer mapping rather than through a literal YES/NO interpretation.

For PIQA and RTE, repeated queries within a group do not correspond to repeatedly sampling a new representation transformation. The group-specific prompt is fixed, and repeated variation is primarily induced by stochastic decoding. This distinction matters when interpreting the Bernoulli abstraction: p_h describes the response distribution induced by the complete generation protocol for group h; fresh prompt paraphrases are not part of this estimand.

B.3 Generation and Sampling Protocols

The model, tokenizer, prompt renderer, parser, and decoding configuration are held fixed within each item-level evaluation. Deterministically keyed seeds make the generated streams reproducible. The statistical theory treats successive accepted outputs as conditionally fresh draws from a stationary group distribution. Independence and stationarity remain modeling assumptions about this pseudorandom generation process rather than properties established by a finite realized stream.

Synthetic hierarchy generation. The synthetic hierarchy study evaluates Qwen3-4B and SmolLM3-3B on graph reachability and XOR-SAT. Each of the four model–task cells contains

250 items and 32 specified uniform-orbit draws per item. The primary synthetic hierarchy study uses deterministic decoding. A separate prespecified sensitivity study uses temperature 0.7; the two Qwen cells pass the parser-quality gate, while the corresponding SmolLM cells are excluded from certificate analysis by that gate.

Stochastic natural-task streams. The primary stochastic BoolQ evaluation uses temperature 0.7 and a strict parser on 100 new disjoint items per model. The PIQA/RTE studies similarly use stochastic response streams under their prespecified multitask protocol. Each endpoint replays a fixed pre-generated stream across policies.

Cached-stream replay. Once generated, natural-task outputs are stored as finite, deterministically indexed streams. Multiple allocation policies can then be evaluated on the same underlying outputs by consuming each group's stream in policy-dependent order. This reduces irrelevant generation noise in policy comparisons while keeping the analysis explicitly bounded to the finite cached stream. The main endpoint comparisons operate within the available specified budgets; supplementary analyses and closest-baseline comparisons reuse the same immutable streams.

Finite-stream scope. A policy may request more observations from a group than remain in its finite cache. We record this separately as stream exhaustion. This issue is particularly relevant to the natural APGAI and P-Tracking comparisons, for which cache-bounded results are reported separately from fully observed stopping-time results.

B.4 Parser Protocols and Abstention Semantics

Parsing is part of the specified observation map and therefore part of the estimand. The evaluation reports the two distinct parser semantics separately.

Total binary-map protocol. In the total binary-map protocol, every raw output is mapped to a binary value. An unparseable output is deterministically encoded as 0. Consequently, the corresponding group parameter is the unconditional mapped mean

$$p_h = \Pr\{g_0(Y) = 1 \mid h\}. \tag{143}$$

Kill-on-abstention protocol. The BoolQ and multitask parse-or-abstain protocol instead uses

$$g(Y) \in \{0, 1, \perp\}. \tag{144}$$

If the parser returns $\perp$, the query is recorded but the policy terminates that item as unresolved before updating the Bernoulli state. Positive certificates obtained before such termination therefore concern the conditional parseable-label distribution

$$p_h = \Pr\left(g(Y) = 1 \mid g(Y) \neq \perp, h\right), \tag{145}$$

as discussed in Appendix A.4.

The strict parser is the primary parser for the main natural BoolQ results. A supplementary leading-token analysis characterizes sensitivity to the parser. Changing the parser can alter both accepted outputs and the estimand, so this analysis estimates observation-map sensitivity rather than an exact paired causal effect.

Unresolved-outcome metadata. Natural-task item records preserve parser-abstention indicators, while the closest-baseline comparisons additionally record stream exhaustion. Unresolved causes are reconstructed deterministically from these flags and the final policy decision, since the two observation protocols use different record schemas. Raw model outputs are retained, and qualitative parser examples follow a fixed selection rule.

B.5 Query Cost and Budget-Censoring Semantics

The experimental efficiency metric is *query cost*. For an item with specified budget B, define

$$C_i(\pi) = \begin{cases} T_i(\pi), & \text{if policy } \pi \text{ certifies by time } T_i(\pi) \leq B, \\ B, & \text{if the item is unresolved at the evaluation endpoint.} \end{cases} \tag{146}$$

Thus unresolved items are assigned the full available budget. This convention prevents a policy from appearing efficient merely because it stops producing certificates on difficult items.

The mean query cost is

$$C(\pi) = \frac{1}{N} \sum_{i=1}^{N} C_i(\pi), \tag{147}$$

and for a comparator B we report relative query reduction

$$\mathrm{QR}(B) = 100 \times \frac{C(B) - C(\textsc{ActiveOrbit})}{C(B)}. \tag{148}$$

Positive values mean that ACTIVEORBIT requires fewer queries under the specified censoring convention.

Certificate rate is always reported separately from query cost. When two policies both certify an item, we additionally compare their final logical region; the main natural cells have complete agreement on common certificates.

Parser killing. Under kill-on-abstention, a parser failure may physically occur before the specified budget is exhausted, but the resulting unresolved item is assigned the full budget in the endpoint cost metric. The endpoint metric therefore measures conservative budget-censored query cost; the literal number of raw generations before parser failure is a different quantity. This observation-process cost also differs from the Bernoulli-update complexity in Proposition 2.

Cached-stream exhaustion. For the main natural comparisons, the specified endpoint lies within the available stream design. In the closest-baseline natural comparisons, some candidate policies exhaust the finite cache before reaching a certificate. The corresponding reports use prespecified candidate-favorable completion bounds, explicitly count exhausted streams, and distinguish these cost bounds from fully observed stopping-time and certificate-rate comparisons.

B.6 Reproducibility and Provenance

All reported experiments use fixed model weights, evaluation configurations finalized before analysis, and deterministically keyed generation seeds. The released artifacts include the exact model identifiers, dataset selections, representation-group definitions, parser configurations, query budgets, generation seeds, cached response streams, and experiment configurations needed to reproduce the reported analyses.

No model is trained or fine-tuned in this work; all experiments use fixed external 4-bit MLX model weights. Supplementary analyses reuse immutable response streams and are identified as stream reanalyses, distinct from independent model-output replications.

C Robustness of the Problem Evidence

On the synthetic mechanism tasks, Figure 1 establishes that aggregate certification can be substantially more permissive than group-wise robustness. Here we examine whether the same distinction appears on a natural-language benchmark and under stochastic decoding. Both checks preserve the central qualitative observation that aggregate and group-wise certificates encode distinct statistical claims.

C.1 Natural BoolQ Certificate Hierarchy

We first repeat the certificate-hierarchy analysis on BoolQ. The evaluation contains 100 balanced validation items for each of Qwen3-4B and SmolLM3-3B, with 32 draws per item. The representation axes are option order and layout.

Table 5 reports the hierarchy after pooling the 200 evaluated items. Aggregate certification is obtained for 71.5% of items. Requiring robustness within either representation axis is substantially

Table 5: **Certificate hierarchy on BoolQ response streams.** The study contains 200 balanced BoolQ validation items in total (100 per model for Qwen3-4B and SmolLM3-3B), with 32 draws per item. Aggregate certification remains substantially more frequent than robustness with respect to either representation axis, and no item satisfies the stricter joint-robust criterion. Joint heterogeneity is certified for 28.5% of items. These results reproduce the aggregate-versus-group-aware distinction on a natural-language task; the rates themselves are specific to this evaluation setting.

Certificate criterion	Certificate rate (%)
Aggregate	71.5
Option-swap robust	54.5
Layout robust	55.0
Joint robust	0.0
Joint heterogeneous	28.5

Table 6: **Temperature-0.7 sensitivity of the synthetic certificate hierarchy.** Each cell contains 50 items with 32 stochastic draws per item. The prespecified parse-quality gate is applied before certificate analysis. Both Qwen3-4B cells satisfy the gate and preserve the main qualitative hierarchy: aggregate certification is more frequent than either group-wise robustness criterion, while joint robustness remains absent. The two SmolLM3-3B cells fall below the prespecified parse threshold (parse rates 91.56% and 90.75%), so no certificate rates are computed for those cells.

Model	Task	Parse rate	Aggregate	Option robust	Parity robust	Joint robust
Qwen3-4B	Graph	$\geq 98\%$	92%	66%	70%	0%
Qwen3-4B	XOR-SAT	$\geq 98\%$	74%	22%	36%	0%
SmolLM3-3B	Graph	91.56%	Excluded by parse-quality gate			
SmolLM3-3B	XOR-SAT	90.75%	Excluded by parse-quality gate			

more restrictive: the option-swap and layout robustness rates are 54.5% and 55.0%, respectively. No item receives the stricter joint-robust certificate across both axes.

At the same time, 28.5% of items receive a joint heterogeneous certificate, showing that certified opposite-side behavior is present in this natural-task setting. As in the main text, the aggregate–robust gap includes both heterogeneous and unresolved items under the stricter group-wise criterion.

The relative prevalence of heterogeneous certificates differs from the synthetic hierarchy matrix, where joint heterogeneity is much rarer. Together, the two studies establish the shared structural conclusion that aggregate and group-conditioned certificates are distinct statistical targets; their different robust/heterogeneous mixtures characterize the respective evaluation settings.

C.2 Temperature-0.7 Sensitivity of the Synthetic Certificate Hierarchy

The primary synthetic hierarchy in Figure 1 uses deterministic decoding. To assess the ordering under stochastic decoding, we also ran a temperature-0.7 sensitivity study on 50 graph and 50 XOR-SAT items for each model, with 32 draws per item.

For Qwen3-4B, both cells satisfy the prespecified parse-quality gate. Table 6 shows that the same hierarchy remains visible under stochastic decoding. On graph reachability, the aggregate certificate rate is 92%, compared with 66% option-wise robustness and 70% parity-wise robustness. On XOR-SAT, the corresponding rates are 74%, 22%, and 36%. Joint robustness remains 0% in both cells.

The stochastic Qwen results extend the aggregate/group-wise separation in Figure 1 beyond deterministic decoding. Together with the BoolQ results above, these experiments support the use of representation-conditioned certification as a distinct evaluation target across both synthetic reasoning tasks and a natural-language benchmark.

Table 7: **Query-efficiency results before and after sample-size expansion.** The smaller Qwen PIQA and RTE cells contain 96 items each; the expanded cells contain 368 prespecified label-balanced PIQA items and the complete 277-item RTE validation set. The direction and approximate magnitude of the query reduction are preserved after expansion, and all paired 95% confidence intervals remain above zero.

Task	Evaluation set	N	Query reduction [95% CI]	Cert. AO / RR
PIQA	Initial	96	28.41% [21.20, 35.02]	97.9 / 93.8%
PIQA	Expanded	368	27.63% [23.94, 31.23]	95.7 / 91.8%
RTE	Initial	96	12.45% [7.22, 18.22]	96.9 / 95.8%
RTE	Full validation	277	12.03% [9.03, 15.20]	97.8 / 96.4%

Table 8: **Independent second-seed replication.** Four outcome-agnostic 32-item sentinel cells are regenerated with an independent generation seed and pooled with their original-seed counterparts. All four pooled query-reduction intervals remain positive, with an overall mean reduction of 27.51%.

Model / task	Two-seed pooled query reduction [95% CI]
Qwen3-4B / PIQA	32.02% [25.54, 37.95]
Qwen3-4B / RTE	17.45% [9.84, 25.28]
SmolLM3-3B / PIQA	32.61% [22.55, 44.26]
Mistral-7B / PIQA	27.96% [19.61, 34.76]
Pooled mean	27.51% [23.38, 31.90]

D Robustness of Query-Efficiency Results

The main natural-task results in Table 1 summarize the largest evaluation cells. Here we test whether the observed query-efficiency pattern is stable to increased sample coverage, an independent generation seed, and changes in the confidence level. We also summarize two additional evaluations that provide complementary evidence for the same allocation effect.

D.1 Sample-Size Extension

The initial multitask evaluation used 96 Qwen items on PIQA and 96 on RTE. The larger-sample analysis keeps the model, representation protocol, query budget, and evaluation rule fixed. PIQA uses a prespecified label-balanced set of 368 validation items, while RTE was expanded to the complete 277-item validation split.

Table 7 compares the original and expanded cells. The relative query reduction remains similar after expansion: Qwen/PIQA changes from 28.41% to 27.63%, while Qwen/RTE changes from 12.45% to 12.03%. The paired confidence intervals remain positive in both expanded cells.

The stability of the point estimates is useful because the two tasks exhibit different absolute levels of query reduction while preserving their respective patterns as coverage increases. The expanded cells provide the Qwen PIQA and RTE endpoints in the main paper.

D.2 Independent Second-Seed Replication

We next test sensitivity to generation randomness. Four 32-item sentinel cells were selected by an outcome-agnostic rule before second-seed generation: Qwen/PIQA, Qwen/RTE, SmolLM/PIQA, and Mistral/PIQA. Each cell was then regenerated with an independent generation seed and analyzed under the same certification protocol.

Table 8 reports the two-seed pooled results. All four cells retain positive query reduction, and every pooled cell-level 95% confidence interval excludes zero. Across the four cells, the pooled mean query reduction is 27.51% $[23.38, 31.90]$.

Table 9: **Sensitivity of the blocker–witness effect to the confidence level.** Across the additional $\delta = 0.05$ and $\delta = 0.20$ evaluations, the query-efficiency advantage from witness-selective allocation remains concentrated in heterogeneous scenarios. The $\delta = 0.05$ and $\delta = 0.20$ runs use six preselected scenarios each, whereas the main $\delta = 0.10$ experiment uses the full 12-scenario suite; the table is therefore intended to compare the direction of the effect rather than the numerical magnitudes across rows.

Confidence level	Scenarios	Mean heterogeneous query reduction
$\delta = 0.05$	6	89.02%
$\delta = 0.10$	12	72.77%
$\delta = 0.20$	6	89.17%

Table 10: **Additional replication evidence for query-efficient allocation.** The empirical-rate replay tests the selector on response structures derived from the fixed synthetic hierarchy study. The prospective stochastic-response experiment evaluates newly generated model streams in four model–task cells. Both evaluations show consistent reductions in the number of queries required for certification.

Evaluation	Setting	Result
Empirical-rate replay	8 cells derived from the synthetic hierarchy study	47.40% mean query reduction; 8/8 matched power
Prospective stochastic-response replication	2 models $\times$ 2 tasks	21.25–34.27% reduction; 4/4 positive

The largest change between individual generation seeds occurs in the Qwen/PIQA sentinel cell (10.65 percentage points), while the pooled estimate remains positive. Together, these replications establish the query-efficiency pattern across two independent decoder seeds.

D.3 Confidence-Level Sensitivity

The mechanism experiment in Figure 3 uses $\delta = 0.10$. We additionally evaluate the same blocker–witness intervention at $\delta = 0.05$ and $\delta = 0.20$ on prespecified scenario subsets.

The qualitative regime pattern is unchanged. At both confidence levels, removing witness-selective allocation produces a query penalty in the heterogeneous scenarios, while the robust scenarios remain consistent with zero first-order advantage from witness selection. The mean heterogeneous query reductions are 89.02% at $\delta = 0.05$ and 89.17% at $\delta = 0.20$.

Thus the central regime asymmetry in Figure 3 persists beyond the nominal $\delta = 0.10$ setting across all three evaluated confidence levels.

D.4 Additional Replication Results

Two additional experiments provide complementary evidence for the query-allocation effect. We summarize them here rather than adding additional main-paper figures.

First, an empirical-rate replay uses groupwise response rates estimated from the fixed synthetic-hierarchy model outputs as the underlying Bernoulli instances. Across all eight replay cells, ActiveOrbit reaches matched decision power and reduces mean query cost by 47.40% relative to the corresponding reference allocation. This experiment tests the selector on rate structures induced by observed model behavior, extending the hand-specified synthetic-probability analysis.

Second, a prospective stochastic-response replication evaluates ActiveOrbit on newly generated graph and XOR-SAT response streams from Qwen3-4B and SmolLM3-3B. All four model–task cells show positive query reduction, ranging from 21.25% to 34.27%. Every paired confidence interval is positive, ActiveOrbit's certificate rate is no lower than the comparator's, and the policies agree on all items certified by both.

Table 11: **Complete component-ablation suite.** All 12 frozen scenarios use 2,000 trials, $\delta = 0.1$, and budget 2,048. ACTIVEORBIT and `all_group_kl` share the same confidence-sequence stopping rule and differ only in allocation. Costs and certificate rates are means across trials. Query reduction is $100(C_{\text{all}} - C_{\text{AO}})/C_{\text{all}}$. All six robust scenarios give 0% reduction, whereas all six heterogeneous scenarios give positive reduction, with a heterogeneous mean of 72.77%.

Scenario	H	Regime	Probability vector p	AO queries	All-group queries	Query reduction	Cert. AO / All
`r3_yes_skewed`	3	Robust	`[.87, .69, .57]`	709.86	709.86	0.00%	99.70% / 99.70%
`r3_no_skewed`	3	Robust	`[.13, .31, .43]`	705.54	705.54	0.00%	99.35% / 99.35%
`r5_yes_one_blocker`	5	Robust	`[.93, .84, .76, .68, .56]`	1081.68	1081.68	0.00%	92.95% / 92.95%
`r5_no_one_blocker`	5	Robust	`[.07, .16, .24, .32, .44]`	1080.12	1080.12	0.00%	92.95% / 92.95%
`r12_yes_two_blockers`	12	Robust	`[.95, .91, .87, .81, .76, .72, .68, .65, .62, .59, .57, .55]`	2031.55	2031.55	0.00%	5.55% / 5.55%
`r12_no_graded`	12	Robust	`[.20, .24, .28, .31, .34, .36, .38, .40, .42, .44, .45, .46]`	2048.00	2048.00	0.00%	0.10% / 0.05%
`h3_boundary_distractor`	3	Heterogeneous	`[.78, .29, .53]`	114.69	1587.64	92.78%	100.00% / 48.70%
`h3_weak_positive`	3	Heterogeneous	`[.57, .23, .45]`	717.38	1588.42	54.84%	98.75% / 63.35%
`h5_one_negative`	5	Heterogeneous	`[.91, .37, .69, .62, .56]`	239.06	1218.48	80.38%	100.00% / 92.10%
`h5_one_positive`	5	Heterogeneous	`[.63, .09, .42, .45, .47]`	284.30	1967.74	85.55%	99.90% / 9.55%
`h12_many_distractors`	12	Heterogeneous	`[.88, .12, .69, .65, .61, .58, .55, .45, .42, .39, .36, .33]`	109.31	1955.63	94.41%	100.00% / 12.80%
`h12_weak_positive`	12	Heterogeneous	`[.56, .18, .49, .48, .47, .46, .44, .42, .40, .38, .36, .34]`	1460.07	2047.15	28.68%	65.30% / 0.10%
Robust mean						0.00%	–
Heterogeneous mean						72.77%	–
Overall mean						36.39%	–

Taken together, the sample-size extension, independent seed replication, confidence-level sensitivity, and additional replications show that the query-efficiency results are stable across cells, sample sizes, seeds, confidence levels, and response-generation settings.

E ADDITIONAL MECHANISM ABLATIONS

The main mechanism experiment in Figure 3 isolates witness-selective allocation by comparing ACTIVEORBIT with an all-group evidence tracker under the same confidence-sequence stopping rule. Here we report the complete scenario-level results and two additional diagnostics that clarify which components account for the observed query-efficiency pattern.

E.1 FULL COMPONENT-ABLATION SCENARIO TABLE

The component-ablation suite contains 12 prespecified Bernoulli scenarios, split evenly between robust and heterogeneous instances. Each scenario is evaluated with 2,000 trials, $\delta = 0.1$, and a maximum budget of 2,048 queries. The comparison between full ACTIVEORBIT and the all-group evidence tracker changes only the allocation rule: both policies use the same groupwise confidence sequences and the same robust/heterogeneous stopping conditions.

Table 11 reports the complete suite. Across the six robust scenarios, the two policies have identical mean query cost in every case, yielding 0% query reduction throughout the robust regime. Across all six heterogeneous scenarios, ACTIVEORBIT uses fewer queries, with a mean reduction of 72.77%. The overall mean query reduction across the 12 fixed scenarios is 36.39%.

The scenario-level table also illustrates the distinction between first-order *regime structure* and finite-budget magnitude. For example, in the 12-group heterogeneous weak-positive-witness scenario, the observed query reduction is 28.68% even though its first-order dense-versus-witness expression is approximately 98.12%. At the available budget, ACTIVEORBIT certifies 65.3% of trials, com-

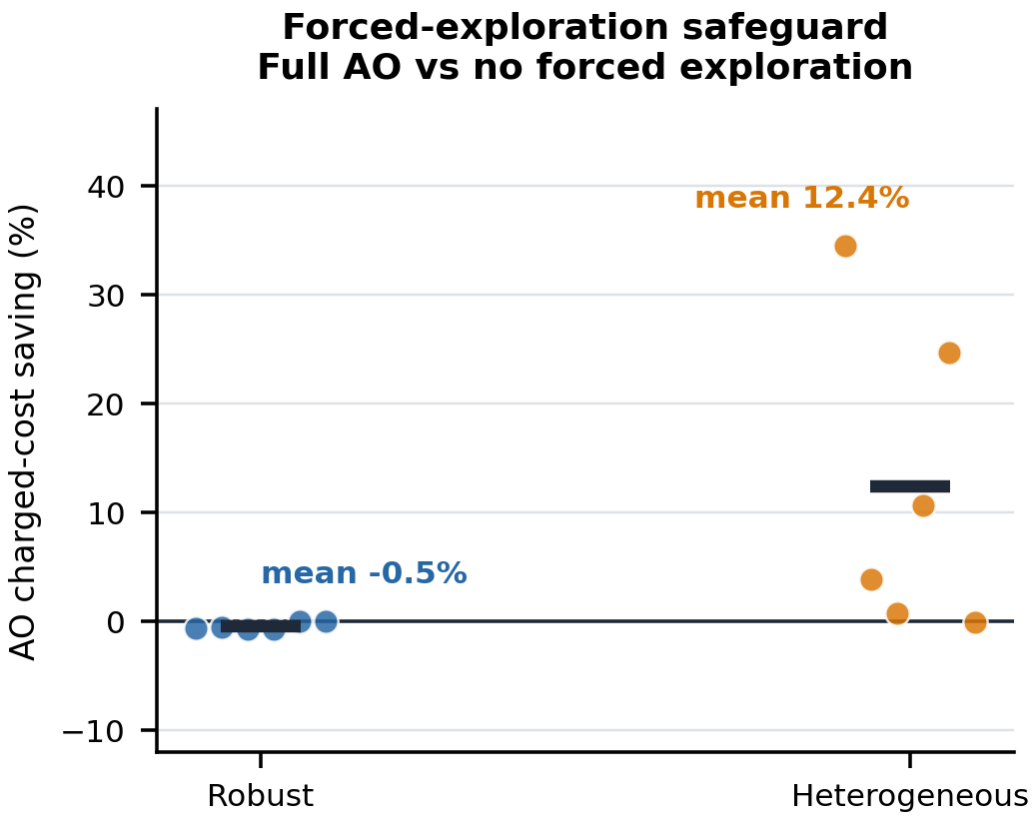

Figure 5: **Contribution of forced exploration on the 12-scenario component suite.** The comparison uses the same prespecified scenarios as Figure 3, with 2,000 trials per scenario, $\delta = 0.1$, and budget 2,048. The vertical axis reports relative query reduction of full ACTIVEORBIT over the no-forced-exploration variant. Across the complete suite, forced exploration yields a mean query reduction of 5.96%; within heterogeneous scenarios the mean is 12.38%. The magnitude is substantially smaller than the effect of removing witness-selective allocation in Figure 3(a), indicating that two-witness allocation is the primary mechanism behind the large heterogeneous-regime gain.

pared with 0.1% for the all-group variant. The finite-budget magnitude additionally depends on mixture-boundary corrections and on how quickly the required evidence becomes attainable within the available budget.

E.2 FORCED EXPLORATION

ACTIVEORBIT combines its blocker/witness rule with the sublinear forced-exploration floor. To determine how much of the empirical advantage comes from this component, we compare the full method with an otherwise identical variant that removes the forced-exploration branch.

Across the full suite, the query reduction of the complete method relative to the no-exploration variant is 5.96%. Within heterogeneous scenarios, the mean reduction is 12.38%. By comparison, replacing witness-selective allocation with the all-group evidence tracker produces a 72.77% mean reduction in the same heterogeneous regime.

This separation is useful for interpreting the method. Forced exploration ensures that every representation group continues to receive samples and supports the consistency argument in Proposition 2, while the much larger empirical heterogeneous-regime effect comes from concentrating ordinary queries on opposite witnesses. Forced exploration therefore provides stabilization, while witness-selective ordinary allocation accounts for the main mechanism gain.

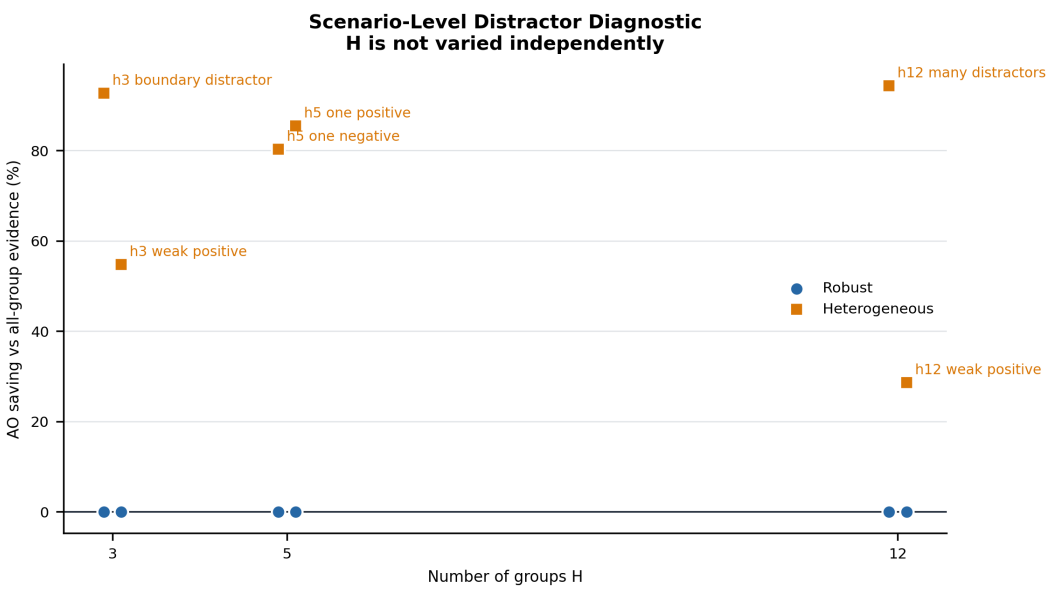

Figure 6: **Scenario-level relationship between group structure and witness-selective query reduction.** Points are the component-ablation scenarios with $H \in \{3, 5, 12\}$. The figure reorganizes existing scenario results to visualize how the benefit of witness-selective allocation varies with scenario structure. Scenarios with the same H exhibit substantially different query reductions, showing the role of instance geometry beyond group count. Because probability gaps and constructions also vary, the figure characterizes instance-level variation; an isolated H-scaling effect would require holding those factors fixed.

E.3 Scenario-Level Distractor Analysis

Proposition 2 predicts that the heterogeneous benefit depends on the inverse-KL difficulty carried by *nonwitness* groups:

$$S_{\infty}^{\mathrm{het}}(p) = 1 - \frac{(d_{+}^{\star})^{-1} + (d_{-}^{\star})^{-1}}{\sum_h d_h^{-1}}. \tag{149}$$

This identifies the relevant quantity as the first-order difficulty outside the two required witnesses, beyond the number of groups alone.

Figure 6 reorganizes the component-ablation scenarios by group count and observed query reduction.

The scenarios span $H \in \{3, 5, 12\}$, and query reduction varies non-monotonically with H. In particular, scenarios with the same number of groups can have markedly different reductions. This behavior is consistent with the characteristic-time expression: an additional group matters when it contributes substantial inverse-KL difficulty to the dense all-group allocation but is unnecessary for the opposite-witness certificate.

The 12-group heterogeneous weak-positive-witness scenario is especially informative in this respect. Its observed finite-budget reduction is 28.68%, substantially below both the heterogeneous-suite mean and its first-order prediction despite the large group count. The evidence therefore supports an *instance-dependent distractor* account: the value of witness-selective allocation depends on the arrangement and difficulty of the nonwitness groups, not on group count alone.

F Adaptive Baseline Details

The main paper compares ActiveOrbit with both general-purpose threshold-aware allocation rules—APT (Locatelli et al., 2016), D-tracking (Garivier & Kaufmann, 2016), and LUCB (Kalyanakrishnan et al., 2012)—and closer problem-adapted baselines based on APGAI (Jourdan

Table 12: **Full strong-baseline holdout.** The suite contains 18 prespecified synthetic scenarios (9 robust and 9 heterogeneous), with $H \in \{2, 4, 8\}$, 5,000 trials per scenario, and budget 4,096. Entries are mean relative query reductions of ACTIVEORBIT with respect to each comparator. "AO strictly lower" counts scenarios in which the mean query cost of ACTIVEORBIT is strictly smaller, excluding ties. The reduction is largest relative to uniform round robin and APT, while the gap narrows for D-tracking and LUCB, which already exploit threshold geometry.

Comparator	Overall	Robust	Heterogeneous	AO strictly lower
APT	32.79%	7.30%	58.27%	18/18
D-tracking	10.64%	5.43%	15.86%	18/18
LUCB	7.32%	0.00%	14.64%	5/18
Round robin	57.42%	58.33%	56.51%	18/18

et al., 2026) and P-Tracking (Tabata et al., 2023). This appendix provides the full baseline breakdown and the corresponding scenario-level analyses.

All synthetic comparisons in this section use the same certificate semantics and confidence-sequence stopping rule. The policies therefore differ in how they allocate queries rather than in what constitutes a successful robust or heterogeneous certificate.

F.1 FULL STRONG-BASELINE HOLDOUT

The strong-baseline holdout contains 18 prespecified Bernoulli scenarios: nine robust and nine heterogeneous, with $H \in \{2, 4, 8\}$. Each scenario is evaluated with 5,000 trials and a maximum budget of 4,096 queries.

We compare ACTIVEORBIT with APT, D-tracking, LUCB, and round robin. Table 12 reports the full regime-level breakdown. The relative query reduction is defined as in Eq. equation 148, so positive values indicate lower mean query cost for ACTIVEORBIT.

The regime breakdown reinforces the blocker–witness interpretation. Relative to APT, the query reduction increases from 7.30% on robust instances to 58.27% on heterogeneous instances. The corresponding D-tracking reductions are 5.43% and 15.86%. For LUCB, the robust mean is tied at 0.00%, whereas the heterogeneous mean reduction is 14.64%.

These results also clarify the role of stronger adaptive baselines. The practical gap decreases as the comparator becomes more responsive to group-specific threshold difficulty. These results locate ACTIVEORBIT's contribution more precisely: the blocker/witness rule is a simple allocation specialized to the certificate's logical structure, while other geometry-aware adaptive rules also capture part of the advantage over round robin.

F.2 CLOSEST ADAPTED BASELINES

We next compare against allocation rules derived from two especially close pure-exploration families: APGAI-style good-arm identification and P-Tracking-style target-allocation tracking.

To make these comparisons align with the ternary certification problem, we adapt only the *allocation* layer while retaining the common representation-conditioned confidence-sequence stopping rule. This design directly compares how the policies distribute queries for the same certificate target. The resulting estimand is allocation-layer performance under shared stopping, distinct from end-to-end evaluations of the published algorithms.

The APGAI comparisons include two variants. DUAL-APGAI-FE combines positive-side and negative-side good-arm scores with forced exploration. The exploration-matched variant uses the same exploration schedule as ACTIVEORBIT, making the comparison focus more directly on the allocation score itself. The P-Tracking comparison uses a ternary extension that tracks target counts appropriate to the three-region problem.

The aggregate gaps are smaller than those against APT or round robin, as expected for baselines that more directly track the relevant pure-exploration geometry. Nevertheless, the paired reductions

Table 13: **Closest adapted allocation baselines on the same 18 prespecified scenarios.** All methods use the common confidence-sequence stopping rule, so the comparison isolates allocation behavior for the same ternary certificate target. Intervals are paired over trials. ACTIVEORBIT has lower mean query cost in 17/18 scenarios relative to Dual-APGAI-FE and in all 18 scenarios relative to the exploration-matched Dual-APGAI and ternary P-Tracking variants.

Adapted comparator	AO query reduction [95% CI]	AO strictly lower
Dual-APGAI-FE	5.12% [4.87, 5.36]	17/18
Exploration-matched Dual-APGAI	8.93% [8.65, 9.21]	18/18
Ternary P-Tracking	15.76% [15.48, 16.03]	18/18

remain positive: 5.12% against Dual-APGAI-FE, 8.93% against the exploration-matched Dual-APGAI variant, and 15.76% against ternary P-Tracking.

Figure 7 shows the corresponding scenario-level comparisons, separated into robust and heterogeneous regimes. The point estimates for exploration-matched Dual-APGAI favor ACTIVEORBIT in all 18 scenarios, although one individual scenario has a paired confidence interval that overlaps zero. The forest therefore shows both the consistency of the aggregate pattern and the finite-trial variation across individual instances.

These experiments compare problem-adapted realizations of the allocation principles under a common certification rule. They support the conclusion that ACTIVEORBIT is a competitive lightweight realization of the blocker/witness allocation geometry.

F.3 NATURAL CACHED-STREAM BASELINE COMPARISONS

We also replay the closest adapted baselines on the natural LLM response streams used elsewhere in the paper. Because adaptive policies may consume the group-specific caches in different orders, some replays reach the end of an available group stream before the policy itself would stop. For these cases, the analysis reports completion bounds in addition to fully observed outcomes.

This distinction lets us retain all available natural-stream evidence without equating a cache-bounded replay with an unconstrained new-generation run.

Direct Dual-APGAI. The direct no-forced-exploration Dual-APGAI replay produces a mixed result. Under the baseline-favorable completion rule, one of the seven headline cells favors ACTIVEORBIT, two favor Dual-APGAI under full-budget completion, and four remain inconclusive. The aggregate baseline-favorable query reduction for ACTIVEORBIT is

$$-1.44\% \; [-3.56, \; 0.35]. \tag{150}$$

The direct no-exploration comparison is therefore mixed: its aggregate interval includes zero and its point estimate favors Dual-APGAI.

Exploration-aligned Dual-APGAI variants. When the APGAI adaptation includes forced exploration, the conservative baseline-favorable completion analysis gives a mean ACTIVEORBIT query reduction of

$$4.67\% \; [4.04, \; 5.34], \tag{151}$$

with positive bounds in all seven headline cells. Three of these cells are fully observed without stream exhaustion.

Under the exploration-matched Dual-APGAI variant, the corresponding baseline-favorable mean reduction is

$$2.00\% \; [1.26, \; 3.02], \tag{152}$$

again with positive lower bounds across the headline cells.

Ternary P-Tracking. The ternary P-Tracking natural replay gives a baseline-favorable mean ACTIVEORBIT query reduction of

$$6.73\% \; [5.31, \; 8.21]. \tag{153}$$

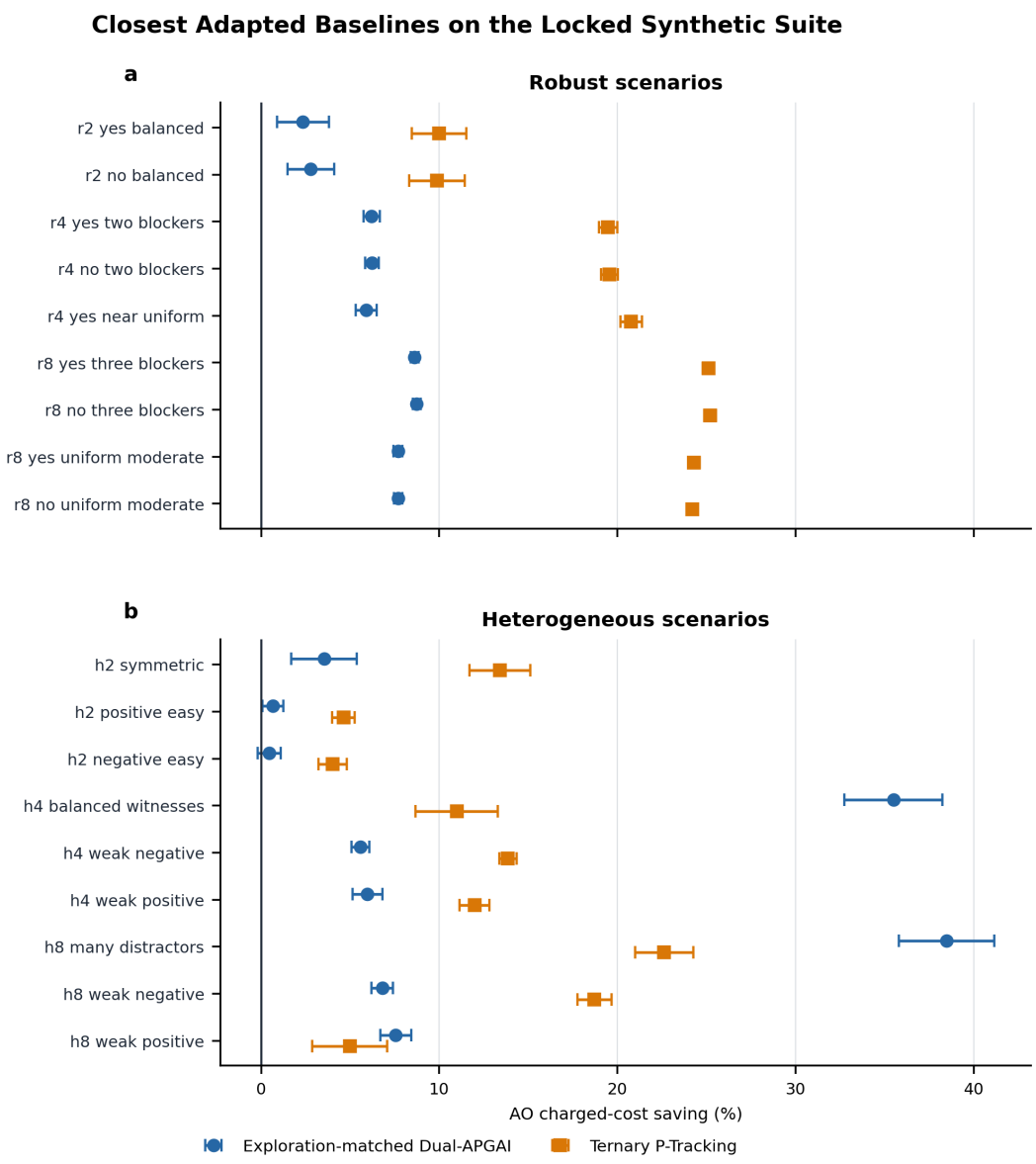

Figure 7: **Scenario-level comparison with the closest adapted allocation baselines.** Each point reports the relative query reduction of ACTIVEORBIT on one of the 18 prespecified scenarios; error bars are paired 95% confidence intervals from the trial-level comparisons. Panels separate robust and heterogeneous scenarios. The plotted comparators are Dual-APGAI-FE, exploration-matched Dual-APGAI, and ternary P-Tracking, all evaluated with the common confidence-sequence stopping rule. Aggregate reductions are 5.12%, 8.93%, and 15.76%, respectively. One exploration-matched Dual-APGAI scenario has an interval overlapping zero, while the overall paired interval remains positive.

Across the replay, 633 policy–item streams exhaust their available cached responses before a fully observed stopping time is obtained. The headline comparison therefore provides a conservative cache-bounded cost result.

The synthetic and natural comparisons serve complementary purposes. On the 18 prespecified synthetic scenarios, complete streams allow direct paired stopping-cost comparisons and show positive aggregate reductions against all three closest adapted baselines. On the natural streams, the conservative replay analysis indicates that the result depends on the precise exploration rule: the direct no-

Table 14: **Natural cached-stream comparisons against the closest adapted allocation baselines.** The same LLM response streams are replayed under each policy. When a candidate policy requires responses beyond an available group cache, the reported baseline-favorable completion rule gives a conservative comparison for ACTIVEORBIT. The direct no-exploration Dual-APGAI replay is mixed, whereas the forced-exploration, exploration-matched, and ternary P-Tracking comparisons yield positive aggregate conservative query-reduction bounds.

Comparator	AO query reduction	Headline cells	Interpretation
Direct Dual-APGAI	-1.44% [-3.56, 0.35]	7	Mixed direct replay; aggregate interval includes zero.
Dual-APGAI-FE	4.67% [4.04, 5.34]	7	Positive baseline-favorable completion bound; three cells fully observed.
Exploration-matched Dual-APGAI	2.00% [1.26, 3.02]	7	Positive baseline-favorable completion bound under matched exploration.
Ternary P-Tracking	6.73% [5.31, 8.21]	7	Positive baseline-favorable completion bound from the cached-stream replay.

exploration APGAI comparison is mixed, while exploration-aligned APGAI and ternary P-Tracking yield small positive aggregate bounds.

Taken together, these results support ACTIVEORBIT as a simple, competitive allocation rule for the representation-conditioned certificate objective. Its largest practical gains arise relative to uniform or less geometry-aware allocation, while comparisons with the closest adaptive baselines establish competitiveness with smaller, exploration-dependent gaps.

G NATURAL-STREAM ANALYSES

The main natural-task experiments report endpoint query cost and certificate rate, while Figure 4 shows how these quantities evolve with budget on the two largest final cells. Here we provide the corresponding full multi-cell budget view and a decision-family analysis that examines whether the natural streams exhibit the same qualitative robust/heterogeneous asymmetry suggested by the theory and mechanism ablation.

G.1 FULL FINITE-BUDGET CURVES

Figure 8 extends Figure 4 to all four PIQA/RTE cells used in the multitask evaluation: Qwen/PIQA, Qwen/RTE, SmolLM/PIQA, and Mistral/PIQA. For each cell, we replay ACTIVEORBIT, D-tracking, LUCB, and round robin over budgets from 16 to 192 queries.

The top row for each cell reports certificate rate, while the bottom row reports mean query cost under the endpoint convention of Appendix B.5. Across the four cells, the adaptive policies generally acquire certificates earlier than round robin and require fewer queries over the practical finite-budget range. ACTIVEORBIT, D-tracking, and LUCB often follow similar trajectories, consistent with the fact that all three respond to group-specific threshold difficulty.

The qualitative pattern is consistent across tasks and model families. At very small budgets, few policies have accumulated enough evidence to certify many items. As the budget increases, the adaptive policies begin to separate from round robin in both certificate rate and query cost. At larger budgets, certificate-rate differences can narrow as more policies resolve the same items, while differences in cumulative query cost remain.

The full four-cell view also reinforces the interpretation of Figure 4: several adaptive policies exploit related threshold geometry, with ACTIVEORBIT providing a lightweight blocker/witness realization of that principle.

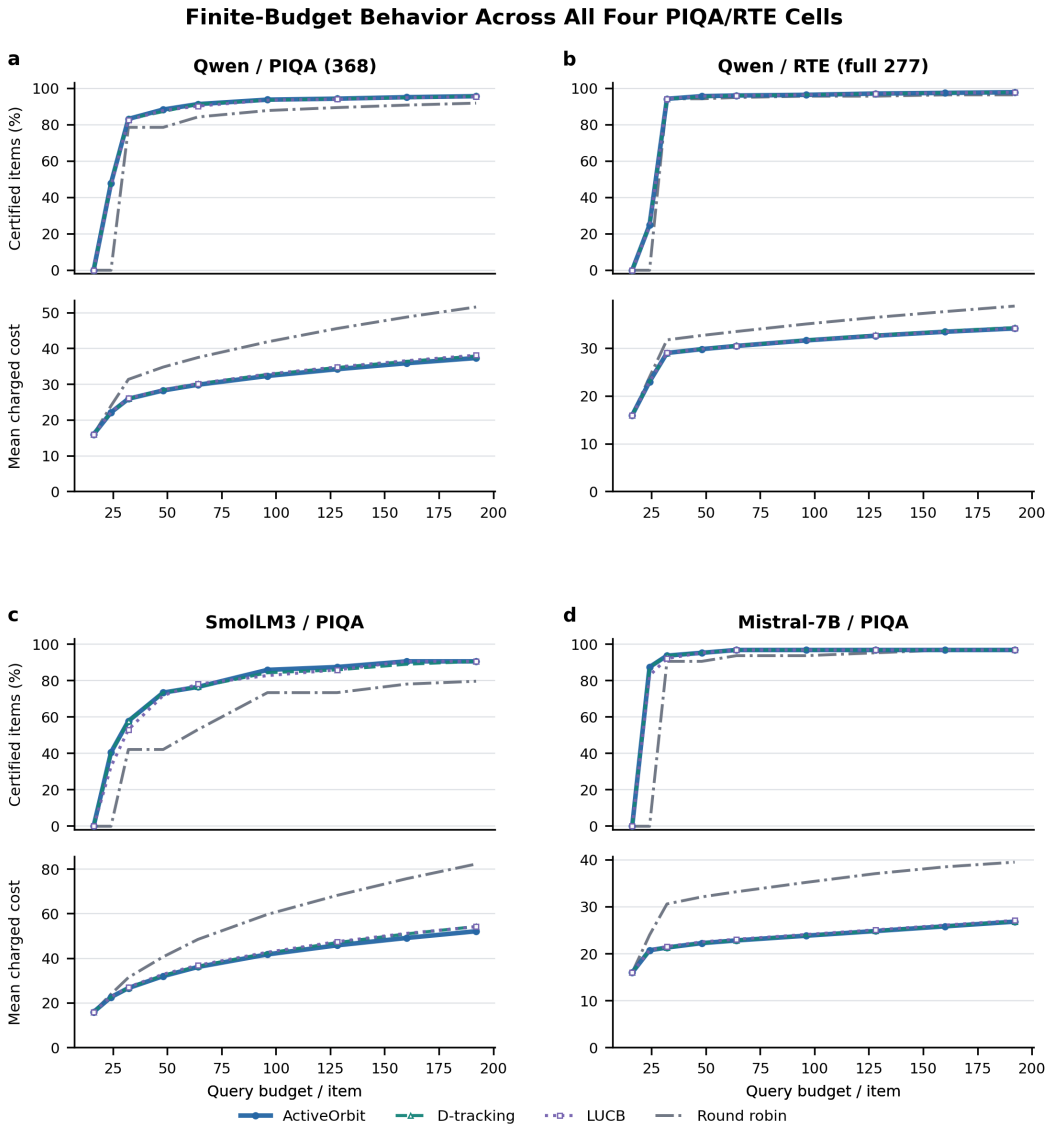

Figure 8: **Full finite-budget analysis across the four PIQA/RTE cells.** The four cells are Qwen3-4B/PIQA, Qwen3-4B/RTE, SmolLM3-3B/PIQA, and Mistral-7B/PIQA. Curves compare ACTIVEORBIT, D-tracking, LUCB, and round robin over budgets 16–192. Certificate-rate panels show the fraction of items resolved by each budget; query-cost panels assign unresolved items the current available budget. Across cells, the adaptive policies typically separate from round robin at moderate budgets, while ACTIVEORBIT, D-tracking, and LUCB often track one another closely. The figure complements the endpoint table by showing when the query-efficiency differences emerge as budget increases.

G.2 DECISION-FAMILY ANALYSIS

The blocker/witness analysis predicts a qualitative regime difference: heterogeneous certificates can terminate once sufficient evidence accumulates for an opposite witness pair, whereas robust certificates require support from every representation group. We therefore examine the natural response streams after grouping items by the logical region returned by both ACTIVEORBIT and round robin.

Across the common-certificate subset, the heterogeneous items show consistently large AO-versus-round-robin query reductions: 35.92% for Qwen/BoolQ, 41.19% for Qwen/PIQA, 36.41% for Qwen/RTE, 43.92% for SmolLM/PIQA, and 37.81% for Mistral/PIQA. The corresponding robust

Table 15: **Natural-stream query reduction within commonly certified heterogeneous items.** Items are grouped by the final logical region returned by both ACTIVEORBIT and round robin, and the table reports the relative AO-versus-RR query reduction within the heterogeneous subset. All five reported cells show reductions above 35%. The corresponding robust subsets generally show smaller reductions, consistent with the blocker–witness distinction.

Model / task	Heterogeneous AO-vs-RR query reduction
Qwen3-4B / BoolQ	35.92%
Qwen3-4B / PIQA	41.19%
Qwen3-4B / RTE	36.41%
SmolLM3-3B / PIQA	43.92%
Mistral-7B / PIQA	37.81%

Table 16: **Query reduction under deterministic decoding.** The temperature-0 BoolQ study contains 50 disjoint items per model with budget 256. The mean query reduction is 7.72%, below the prespecified 10% aggregate criterion. Under exact repeated decoding, within-group response streams contain much less sampling variation, leaving less repeated evidence for an adaptive allocation rule to exploit.

Model	AO-vs-RR query reduction
Qwen3-4B	4.08%
SmolLM3-3B	11.35%
Mean	7.72%

reductions are generally smaller; for example, Qwen/RTE shows a 3.26% reduction among commonly certified robust items.

This pattern is qualitatively aligned with the mechanism experiment: the largest reductions occur on items whose final certificate is heterogeneous, where the stopping condition can be supported by opposite witnesses rather than by all groups.

Because the grouping uses the policies' final common certificate, it provides a descriptive decomposition of natural-stream costs. Figure 3 supplies the direct mechanism test, while this analysis shows the same regime asymmetry in the natural evaluation streams.

H OBSERVATION-PROCESS ANALYSES

The sequential certification problem depends not only on the allocation rule, but also on the observation process that generates repeated group-conditioned responses. Two aspects are especially relevant in practice: whether repeated queries provide meaningful stochastic variation, and how raw model outputs are mapped into binary observations. This section examines both effects on the BoolQ experiments.

H.1 DETERMINISTIC DECODING AS A BOUNDARY CASE

The main stochastic BoolQ experiment uses repeated sampling to accumulate evidence within each representation group. To examine the opposite regime, we also ran a prespecified temperature-0 BoolQ study on 50 disjoint items per model with a query budget of 256.

Table 16 reports the resulting query reductions relative to round robin. Qwen3-4B yields a 4.08% reduction and SmolLM3-3B yields 11.35%, for an unweighted mean of 7.72%. The prespecified aggregate criterion required a mean reduction above 10%, so this study does not meet that criterion.

This result helps characterize the information regime in which active certification is useful. When the prompt, model, and decoding rule are all deterministic, repeated queries to a fixed group can become nearly constant. In that regime, additional queries contribute little new within-group information, so the opportunity for adaptive allocation is correspondingly smaller.

Table 17: **SmolLM/BoolQ parser sensitivity.** Replacing the strict parser with a leading-token parser increases the fraction of outputs mapped to a binary observation and substantially increases the fraction of items certified within budget. The corresponding mean query costs decrease for both ACTIVEORBIT and round robin. The primary paper results use the strict parser; the leading-token result is used to characterize sensitivity to the observation map.

Metric	Strict parser	Leading-token parser
Parse rate	97.28%	99.93%
AO certificate rate	34%	84%
AO mean queries	180.12	75.01
RR mean queries	214.52	108.06
AO-vs-RR query reduction	16.04%	30.58%

Stochasticity is therefore part of the experimental setting: the blocker–witness geometry is unchanged at temperature 0, while the practical value of sequential allocation depends on a response process in which repeated sampling accumulates information.

H.2 STRICT VS. LEADING-TOKEN PARSING

The parser is another component of the observation process. The primary stochastic BoolQ results use the strict parser. We additionally evaluate a leading-token parser on the same 100 SmolLM/BoolQ items as a sensitivity analysis.

Table 17 summarizes the comparison. The parse rate increases from 97.28% to 99.93%. Under the leading-token parser, ACTIVEORBIT's certificate rate rises from 34% to 84%, while its mean query cost decreases from 180.12 to 75.01. Round robin changes from 214.52 to 108.06 mean queries. The resulting AO-versus-RR query reduction under the leading-token parser is 30.58%.

The magnitude of this change illustrates why parsing must be treated as part of the statistical specification. The confidence sequence operates on the binary observations produced by the parser, so changing that map can change both certifiability and the representation-conditioned means being estimated.

This analysis is near-paired: 97.25% of raw outputs align exactly across the two runs, but the predefined exact-output pairing criterion is unmet. It therefore characterizes parser sensitivity without attributing the full numerical difference causally to parser choice.

More generally, these results reinforce the formulation in Section 2.3: the observation map

$$g(Y) \in \{0, 1, \perp\} \tag{154}$$

is part of the certification problem itself.

H.3 QUALITATIVE PARSER EXAMPLES

To make the distinction between parser protocols concrete, we additionally inspect a small set of raw BoolQ generations. The examples are selected from the response files by a deterministic rule.

For presentation, we use the following deterministic selection rule: sort records by item identifier, representation group, and draw index, then report the first examples satisfying each of the following categories: (i) a strict-parser abstention, and (ii) a strict-reject / leading-token-accept case. Additional examples, if space permits, are obtained by continuing in the same order.

These examples make the observation interface tangible; the quantitative evidence about parser behavior comes from the full 100-item sensitivity analysis in Table 17, rather than from individual outputs.

I EMPIRICAL VALIDATION AT AND AWAY FROM THE BOUNDARY

Theorem 1 gives the formal validity guarantee for the sequential certificate. We additionally use Monte Carlo simulations to verify that the implemented confidence-sequence calculations and multi-

Selection type	Item / group / draw	Raw model output	Strict	Leading-token
Both accept (1)	item 0, group 0, draw 0	`A) Yes.`	1	1
Both accept (0)	item 0, group 1, draw 0	`A) No.`	0	0
Both reject (prompt echo)	item 1, group 3, draw 19	`Return exactly one letter: A.`	⊥	⊥
Leading recovers (1)	item 2, group 2, draw 0	`A\n\nExplanation: The Enchanted Forest is operated by the Tofte`	⊥	1
Leading recovers (0)	item 5, group 2, draw 0	`B\n\nExplanation:\nNew Zealand is a separate entity from the Australian continent, known`	⊥	0

Table 18: **Mechanically selected parser examples from the frozen SmolLM3 BoolQ stream.** Records are sorted by item identifier, representation group, and within-group draw index. The table reports the earliest record in each of five predeclared outcome categories: joint acceptance with registered observation 0 or 1, strict rejection recovered by the leading-token parser as 0 or 1, and joint rejection of a prompt echo. Newlines in raw outputs are rendered as `\n`; all other model text is retained verbatim up to LaTeX escaping. The examples are not selected for visual impact.

Table 19: **Monte Carlo validation of the implemented beta-mixture confidence sequence.** Each Bernoulli setting uses 100,000 simulated sequences and prefixes of up to 32 observations. At the decision boundary $p = 1/2$, the recorded anytime failure rate is 5.055%, and all evaluated settings remain below the nominal 10% item-level error level.

Validation setting	Trials	Recorded anytime failure rate
$p = 1/2$	100,000	5.055%
All evaluated probabilities	100,000 / probability	$< 10\%$

group stopping logic exhibit the expected finite-sample behavior. These checks include both off-boundary settings and the exact decision boundary, where certification may remain unresolved for long periods.

I.1 Beta-Mixture Monte Carlo Validation

We first validate the one-group Jeffreys beta-mixture confidence sequence used as the primitive for all group-wise certificates. For each evaluated Bernoulli probability, we simulate 100,000 independent sequences and monitor the confidence sequence for up to 32 observations.

At the exact boundary $p = 1/2$, the recorded anytime failure rate is 5.055%. Across all evaluated probabilities, the empirical failure rates remain below the nominal 10% item-level error level.

This validation tests the numerical implementation of the confidence-sequence primitive used throughout the experiments. Its behavior is consistent with the time-uniform error control assumed by the certificate construction.

I.2 Stratified Certificate Monte Carlo

We next validate the complete robust/heterogeneous decision logic in addition to the single confidence-sequence primitive. The simulation suite includes grouped Bernoulli instances with $H \in \{2, 4\}$, including exact-boundary coordinates, and evaluates 50,000 trials per scenario.

Across the evaluated scenarios, the maximum recorded false-conclusion rate is 0.028%. Boundary configurations are overwhelmingly returned as unresolved rather than being forced into one of the three positive certificate regions.

Table 20: **Monte Carlo validation of the full three-region decision rule.** The evaluation covers $H = 2$ and $H = 4$ grouped Bernoulli instances, including coordinates exactly at the $1/2$ threshold. Across 50,000 trials per scenario, the maximum observed false-conclusion rate is 0.028%. Exact-boundary cases are typically left unresolved, as intended by the open certificate regions.

Validation property	Result
Group counts	$H \in \{2, 4\}$
Trials per scenario	50,000
Includes exact-boundary cases	Yes
Maximum false-conclusion rate	0.028%
Boundary behavior	Predominantly unresolved

Table 21: **Exact-boundary stress test of the confidence-sequence primitive.** For an item-level error level $\delta = 0.1$, the corresponding groupwise levels are δ/H. Each setting uses 100,000 Bernoulli$(1/2)$ trials with prefixes monitored up to 4,096 observations. The observed crossing rates remain below their corresponding groupwise error levels.

H	Per-group δ_h	Trials	Boundary crossing rate
2	0.0500	100,000	3.793%
4	0.0250	100,000	1.929%
8	0.0125	100,000	1.041%

This experiment checks the composition of groupwise confidence sequences with the robust-positive, robust-negative, and heterogeneous stopping conditions. In particular, exact-boundary cases remain predominantly unresolved under the implemented stopping logic.

I.3 Exact-Boundary Primitive Validation

Finally, we stress-test the confidence-sequence primitive specifically at $p = 1/2$ over a substantially longer prefix horizon. We consider the per-group error levels induced by an item-level $\delta = 0.1$ allocation,

$$\delta_h = \frac{\delta}{H}, \tag{155}$$

for $H \in \{2, 4, 8\}$. Each setting uses 100,000 Monte Carlo trials and monitors prefixes through 4,096 observations.

The observed boundary-crossing rates are 3.793%, 1.929%, and 1.041% for $H = 2$, 4, and 8, respectively.

The largest observed ratio between the empirical boundary-crossing rate and its nominal groupwise error level is 0.8328. Thus the longer-horizon boundary validation is consistent with the intended time-uniform behavior across all three tested values of H.

This experiment validates behavior at the decision boundary, where finite-time termination is outside the guarantee. Because the robust and heterogeneous certificate regions are defined by strict inequalities, a coordinate with $p_h = 1/2$ may remain unresolved even as the query budget grows. The resulting high unresolved frequency is the expected behavior under the open certificate regions.

J Negative Results and Claim Boundaries

Several prespecified experiments and supplementary analyses produced outcomes that determine how broadly the main results apply. We consolidate them here to make the paper's evidential boundaries explicit.

J.1 Prespecified and Scientifically Informative Negative Results

Table 22 summarizes the principal outcomes that directly calibrate the scope of the claims made in the main paper.

Table 22: **Scientifically informative negative results and their claim boundaries.** Each outcome identifies the precise conclusion supported by the corresponding experiment. Together they delimit the main claims that the aggregate and group-conditioned certification targets differ, certificate semantics induce blocker/witness query geometry, and ACTIVEORBIT exploits that geometry on the evaluated stochastic response streams.

Result	Status	Observation	Claim boundary
SmolLM temperature-0.7 synthetic sensitivity	Parse-quality gate	The graph and XOR-SAT cells obtain parse rates of 91.56% and 90.75%, below the prespecified parser-quality threshold. The cells are therefore stopped before certificate analysis.	Certificate results cover model–decoding cells satisfying the specified observation-quality requirements; the two excluded cells document the protocol's variation across model–decoding combinations.
Temperature-0 BoolQ	Prespecified endpoint	AO-vs-RR query reductions are 4.08% for Qwen3-4B and 11.35% for SmolLM3-3B, with a mean of 7.72%. The prespecified aggregate $> 10\%$ query-reduction criterion is not met.	The supported query-reduction result concerns the evaluated stochastic streams; the deterministic setting falls below the $> 10\%$ criterion and is consistent with the observation-process analysis in Appendix H.1: deterministic repetition provides less new within-group sampling information.
Strict vs. leading-token parser	Sensitivity analysis	On SmolLM/BoolQ, the leading-token parser substantially changes parse rate, certificate rate, and query cost. Raw outputs align exactly in 97.25% of cases, but the predefined exact-output pairing condition is not satisfied.	The comparison establishes observation-map sensitivity; an exact causal effect of replacing the parser would require outputs satisfying the exact-pairing criterion.
Direct no-FE Dual-APGAI natural replay	Cached-stream comparison	Under baseline-favorable completion, only 1 of 7 headline cells favors ACTIVEORBIT; 2 favor the comparator under full-budget completion and 4 remain inconclusive. The aggregate AO query reduction is -1.44% $[-3.56, 0.35]$.	The evidence supports competitiveness with geometry-aware adaptive baselines, with no uniform advantage over APGAI-style allocation on natural streams.
Ternary P-Tracking natural replay	Cached-stream comparison	The baseline-favorable analysis gives an AO query reduction of 6.73% $[5.31, 8.21]$, but the replay contains 633 stream-exhaustion events and no headline cell with zero exhaustion.	The natural P-Tracking result establishes a cache-bounded cost bound; a fully observed stopping-time or certificate-rate comparison requires nonexhausted streams.
12-group heterogeneous weak-positive-witness scenario	Finite-budget comparison	The first-order dense-versus-witness expression gives an approximately 98.12% asymptotic query reduction, whereas the observed reduction at budget 2,048 is 28.68%.	The first-order analysis predicts allocation support, regime asymmetry, and limiting query complexity; lower-order and censoring effects determine the finite-budget magnitude.

The ledger separates three kinds of evidential boundary. First, observation-process results show that decoding and parsing determine what repeated evidence is available to the sequential procedure. Second, the closest-baseline replays support competitiveness while ruling out a uniform-superiority interpretation across every adaptive allocation. Third, the 12-group heterogeneous weak-positive-witness scenario separates the asymptotic mechanism claim from an exact finite-budget prediction.

The direct mechanism evidence comes from the all-group evidence-tracker intervention in Figure 3, while the natural-task experiments establish finite-budget query-efficiency behavior under the specified observation protocols.