

Sequential Sampling for Binary Classification: Two LLMs are (Almost) All You Need

Guokai Li, Jiaxin (Alys) Liang

McGill University, Montreal, Quebec H3A 1G5, Canada
{guokai.li, alys.liang}@mcgill.ca

Mo Liu

University of North Carolina, Chapel Hill, NC 27599, USA
mo.liu@unc.edu

Yanzhe (Murray) Lei

Queen’s University, Kingston, ON K7L 3N6, Canada
yl64@queensu.ca

Stefanus Jasin, Fenghua Yang, Preet Baxi

University of Michigan, Ann Arbor, MI 48109, USA
{sjasin, yfenghua, preetb}@umich.edu

We study how to orchestrate multiple large language models (LLMs) for reliable binary classification when models differ in accuracy, monetary cost, and response time. A decision-maker adaptively chooses which LLM to query, updates its belief after each response, and stops once sufficient confidence has been reached. Our results reveal a sharp contrast between exact and near-optimal orchestration. In the worst case, an exactly optimal policy may need to query every available LLM. Yet under high reliability requirements, a simple policy using at most two LLMs is asymptotically optimal. The policy identifies one specialist for each possible outcome and adaptively switches between them according to the accumulated evidence. We derive a universal lower bound on the achievable cost and show that this two-LLM policy approaches the bound at the best possible rate. Two LLMs are not merely sufficient—they can be essential. When models have complementary strengths, relying on the best single LLM can be arbitrarily more costly than using two specialists. How the models are queried also matters: committing to a fixed query plan can waste substantial resources because it cannot stop early or redirect queries as evidence accumulates. In a weak-signal regime, we analytically show that the best static design costs four times as much as the optimal sequential policy. Numerical experiments using 18 LLMs on a public benchmark show that these benefits remain substantial at practically relevant confidence levels. The main message is simple: near-optimal LLM orchestration need not require coordinating a large portfolio of models. Identifying two complementary specialists and switching between them as evidence evolves can capture most of the value of adaptive inference.

1. Introduction

A defining feature of modern AI systems is their reliance on *test-time compute*: allocating more computational resources at inference to improve the quality of the output. While classical machine learning invested almost all effort in training, the frontier of large language model (LLM) deployment has shifted decisively toward inference-time strategies. Reasoning models such as OpenAI’s o1 routinely generate dozens of candidate solutions per query, then aggregate them via majority voting or verification (Snell et al. 2025, Wu et al. 2025). Agentic workflows route tasks through

cascades of specialized models and tools, each invocation adding cost and latency (Wang et al. 2025a). Indeed, this multi-source orchestration is becoming the default architecture for frontier systems: OpenAI’s GPT-5, for instance, integrates multiple LLM models behind a single router that dynamically selects which model to invoke for each query (OpenAI 2025). These developments have made *scaling test-time compute* one of the most pressing questions in AI, and the question is fundamentally operational: The key challenge is not to select a single model to deploy, but *how to orchestrate multiple, heterogeneous information sources sequentially* to reach a reliable decision as quickly and cheaply as possible.

More broadly, many digital platforms face a similar challenge of orchestrating multiple information sources. For example, in a modern payment-fraud pipeline, a single transaction may pass through device fingerprinting, behavioral scoring models, merchant risk engines, and third-party verification services before a final verdict is reached (Abdallah et al. 2016). In cybersecurity, a single suspicious event may pass through multiple detection layers—anomaly detectors, intrusion-detection systems, endpoint scanners, and forensic analysis tools—before an analyst decides whether to escalate (Khraisat et al. 2019). Content-moderation platforms screen posts through keyword filters, lightweight classifiers, contextual language models, and human reviewers, escalating ambiguous items through successive rounds of review (Gorwa et al. 2020). In all of these examples, the decision-maker faces the same sequential problem: For each incoming case there is a single, fixed but unknown ground truth, and the system must resolve this uncertainty by querying a sequence of tools, each of which returns a noisy signal at some cost and after some delay.

Despite the practical importance of this orchestration problem, the existing literature has not yet provided enough understanding. Recent work on test-time compute has shown that strategies such as repeated sampling, adaptive self-consistency, verifier-guided selection, self-correction, search, and debate can substantially improve reasoning quality (Wang et al. 2022, Aggarwal et al. 2023, Li et al. 2024, Chen et al. 2024a, Snell et al. 2025, Zhang et al. 2025, Huang et al. 2026a). However, these works largely focused on testing the empirical performances without providing structural understanding for the sequential decision of *which* resource to query and *when* to stop.

At the other end, classical sequential testing and experiment-design models provide the mathematical backbone for adaptive information acquisition (Wald 1945, Chernoff 1959, Naghshvar and Javidi 2013). However, these works were not developed for test-time compute in LLMs and therefore do not capture practical considerations such as heterogeneous query costs, response latency, and asymmetric diagnostic power. These considerations are critical when optimizing system performance. Figure 1 illustrates this heterogeneity among selected LLMs by comparing model intelligence, cost efficiency, and output speed. The horizontal axis reports the cost per Intelligence Index task in USD on a logarithmic scale, the vertical axis reports the Artificial Analysis Intelligence

Index (Artificial Analysis 2026), and the color represents output speed in tokens per second. The figure highlights that model capability does not translate into a uniform ordering across operational dimensions: models with similar Intelligence Index values can differ substantially in cost efficiency and generation speed.

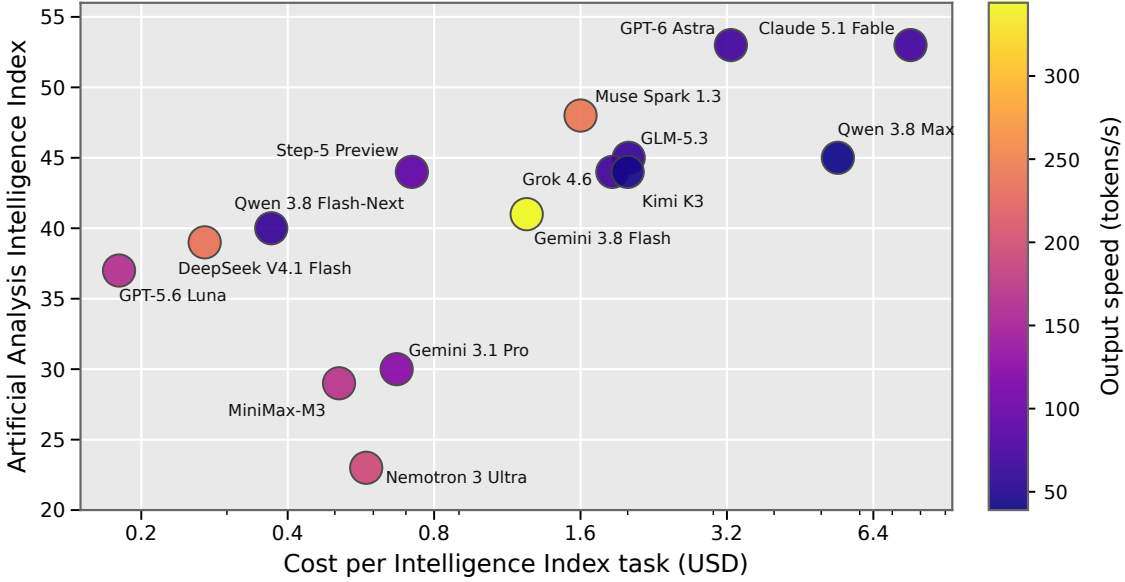


Figure 1 Cost–intelligence–speed trade-offs for selected LLMs. Model names are displayed next to their corresponding bubbles.

This paper takes a first step toward filling this gap. We formulate a sequential hypothesis testing model in which a decision-maker faces a *binary latent state* $\theta \in \{A, B\}$ and may adaptively query any of M information sources. Each source j is characterized by three primitives: A per-query monetary cost c_j , a random waiting time (i.e., latency) with a mean μ_j , and a pair of asymmetric accuracies $(\gamma_{j,A}, \gamma_{j,B})$ with $\gamma_{j,\theta} \in (1/2, 1)$ that determine its diagnostic power under each hypothesis $\theta = A$ or B . The decision-maker updates a posterior belief after every query in a Bayesian manner and stops once the posterior error probability falls below a pre-defined threshold α . The objective is to minimize the sum of expected monetary costs and a convex penalty on total waiting time, subject to the error constraint. Although the model is deliberately stylized, it captures the essential features shared by all the examples described above. Moreover, we are able to conduct a sharp asymptotic analysis in the high-confidence regime where error tolerance α approaches zero. This high-confidence regime is operationally relevant since digital platforms support many queries in real-time, and practical systems increasingly operate under stringent reliability requirements.

Our model does not attempt to capture open-ended answer generation or full agent-environment interaction. Instead, we focus on the task-level decision of how to spend additional test-time compute on one question with two possible answers when the solution quality needs to be improved.

In many deployed AI systems, this is exactly where decision-makers face the most consequential trade-offs between reliability, latency, and cost: A platform may need to decide whether to accept or reject a generated answer, whether one response is better than the other, whether a code patch is ready to execute, whether an agent’s action is safe enough to take, or whether a case should be escalated to a more expensive LLM or a human reviewer. In such settings, the latent state is effectively binary, but the platform may have access to several heterogeneous information sources: different LLMs, prompting strategies, verifiers, or human checks. These sources differ in price, latency, and statistical power.

Our analysis begins by asking what any admissible policy must achieve in order to meet the target error level. More precisely, we establish a universal lower bound on the total expected cost incurred by any admissible policy. Using a change-of-measure argument and martingale techniques, we show that the posterior-threshold stopping rule forces every policy to accumulate, under each hypothesis, on the order of $\log(1/\alpha)$ units of statistical evidence. This requirement is further translated into two deterministic information-budget constraints, one for each hypothesis. As a result, we propose a convex program whose optimal value constitutes the universal lower bound. The convex program has a natural interpretation as a query-allocation problem: It determines how the expected number of queries should be distributed across the M information sources under each hypothesis in order to meet the error requirements while minimizing the combined monetary cost and delay penalty. Importantly, this lower-bound program reveals a simple structural property. Because the program contains only two linear information constraints and a convex objective, its optimal solution can be supported on at most one source for each hypothesis. Consequently, there exists an optimal policy that relies on at most two information sources, regardless of how many are available.

Inspired by this property, we then propose a simple class of *sign-based two-source* policies: At each step, the decision-maker tracks the cumulative log-likelihood ratio and queries an A -specialist when the evidence favors hypothesis A , and a B -specialist when it favors B , stopping when the likelihood ratio crosses the posterior thresholds. Despite its simplicity, we show that an appropriately chosen pair of specialists achieves the universal lower bound up to an additive remainder. Moreover, the order of the remainder is tight: Even the optimal policy cannot improve the rate in general. In particular, we show that, as the evidence accumulates when the policy proceeds, an asymptotically optimal policy effectively alternates between two specialists; as the evidence becomes decisive, it commits to a single one. We further provide guidance on how to construct an asymptotically optimal two-source policy.

We further show that the proposed policy achieves a proper balance between cost performance and computational complexity. We first show that there exists a family of problem instances where any exact optimal policy must query all of the sources at least once with high probability. In other

words, the computation cost is prohibitively high if one wants to achieve the minimum cost. We then show that using one source alone can incur a cost that is arbitrarily worse than the minimum cost, even in the high confidence regime, especially when different sources differ in their efficiency in gathering information. Lastly, we also show that there is a significant benefit in sequentially querying relative to a static design that chooses all sources' query counts before observing any responses. Taken together, these results suggest that adaptively querying two information sources is necessary to achieve near-optimal cost without incurring excessive computation costs.

Our work contributes to the emerging intersection of Operations Research (OR) and AI systems. The major contributions of this paper can be summarized as follows:

- From a modeling perspective, we introduce a sequential hypothesis testing framework that captures a class of fundamental problems arising naturally in modern AI systems but is not well understood in the existing literature. In contrast with classical sequential testing, our model explicitly incorporates monetary cost, stochastic waiting time, and stopping decisions within one unified objective. In contrast with most existing work on test-time compute in computer science, we are able to obtain clean and insightful theoretical results.
- From a theoretical perspective, we provide a sharp asymptotic characterization of optimal performance in the high-confidence regime. The lower bound shows that the error rate requirement can be summarized by two information budgets, one for each hypothesis, while the upper bound shows that a simple sign-based policy attains this lower bound up to the best possible additive remainder. Classical sequential design results establish asymptotic optimality when the objective is expected sample size or a closely related stopping criterion. Our results go further by allowing source-specific costs, stochastic waiting times, and nonlinear delay penalties, and by showing that these operational features can be incorporated without losing sharp analytical structure.
- From a practical perspective, we show that while adaptively querying two sources can achieve asymptotically optimal cost, either dropping one more source or statically choosing the query counts will significantly increase the cost. This is a strong and actionable design principle. A near-optimal AI system does not need to orchestrate a large collection of sources in a complicated way; it needs to identify one source that is especially effective for each hypothesis, respectively, and then switch between them as evidence evolves. This insight helps explain why simple cascading may work well in practice, and suggests that the complexity of effective AI decision systems need not grow with the number of the available sources.

1.1. Related Literature

Our work connects to and draws on several streams of research. We will discuss them in turn and clarify how our contribution differs from or extends the existing results.

Our work is closely related to the literature on sequential hypothesis testing and sequential information acquisition. Wald (1945) introduced the celebrated sequential probability ratio test (SPRT) and established its optimality for testing two simple hypotheses with a single source of observation. Chernoff (1959) further extended these findings to the sequential design of experiments, where the decision-maker adaptively chooses which experiment to run. Later work (Naghshvar and Javidi 2013) further developed active sequential hypothesis testing formulations that optimize information acquisition across multiple actions. Dayanik and Sezer (2012) study a continuous-time Bayesian binary testing problem with multiple simultaneously observed sources and characterize the optimal stopping region. Song and Fellouris (2017) and Tsopelakos and Fellouris (2023) study sequential multiple testing and anomaly detection across multiple independent data streams, where the objective is to identify which streams are anomalous under global error constraints and sampling limitations. There are also some works that incorporate operational frictions, such as switching costs (Vaidhiyan and Sundaresan 2015) and delays (Lambez and Cohen 2022). This literature provides the conceptual backbone of our analysis. Our model differs, however, in several important respects. We study a setting where the decision-maker chooses both which source to query and when to stop; each source has its own monetary cost, stochastic waiting time with a convex penalty on total waiting time, and asymmetric diagnostic power.

Our work also contributes to the rapidly growing computer-science literature on test-time compute for LLMs. Self-consistency (Wang et al. 2022) established repeated sampling and answer aggregation as a powerful test-time strategy. Adaptive self-consistency and early-stopping methods show that it is often inefficient to work with a fixed budget and that one can reduce the number of model calls by stopping once sufficient agreement emerges (Aggarwal et al. 2023, Li et al. 2024, Huang et al. 2026b). Other works study broader test-time scaling laws and compute-optimal inference strategies, showing that the value of additional calls depends on the complexity of the prompting technique and on how computation is split across generation, verification, and aggregation (Chen et al. 2024a, Snell et al. 2025, Huang et al. 2026a, Singhi et al. 2025). Recent works on verifier-based reasoning and reward-model approaches similarly use additional test-time compute to search over, rank, or refine candidate solutions (Zhang et al. 2025, Setlur et al. 2025). Agentic methods such as ReAct, Tree of Thoughts, CRITIC, and debate further illustrate that performance improvement often comes from repeated calls, searches, critiques, and external feedback rather than from a single forward pass (Yao et al. 2023b,a, Gou et al. 2024, Du et al. 2024). Our paper is complementary to these works. Rather than proposing another protocol for repeated sampling, search, or self-correction, we propose a model that captures the key trade-off of how to orchestrate heterogeneous information sources sequentially when error rate, monetary cost, and latency all matter.

More broadly, the machine learning literature on model-cascade and budgeted-prediction studies how to escalate examples through sequences or trees of predictors with increasing computational cost, typically using an offline-designed cascade or partition of the input space (Viola and Jones 2001, Xu et al. 2013). Closely related developments in LLM-as-judge have moved beyond single judges to study cascaded selective evaluation, panels of diverse evaluators, and debate-based assessment (Kim et al. 2024, Verga et al. 2024, Jung et al. 2025). In the LLM setting, FrugalGPT revisits this idea with cascades over heterogeneous APIs (Chen et al. 2024b). A related but distinct line studies LLM routing, where a router makes a one-shot assignment of each query to a cheaper or stronger model based on predicted difficulty or quality–cost trade-offs (Ong et al. 2025, Ding et al. 2024). Relative to these works, our model allows for sequential querying and re-routing as evidence accumulates; we also provide asymptotic optimality guarantees for the proposed policy.

Lastly, our work also lies at the intersection of OR and LLMs. This emerging literature uses OR tools such as queueing theory, online scheduling, robust optimization, and stochastic modeling to study how LLM systems should be run efficiently at inference time. Recent papers analyze how to batch and schedule many concurrent LLM requests under memory constraints, uncertain response lengths, and latency objectives (Ao et al. 2025, Jaillet et al. 2025, Li et al. 2025, Wang et al. 2025b, Chen et al. 2025, Bari et al. 2025). Collectively, these papers show that OR methods can provide new algorithmic insights and rigorous performance guarantees for important deployment problems in generative AI. The focus is mostly on how a shared system should allocate hardware and service capacity across many jobs. By contrast, we study a task-level problem within a single job: how to allocate sequential test-time compute across heterogeneous information sources until one instance can be certified with high confidence. Huang et al. (2026b) is more closely related to our paper. They study Bayesian stopping for repeated sampling from a single LLM in order to identify the most consistent answer under informative priors. We, instead, study a binary hypothesis testing problem with multiple heterogeneous sources, each with its own diagnostic power, monetary cost, and stochastic response time. Guo et al. (2026) study classification tasks routed through LLM cascades in a steady-state system with endogenous queueing delays. Their analysis optimizes routing while accounting for congestion across jobs, whereas ours optimizes successive queries within a job while accounting for uncertainty in both evidence and accumulated response time. Belloni et al. (2026) formulate contextual LLM cascading as an online Pandora’s box problem and develop a policy with sublinear regret that learns which models to query and which generated output to deploy using feedback observed only for the deployed output. While their framework selects among generated outputs and learns across requests, we study sequential evidence accumulation for binary classification, minimizing query and waiting costs subject to accuracy requirements and establishing conditions under which two models suffice for asymptotic optimality. These modeling differences lead to a different mathematical structure and a different set of insights.

1.2. Organization

The remainder of the paper is organized as follows. Section 2 introduces the model and its assumptions. Section 3 shows that exact optimal policies may require every available model. Section 4 develops the deterministic lower bound and the two-LLM policy, proves asymptotic optimality, and establishes the tight order of the gap to the benchmark. Sections 5 and 6 examine the value of a second model and the value of sequential querying, respectively. Section 7 reports the numerical experiments, and Section 8 concludes. Proofs and additional computational details appear in the appendix.

2. Model

A decision-maker must identify a fixed but unknown ground truth $\theta \in \{A, B\}$ by sequentially querying a collection of information sources. For concreteness, we refer to these sources as LLMs, although the model applies to any collection of noisy binary classifiers. Let $[M] := \{1, \dots, M\}$ denote the available LLMs, and let $D \in \{A, B\}$ denote the final decision. The decision-maker's prior is

$$P(\theta = A) = \xi_A \in (0, 1), \quad P(\theta = B) = \xi_B := 1 - \xi_A \in (0, 1).$$

To reach a satisfactory decision, the decision-maker adaptively queries the LLMs sequentially until the posterior error probability falls below a threshold $\alpha \in (0, 1/2)$. At round $t = 1, 2, \dots$, the decision-maker selects an LLM $j_t \in [M]$ using the observations available so far. The selected LLM returns an output $Y_t \in \{A, B\}$ after a random waiting time $T_{j_t, t}$. Queries are made one at a time, and an LLM may be queried repeatedly. Each LLM j has three characteristics:

1. A per-query monetary cost $c_j > 0$.
2. Asymmetric accuracies $\gamma_{j,A}, \gamma_{j,B} \in (1/2, 1)$: the probability of returning A under $\theta = A$ is $\gamma_{j,A}$, whereas the probability of returning B under $\theta = B$ is $\gamma_{j,B}$.
3. A nonnegative waiting time with known mean $\mu_j := \mathbb{E}[T_{j,t}] \in (0, \infty)$.

Allowing $\gamma_{j,A} \neq \gamma_{j,B}$ captures differences in a model's ability to distinguish the two hypotheses. For example, a model may reliably identify violations but reject compliant cases more often. The relevant trade-off is therefore not captured by a single average accuracy. We treat the accuracies and waiting-time distributions as known inputs for the class of tasks under consideration; they can be learnt from historical observation, so we do not explicitly model the process of learning this information.

2.1. Admissible Policies and Posterior Dynamics

Let $\mathcal{H}_t := \sigma(j_1, Y_1, \dots, j_t, Y_t)$ denote the history of model choices and outputs, and let $\mathcal{G}_t := \sigma(j_1, Y_1, T_{j_1,1}, \dots, j_t, Y_t, T_{j_t,t})$ also include realized waiting times. For randomized policies, we augment these histories with an independent random seed available to the decision-maker. Under our

assumption (see Assumption 1), waiting times provide no additional evidence about θ once the choices and outputs are observed, so the posterior is the same under $\mathcal{H}_t$ and $\mathcal{G}_t$. Nevertheless, realized delays may affect which LLM the decision-maker queries next.

We consider non-anticipative policies that stop at the first round when either hypothesis has a posterior probability of at least $1 - \alpha$, where $\alpha \in (0, 1/2)$. Given the stopping criteria, the total query time is

$$\tau := \inf\{t \geq 0 : P(\theta = A \mid Y_{1:t}, j_{1:t}) \geq 1 - \alpha \text{ or } P(\theta = B \mid Y_{1:t}, j_{1:t}) \geq 1 - \alpha\}, \quad (1)$$

which is also a stopping time. The corresponding decision rule is

$$D := \begin{cases} A, & \text{if } \mathbb{P}(\theta = A \mid Y_{1:\tau}, j_{1:\tau}) \geq 1 - \alpha, \\ B, & \text{if } \mathbb{P}(\theta = B \mid Y_{1:\tau}, j_{1:\tau}) \geq 1 - \alpha. \end{cases} \quad (2)$$

This rule ensures $\mathbb{P}(D \neq \theta \mid \mathcal{G}_\tau) \leq \alpha$, and hence the probability of making a wrong inference satisfies $\mathbb{P}(D \neq \theta) \leq \alpha$. If the prior already meets the confidence requirement, then $\tau = 0$. Our analysis concerns the nontrivial case $\alpha < \min\{\xi_A, \xi_B\}$, which holds eventually as $\alpha \rightarrow 0$ for any fixed prior.

We now formally define the class of admissible policies.

DEFINITION 1 (ADMISSIBLE POLICY). For a given $\alpha \in (0, 1/2)$, an admissible policy $\pi = \{\phi_t\}_{t \geq 1}$ selects $j_t = \phi_t(\mathcal{G}_{t-1}) \in [M]$ nonanticipatively and uses the stopping time and terminal decision in (1) and (2), respectively. Independent policy randomization is included in the history as well.

Let $\Pi(\alpha)$ denote this class. The stopping criterion is prescribed; the control decision is which LLM to query before the criterion is met. The realized stopping time remains endogenous because it depends on the chosen LLMs and their outputs.

In characterizing the posterior dynamic, we choose to work with the log-likelihood ratio (LLR) instead of the posterior probability directly. This is due to three reasons. First, the cumulative LLR accumulates additively across queries (see LLR), with each query to LLM j contributing a specific gain regardless of the current belief state. In contrast, the posterior updates nonlinearly. Second, as we show in Lemma 1, the posterior threshold stopping rule translates into a simple linear boundary crossing condition on the cumulative LLR, connecting our problem to classical random walk theory. Third, the expected increment of the log-likelihood ratio per query under each hypothesis is a fixed property of LLM j that does not depend on the cumulative LLR. This last point is crucial: it means the information contributed by each LLM can be characterized by two numbers, independent of how the process has evolved, which is what allows the optimal policy to be analyzed through the deterministic convex program developed in Section 4.1.

We start by defining the LLR for a single LLM. Define the prior log-odds in favor of A over B as

$$\delta := \log \frac{\xi_A}{\xi_B},$$

which serves as the initial offset for the log-likelihood ratio process that accumulates evidence in favor of A versus B . For any single LLM $j \in [M]$ and output $y \in \{A, B\}$, the LLR of $\theta = A$ versus $\theta = B$ can be written as

$$\ell_j(y) := \log \frac{\mathbb{P}(Y = y \mid \theta = A, j)}{\mathbb{P}(Y = y \mid \theta = B, j)} = \begin{cases} \log \frac{\gamma_{j,A}}{1 - \gamma_{j,B}}, & \text{if } y = A, \\ \log \frac{1 - \gamma_{j,A}}{\gamma_{j,B}}, & \text{if } y = B. \end{cases} \quad (3)$$

Since $\gamma_{j,A}, \gamma_{j,B} \in (1/2, 1)$, we have $\ell_j(A) > 0$ and $\ell_j(B) < 0$. Recall that j_t denotes the LLM chosen at round t , and Y_t is its output. Then, starting from $t = 1$, the *cumulative log-likelihood ratio (LLR)* up to t can be written as

$$L_t := \sum_{s=1}^t \ell_{j_s}(Y_s). \quad (\text{LLR})$$

For ease of exposition, let $\mathbb{P}_\theta(\cdot) := \mathbb{P}(\cdot \mid \theta)$ and $\mathbb{E}_\theta[\cdot] := \mathbb{E}[\cdot \mid \theta]$ for $\theta \in \{A, B\}$. Unsubscripted probabilities and expectations average over the prior, so

$$\mathbb{P}(\cdot) = \xi_A \mathbb{P}_A(\cdot) + \xi_B \mathbb{P}_B(\cdot), \quad \mathbb{E}[\cdot] = \xi_A \mathbb{E}_A[\cdot] + \xi_B \mathbb{E}_B[\cdot].$$

Using L_t , we can interpret stopping rule via a threshold structure. Formally, the following lemma translates the posterior confidence requirement into two stopping boundaries.

LEMMA 1. *Under Assumption 1, for every time $t \geq 1$, we have*

$$\mathbb{P}(\theta = A \mid Y_{1:t}, j_{1:t}) = \frac{e^{\delta + L_t}}{1 + e^{\delta + L_t}}, \quad \mathbb{P}(\theta = B \mid Y_{1:t}, j_{1:t}) = \frac{1}{1 + e^{\delta + L_t}}. \quad (4)$$

Moreover, the stopping time and the terminal decision can both be represented equivalently using a threshold structure

$$\tau := \inf\{t \geq 0 : L_t \geq A_\alpha \text{ or } L_t \leq -B_\alpha\}, \quad D := \begin{cases} A, & \text{if } L_\tau \geq A_\alpha, \\ B, & \text{if } L_\tau \leq -B_\alpha, \end{cases} \quad (\text{Threshold Decision Rule})$$

where $A_\alpha := \log \frac{1-\alpha}{\alpha} - \delta$ and $B_\alpha := \log \frac{1-\alpha}{\alpha} + \delta$.

The structure of the threshold policy has an intuitive explanation: querying will continue until sufficient evidence has been accumulated towards a hypothesis, measured by cumulative information gain crossing a certain threshold. In particular, L_t measures the evidence supplied by the queries, while $\delta + L_t$ incorporates the prior. A prior favoring A reduces the additional evidence needed to declare A and increases the evidence needed to declare B .

To quantify an LLM's contribution to the query process, we associate each LLM $j \in [M]$ with a pair of *expected information gains* under different ground truth θ :

$$\begin{aligned} I_{j,A} &:= \mathbb{E}_A[\ell_j(Y)] = \sum_{y \in \{A,B\}} \mathbb{P}(Y = y \mid \theta = A, j) \cdot \ell_j(y), \\ I_{j,B} &:= -\mathbb{E}_B[\ell_j(Y)] = - \sum_{y \in \{A,B\}} \mathbb{P}(Y = y \mid \theta = B, j) \cdot \ell_j(y). \end{aligned}$$

For simplicity, we will call these quantities *information rates*, measured per query rather than per unit of time. Observe that both $I_{j,A}$ and $I_{j,B}$ are strictly positive because $\gamma_{j,A}, \gamma_{j,B} \in (1/2, 1)$. Under A , the process L_t has a positive drift; under B , it is $-L_t$ that has a positive drift. Each rate depends on both accuracy parameters because an output is informative only to the extent that it distinguishes the hypotheses.

The definition of information rate is consistent with the KL divergence: If $P_{j,A}$ and $P_{j,B}$ are the output distributions of LLM j under the two hypotheses, then

$$I_{j,A} = \text{KL}(P_{j,A} \parallel P_{j,B}) > 0, \quad I_{j,B} = \text{KL}(P_{j,B} \parallel P_{j,A}) > 0,$$

where $\text{KL}(P \parallel Q)$ denotes the Kullback–Leibler divergence.

We further define *cost-efficiency* and *time-efficiency* for each $\theta \in \{A, B\}$ as

$$\kappa_{j,\theta} := \frac{c_j}{I_{j,\theta}}, \quad \eta_{j,\theta} := \frac{\mu_j}{I_{j,\theta}}, \text{ respectively}$$

where c_j is the per-query monetary cost, and μ_j is the mean waiting time. These quantities are central to the design of our proposed policy. Smaller values indicate greater efficiency. In particular, the fastest model per query need not be the fastest at resolving a hypothesis: a slower but more informative model can have a smaller $\eta_{j,\theta}$. These ratios, rather than accuracy or latency alone, guide the policy developed in Section 4.

2.2. The Objective Function: Balancing Query Cost and Latency

The objective of the decision maker is to achieve the accuracy threshold $1 - \alpha$ at a low cost and high speed. For a policy π with stopping time τ , the total query cost and total waiting time are defined as

$$C_\pi := \sum_{t=1}^{\tau} c_{j_t}, \quad W_\pi := \sum_{t=1}^{\tau} T_{j_t,t}.$$

To characterize the cost of waiting, we let $g: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a convex and non-decreasing function with $g(0) = 0$. The objective is to minimize the total weighted cost of a policy $\pi \in \Pi(\alpha)$ at error tolerance α , defined as

$$\mathcal{R}(\pi; \alpha) := \mathbb{E}[C_\pi + g(W_\pi)] = \mathbb{E} \left[\sum_{t=1}^{\tau} c_{j_t} + g \left(\sum_{t=1}^{\tau} T_{j_t,t} \right) \right] \quad (\text{Objective})$$

Minimizing $\mathcal{R}(\pi; \alpha)$ strikes a trade-off between information gain, monetary cost, and total waiting time. To characterize such a trade-off, it is essential to carefully analyze the behavior of L_t and the stopping time τ and further establish their connections to the objective $\mathcal{R}(\pi, \alpha)$.

Our model is analogous to Wald’s sequential probability ratio test (SPRT). However, there are at least two key features that distinguish our setting from classical sequential testing models. First,

in standard formulations of sequential hypothesis testing, there is a single information channel, and the only control decision concerns *when* to stop. In contrast, our framework involves multiple heterogeneous and asymmetric information sources. The decision-maker must decide not only *when* to stop but also *which* LLM to query. While the statistical characteristics of each LLM are known, their costs, waiting time distributions, and diagnostic efficiencies differ across hypotheses. The challenge, therefore, lies in optimally orchestrating multiple LLMs whose contributions vary across states. This results in a high-dimensional stochastic control problem in which both the stopping time and the flow of information are endogenous and depend on the entire query history. Second, the structure of the objective function introduces an additional layer of complexity. Query costs scale linearly with the number of queries, but the random total waiting time is penalized through a convex function. This nonlinearity prevents a direct application of standard optimal stopping or dynamic programming arguments. In particular, a near-optimal policy must carefully control the distribution of the stopping time and the accumulated waiting time, rather than only their expectations.

REMARK 1. It suffices to consider policies with $\mathbb{E}[\tau] < \infty$. Indeed, if $\mathbb{E}[\tau] = \infty$, then $c_j > 0$ implies,

$$\mathbb{E}[C_\pi] \geq (\xi_A \mathbb{E}_A[\tau] + \xi_B \mathbb{E}_B[\tau]) \cdot \min_{j \in [M]} c_j = \infty,$$

A policy with an infinite expected number of queries, therefore, has infinite cost. Under the assumptions below, repeatedly querying any fixed LLM yields a finite-cost threshold policy. Thus, infinite-cost policies cannot be optimal.

2.3. Assumptions and Discussion

We impose the following assumptions on the information and waiting-time processes.

ASSUMPTION 1. *We assume:*

(i) **OUTPUT INDEPENDENCE.** *Given the ground truth and the selected LLM, the next output has a history-independent distribution:*

$$\mathbb{P}(Y_t = y \mid \theta, j_t = j, \mathcal{G}_{t-1}) = \mathbb{P}(Y = y \mid \theta, j), \quad y \in \{A, B\}, \quad t \geq 1.$$

(ii) **STATIONARY WAITING TIMES.** *For each LLM $j \in [M]$, the waiting times $(T_{j,t})_{t \geq 1}$ are i.i.d. with mean $\mu_j > 0$ and finite variance $\sigma_j^2 \geq 0$.*

(iii) **SEPARATION OF INFORMATION AND LATENCY.** *Waiting times are generated independently of the ground truth, output-generating randomness, and policy randomization. In particular, for a measurable set $E \subseteq \mathbb{R}_+$,*

$$\mathbb{P}(T_{j_t,t} \in E \mid \theta, \mathcal{G}_{t-1}, j_t, Y_t) = \mathbb{P}(T_{j_t,t} \in E \mid j_t).$$

(iv) **CROSS-LLM INDEPENDENCE.** *The waiting-time sequences associated with different LLMs are mutually independent.*

These assumptions capture the idea that each query behaves as a fresh experiment whose output realization depends only on the selected LLM and the ground truth θ . Part (i) ensures that evidence accumulates additively over time, which yields a tractable likelihood-ratio representation. Parts (ii)–(iv) separate informational uncertainty from system-level latency. In particular, inference delays are driven by computational and network factors, not by which hypothesis is true or which label is produced. This separation allows us to analyze information acquisition and delay costs in parallel and underpins the martingale and concentration arguments developed later.

We next formalize the tail property imposed on the waiting-time distributions.

ASSUMPTION 2 (Sub-Gaussian Waiting Times). *There exists a constant $\psi > 0$ such that, for every $j \in [M]$, $t \geq 1$, and $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda(T_{j,t} - \mu_j)}] \leq \exp\left(\frac{\psi^2 \lambda^2}{2}\right).$$

Assumption 2 controls the tails of individual response times. The concentration of the total delay also depends on the random query count and the routing policy; it does not follow from the individual tail bound alone.

Lastly, to enable our asymptotic analysis in Section 4, we introduce a mild regularity assumption on the waiting-time cost function $g(\cdot)$.

ASSUMPTION 3 (Power Waiting Cost). *The waiting penalty is $g(x) = cx^\rho$, where $c \geq 0$ and $\rho \geq 1$.*

Assumption 3 captures one representative class of convex non-decreasing functions that characterize the waiting time penalty. It can be further generalized such that $g(x)$ is a non-decreasing, regularly varying function with order $\rho \geq 1$ (see Bingham et al. 1989). Similar results can be derived following the proof framework in this paper.

The implications of Assumption 2 and Assumption 3 are that, under any reasonable policy π , the total waiting time W_π should scale on the same order as the number of queries τ which, as we will later show, grows only logarithmically in $1/\alpha$. If $g(\cdot)$ were allowed to grow exponentially fast, then even moderate random fluctuations in W_π would make the waiting time penalty dominate the objective function, rendering the query cost meaningless. Assumption 3 rules out such explosive growth, while sub-Gaussian tails prevent large deviations of W_π . Together, they ensure that the waiting-time penalty remains commensurate with query costs and information accumulation, so that $g(W_\pi)$ cannot overwhelm C_π in the asymptotic regime $\alpha \rightarrow 0$. This matches how LLMs are used

in practice: users and system builders care about both the per-call price (e.g., pay-per-token or per-request fees) and the time-to-response, and adoption decisions are typically made by trading these two considerations off against each other. Moreover, in most production settings, response times are engineered to stay within a reasonable, finite range (via timeouts, load balancing, batching, and autoscaling), so while latency is stochastic, it typically exhibits limited tail risk around a stable mean, which is consistent with sub-Gaussian concentration.

REMARK 2 (NO MONOTONICITY ASSUMPTIONS). In this paper, we do *not* assume any monotonic relationship between c_j , $(\gamma_{j,A}, \gamma_{j,B})$, and μ_j across different LLMs. Thus, an LLM may be relatively cheap but inaccurate, accurate but slow, or exhibit any other combination of characteristics. Our results do not rely on any ordering of the LLMs.

In Table 1, we provide a summary of the notations.

Table 1 Summary of notation

Symbol	Definition
M	Number of LLMs
θ	Ground-truth hypothesis
ξ_θ	Prior probability of hypothesis $\theta \in \{A, B\}$
δ	Prior log-odds in favor of A
$\gamma_{j,\theta}$	Accuracy of LLM j under hypothesis θ
$T_{j,t}$	Waiting time of LLM j at round t
μ_j	Mean waiting time of LLM j
c_j	Per-query cost of LLM j
$\ell_j(y)$	Single-query log-likelihood ratio
L_t	Cumulative log-likelihood ratio through round t
$I_{j,\theta}$	Expected evidence per query in favor of the true hypothesis
$A_\alpha, -B_\alpha$	Upper and lower stopping boundaries for L_t
$\kappa_{j,\theta} = c_j / I_{j,\theta}$	Monetary cost per unit of information
$\eta_{j,\theta} = \mu_j / I_{j,\theta}$	Mean waiting time per unit of information
τ, C_π, W_π	Number of queries, total query cost, and total waiting time

3. Exact Optimal Policies May Query All the LLMs

The availability of many LLMs creates a natural design question: can the decision-maker select a small subset before querying begins without sacrificing optimality? At a fixed confidence requirement, a model’s usefulness depends not only on its average information contribution, but also on whether its possible outputs provide enough evidence to stop from the current belief. We show that these state-dependent advantages can make every available LLM necessary for exact optimality.

Let $N_j(t)$ denote the number of queries sent to LLM j through round t and define $\bar{\alpha} := (1 + e^4)^{-1}$. The following theorem constructs instances in which every exact optimal policy queries all available models with high probability under hypothesis A .

THEOREM 1 (Worst-case use of all available LLMs). *For each $\alpha \in (0, \bar{\alpha}]$, define $H_\alpha := \log \frac{1-\alpha}{\alpha}$ and $M_\alpha := \lfloor \sqrt{H_\alpha} \rfloor$. Then there exists a problem instance with M_α LLMs such that every exact optimal policy $\pi^* \in \arg \min_{\pi \in \Pi(\alpha)} \mathcal{R}(\pi; \alpha)$ satisfies*

$$\mathbb{P}_A(N_j(\tau) \geq 1 \text{ for all } j \in [M_\alpha]) \geq 1 - M_\alpha \frac{\alpha}{1-\alpha}.$$

Moreover, $M_\alpha \rightarrow \infty$ and $M_\alpha \frac{\alpha}{1-\alpha} \rightarrow 0$ as $\alpha \rightarrow 0$.

Theorem 1 rules out a constant bound, independent of both the instance and the error tolerance, on the number of LLMs used by an exact optimal policy. Moreover, the need for all models is not confined to rare histories: conditional on A , the probability of using every model approaches one along the constructed family.

We explain the construction to show why different models become useful at different beliefs. Take a symmetric prior, so $\delta = 0$ and $A_\alpha = B_\alpha = H_\alpha$. Set $c_j = 1$ and $T_{j,t} \equiv 1$ for every model, and let $g(w) = w$. Then $\mathcal{R}(\pi; \alpha) = 2\mathbb{E}[\tau]$, so minimizing total cost is equivalent to minimizing the expected number of queries. Equal prices and response times remove cost and speed differences across models. The choice of model therefore turns entirely on how its outputs move the belief toward a stopping boundary. By Lemma 1, querying continues while the LLR lies in $(-H_\alpha, H_\alpha)$.

Let $\lambda_\alpha := H_\alpha/M_\alpha$ and define the ladder of LLR states

$$x_i := i\lambda_\alpha, \quad i = 0, 1, \dots, M_\alpha,$$

where $x_0 = 0$ and $x_{M_\alpha} = H_\alpha$. The accuracy parameters are chosen so that, at each continuation state x_i on this ladder, an A output from any LLM moves the LLR to x_{i+1} . A B output from LLM j instead moves it to $x_i - (M_\alpha + j - 1)\lambda_\alpha$. Appendix B verifies that these transitions arise from accuracies in $(1/2, 1)$. Thus, all models supply the same additional evidence after an A output, but a larger-index model supplies stronger contrary evidence after a B output. This stronger B output is also less likely, given the current belief.

At x_i , LLM $i+1$ is the smallest-index model whose B output reaches the lower stopping boundary. This is the central feature of the construction: as the accumulated evidence moves farther toward A , a larger downward move is needed for a B output to end the test. LLM $i+1$ supplies exactly that move.

A smaller-index model $j \leq i$ produces B with a slightly higher probability, but that response leaves the LLR inside the continuation region, requiring further queries. The construction ensures that the saving from making an A output less likely is too small to compensate for this additional continuation after a B output. A larger-index model $j > i+1$, when available, also stops after a B output, but is less likely to produce it. Its A output leads to the same continuation state x_{i+1} ,

so it increases the probability of requiring more queries without improving the stopping outcome after B . Overshooting the lower boundary is not itself penalized; it simply provides no additional benefit once the confidence requirement is met.

The Bellman comparisons in Appendix B establish that LLM $i + 1$ is uniquely optimal at x_i . At the last continuation state $x_{M_\alpha - 1}$, LLM M_α stops after either output, whereas every smaller-index model still requires additional queries after a B output. The choice $\bar{\alpha} = (1 + e^4)^{-1}$ ensures that the comparisons hold throughout the stated range of error tolerances.

Figure 2 illustrates the construction with $\alpha = (1 + e^{25})^{-1}$, so $H_\alpha = 25$, $M_\alpha = 5$, and $\lambda_\alpha = 5$. At $x_2 = 10$, all five models move to $x_3 = 15$ after an A output, so querying continues. The figure compares three representative models through their B outputs: LLM 2 moves to -20 , still inside the continuation region; LLM 3 reaches the lower boundary -25 ; and LLM 5 moves beyond it to -35 . Thus, LLM 3 is the first model whose B output suffices to stop. The preceding comparisons explain why it is uniquely optimal among all five models, not merely the three shown.

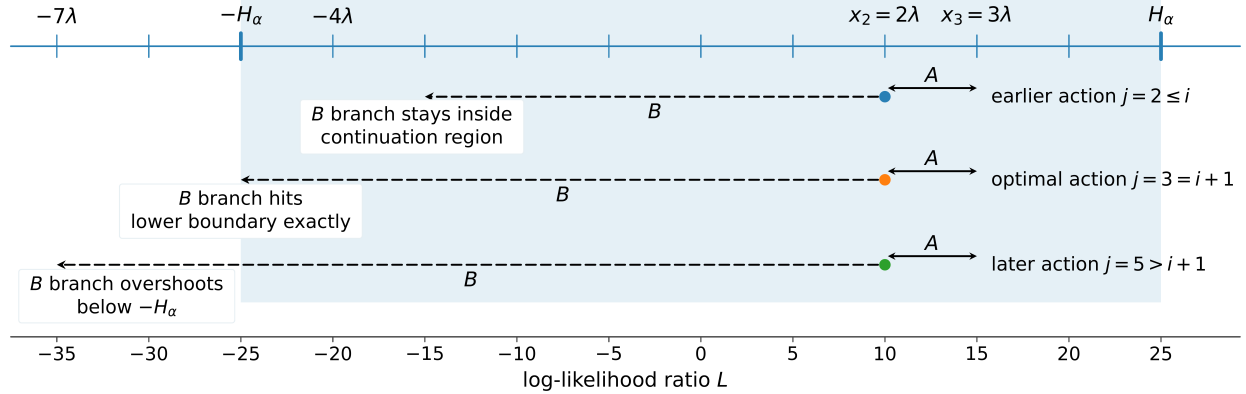


Figure 2 An illustrative example within the family of instances as defined in Theorem 1

To obtain the probability bound, consider the path on which the first M_α outputs are all A . Because each ladder state has a unique optimal model, every exact optimal policy follows

$$0 = x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_{M_\alpha} = A_\alpha.$$

In other words, every exact optimal policy must query LLMs $1, 2, \dots, M_\alpha$ in that order before stopping. Conditional on $\theta = A$, this path has a probability $\prod_{j=1}^{M_\alpha} \gamma_{j,A}$. Our construction further ensures

$$1 - \gamma_{j,A} \leq e^{-H_\alpha} = \frac{\alpha}{1 - \alpha},$$

and therefore

$$\prod_{j=1}^{M_\alpha} \gamma_{j,A} \geq 1 - \sum_{j=1}^{M_\alpha} (1 - \gamma_{j,A}) \geq 1 - M_\alpha \frac{\alpha}{1 - \alpha}.$$

In summary, Theorem 1 shows that, in the non-asymptotic setting, the exact optimal policy may use all available LLMs in the worst case. Computing such an exact optimal policy would, therefore, require solving a high-dimensional stochastic dynamic programming problem, which is intractable in practice. This finding motivates the design goal of the rest of the paper. Since the exact optimal policy with finite- α can be structurally complex, could a simpler adaptive policy have near-optimal performance in a practically relevant regime? Fortunately, as we show later, a policy that only uses two LLMs is asymptotically optimal in the high confidence regime.

4. The Two-LLM Policy

We now develop a simple policy by first identifying the information that any admissible policy must acquire. In particular, we first translate the posterior confidence requirement into deterministic information budgets and construct a deterministic lower bound program for the optimal objective. Motivated by the structure of the optimal solution to this lower bound program, we propose a policy that identifies an efficient LLM for each hypothesis and then routes queries to one of the two LLMs according to the current posterior. We show that, although the proposed policy has to allocate queries without knowing the truth, as $\alpha \rightarrow 0$, its expected cost approaches the optimal cost of the lower bound program in which queries are allocated based on the ground truth and, therefore, is asymptotically optimal among all admissible policies. We further show that the gap between the cost of our policy and the lower bound program is tight.

4.1. A Deterministic Lower Bound

We start with the lower bound. The LLR representation in Section 2.1 makes it possible to separate the amount of evidence required from the way that evidence is acquired. Every query contributes a model-specific expected information gain under each hypothesis. Consequently, although a policy can follow many different histories, its expected evidence at stopping can be summarized by the expected number of times it uses each LLM.

Fix an admissible policy π and LLM j . Define the number of queries sent to that LLM by time t and stopping time τ as

$$N_j^\pi(t) := \sum_{s=1}^t \mathbf{1}\{j_s^\pi = j\}. \quad (5)$$

We further define $n_{j,\theta}^\pi := \mathbb{E}_\theta[N_j^\pi(\tau)]$ as the expected number of queries of LLM j under $\theta \in \{A, B\}$. The two vectors $\mathbf{n}_A^\pi$ and $\mathbf{n}_B^\pi$ describe the allocations generated by the same policy under different ground truths; they are not choices that the policy can make after observing θ .

Under Assumption 1, conditioning on the ground truth θ , at stopping time τ , policies with finite expected query counts satisfy

$$\begin{aligned}\mathbb{E}_A[L_\tau^\pi] &= \sum_{j=1}^M I_{j,A} \mathbb{E}_A[N_j^\pi(\tau)] = \sum_{j=1}^M I_{j,A} n_{j,A}^\pi, \\ \mathbb{E}_B[-L_\tau^\pi] &= \sum_{j=1}^M I_{j,B} \mathbb{E}_B[N_j^\pi(\tau)] = \sum_{j=1}^M I_{j,B} n_{j,B}^\pi.\end{aligned}\tag{6}$$

The negative sign under B aligns evidence with the true hypothesis, so both sums are positive. These identities turn a requirement on the stopped LLR into a requirement on expected query allocations.

The next proposition formalizes this intuition and shows that the expected information gathered from the LLMs must be at least over a certain threshold.

PROPOSITION 1. *There exists a constant $C_0 > 0$, independent of α and π , such that, for all sufficiently small α and every admissible policy satisfying $\mathbb{E}_A[\tau], \mathbb{E}_B[\tau] < \infty$,*

$$\begin{aligned}\mathbb{E}_A[L_\tau^\pi] &= \sum_{j=1}^M I_{j,A} n_{j,A}^\pi \geq S_A(\alpha) := A_\alpha - C_0, \\ \mathbb{E}_B[-L_\tau^\pi] &= \sum_{j=1}^M I_{j,B} n_{j,B}^\pi \geq S_B(\alpha) := B_\alpha - C_0.\end{aligned}\tag{7}$$

Note that as $\alpha \rightarrow 0$, both A_α and B_α are in the order of $\log(1/\alpha)$. As a result, Proposition 1 says that each hypothesis imposes an information budget of order $\log(1/\alpha)$ over any admissible policy. The budgets are slightly below the stopping thresholds (by a constant C_0) because a policy can occasionally stop at the wrong boundary. Under A , for example, those paths contribute negative rather than positive evidence to $\mathbb{E}_A[L_\tau]$. Their probability is $O(\alpha)$ and the LLR overshoot is bounded, so their effect can be absorbed into a constant slack. Thus, $S_\theta(\alpha) = \log(1/\alpha) + O(1)$ for either hypothesis. The prior changes the constant offset in the required evidence, but not its leading logarithmic order. In the remainder of the lower-bound analysis, we assume that α is small enough such that both budgets are positive.

In the following proposition, we further rewrite the objective function in terms of $n_{j,A}$ and $n_{j,B}$.

PROPOSITION 2. *For any policy π with $\mathbb{E}_\theta[\tau] < \infty$ and $\theta \in \{A, B\}$, we have*

$$\mathbb{E}_\theta[C_\pi] = \sum_{j=1}^M c_j n_{j,\theta}^\pi, \quad \mathbb{E}_\theta[W_\pi] = \sum_{j=1}^M \mu_j n_{j,\theta}^\pi\tag{8}$$

If we further use the expected values to replace the random variable in the objective function, we obtain the following deterministic function

$$F(\mathbf{n}_A^\pi, \mathbf{n}_B^\pi) := \sum_{\theta \in \{A, B\}} \xi_\theta \left[\sum_{j=1}^M c_j n_{j,\theta}^\pi + g \left(\sum_{j=1}^M \mu_j n_{j,\theta}^\pi \right) \right].\tag{9}$$

Combining the above results, we propose the following convex problem:

$$\begin{aligned} \Psi_\alpha &:= \min_{\mathbf{n}_A, \mathbf{n}_B} F(\mathbf{n}_A, \mathbf{n}_B) \\ \text{s.t. } &\sum_{j=1}^M I_{j,\theta} n_{j,\theta} = S_\theta(\alpha), \quad \forall \theta \in \{A, B\} \\ &n_{j,A}, n_{j,B} \geq 0 \quad \forall j \in [M]. \end{aligned} \tag{LB}$$

In the following proposition, we establish that Ψ_α is a tractable lower bound of the optimal objective value.

PROPOSITION 3. *For α such that $S_A(\alpha), S_B(\alpha) > 0$, every admissible policy satisfies $\mathcal{R}(\pi; \alpha) \geq \Psi_\alpha$. Moreover, $F(\mathbf{n}_A, \mathbf{n}_B)$ is convex in $(\mathbf{n}_A, \mathbf{n}_B)$.*

Program (LB) has a natural interpretation from the perspective of acquiring information. A query to model j supplies $I_{j,\theta}$ units of expected evidence under θ , at monetary cost c_j and mean delay μ_j . The program determines the least expensive combination that supplies the required budget under each hypothesis. Two further simplifications make this benchmark optimistic. Only expected information is required, rather than an actual sample path that reaches a stopping boundary, and the waiting penalty is evaluated at each conditional mean, rather than at realized delay. Every admissible policy induces allocations feasible for the corresponding inequality-constrained relaxation, and its true cost is no smaller than their deterministic cost. These observations explain why the program is a lower bound. They also identify that an asymptotically optimal policy needs to learn which allocation to follow while keeping the costs of excessive evidence and noisy delay on the lower order. Note that we change the inequality constraints in (7) to equality constraints. This is intuitive since the weighted cost of any non-binding query allocation can be further reduced.

To study the structure of the allocation, it is useful to work with fractions of information rather than fractions of queries. Define

$$w_{j,\theta} = \frac{I_{j,\theta} \cdot n_{j,\theta}}{S_\theta(\alpha)}.$$

The fractions sum to one under each hypothesis. Allocating a fraction $w_{j,\theta}$ to a more informative model requires fewer queries, so this normalization allows models to be compared on a common evidence scale. Using the cost-efficiency and time-efficiency ratios defined in Section 2, let

$$\kappa_\theta(\mathbf{w}_\theta) := \sum_{j=1}^M \kappa_{j,\theta} w_{j,\theta}, \quad \eta_\theta(\mathbf{w}_\theta) := \sum_{j=1}^M \eta_{j,\theta} w_{j,\theta}.$$

These are the average monetary cost and mean waiting time per unit of information. Using these definitions, we can rewrite the objective function of (LB) in terms of $\mathbf{w}_\theta := (w_{1,\theta}, w_{2,\theta}, \dots, w_{M,\theta})$, as follows:

$$G_\alpha(\mathbf{w}_A, \mathbf{w}_B) := \sum_{\theta \in \{A, B\}} \xi_\theta [S_\theta(\alpha) \kappa_\theta(\mathbf{w}_\theta) + g(S_\theta(\alpha) \eta_\theta(\mathbf{w}_\theta))]. \quad (10)$$

Thus, (LB) can be equivalently written as the following compact formulation

$$\Psi_\alpha = \min_{(\mathbf{w}_A, \mathbf{w}_B) \in \Delta_M \times \Delta_M} G_\alpha(\mathbf{w}_A, \mathbf{w}_B) \quad (\text{ALO})$$

where $\Delta_M := \left\{ \mathbf{w} \in \mathbb{R}_+^M : \sum_{j=1}^M w_j = 1 \right\}$ is the probability simplex in $\mathbb{R}^M$. Now, leveraging the Karush-Kuhn-Tucker (KKT) conditions, we can establish the property of (10) when α is relatively small.

PROPOSITION 4. *Let $c > 0$. Under Assumptions 1- 3, there exists a finite constant $\hat{\alpha} > 0$, such that for all $\alpha < \hat{\alpha}$, we have*

$$\min_{(\mathbf{w}_A, \mathbf{w}_B) \in \Delta_M \times \Delta_M} G_\alpha(\mathbf{w}_A, \mathbf{w}_B) = \min_{(j_A, j_B) \in [M] \times [M]} G_\alpha(\mathbf{e}_{j_A}, \mathbf{e}_{j_B}),$$

where $\mathbf{e}_j$ is the j -th unit vector. Furthermore, the optimal entries satisfy

$$j_\theta^* \in \arg \min_{j \in [M]} \{S_\theta(\alpha) \kappa_{j,\theta} + g(S_\theta(\alpha) \eta_{j,\theta})\}, \theta \in \{A, B\}.$$

Proposition 4 establishes a useful simplification of the deterministic lower-bound problem: for sufficiently small α , there exists an optimal allocation that assigns each hypothesis-specific information budget to a single LLM. The selected LLM balances the monetary cost and waiting penalty of supplying that budget. Specifically, if LLM j supplies all $S_\theta(\alpha)$ units of information under hypothesis θ , the relaxed allocation uses $S_\theta(\alpha)/I_{j,\theta}$ queries, incurring monetary cost $S_\theta(\alpha) \kappa_{j,\theta}$ and mean waiting time $S_\theta(\alpha) \eta_{j,\theta}$. The specialist therefore minimizes the combined cost of acquiring the required evidence, rather than being selected by its per-query price, accuracy, or response time alone.

Because both efficiency ratios depend on the hypothesis, the LLM offering the best cost–delay trade-off under A need not be the one offering the best trade-off under B . This gives rise to the two specialists, which may coincide if the same LLM minimizes both criteria. The remaining challenge is to recover the benefit of these hypothesis-specific allocations through a policy that does not know the ground truth.

4.2. Sign-based Two-LLM Policy

Based on the deterministic lower bound in Section 4.1, we design an asymptotically optimal admissible policy as α goes to zero. A classical idea in designing near-optimal control policy when a tractable benchmark is known is to mimic the behavior of the optimal solution to that benchmark. In our setting, the obstacle is informational rather than computational: In (LB), we allocate the queries based on the underlying truth $\theta \in \{A, B\}$; in contrast, an admissible policy is nonanticipating and does not know θ . We use the posterior to bridge this gap. When the evidence favors A , the policy queries the A -specialist; when it favors B , it queries the B -specialist.

Let j_A^* and j_B^* denote the A -specialist and the B -specialist defined by Proposition 4, respectively. In the following definition, we describe a sign-based two-LLM sequential (2LLM) policy.

DEFINITION 2 (SIGN-BASED TWO-LLM SEQUENTIAL POLICY). Fix an error tolerance α and the optimal pair of LLMs (j_A^*, j_B^*) in Proposition 4.

- *Initialization:* Set $L_0 = 0$.
- *LLM selection and update:* Before stopping, at period $t \geq 1$, choose

$$j_t = \begin{cases} j_A^*, & \text{if } \delta + L_{t-1} \geq 0, \\ j_B^*, & \text{if } \delta + L_{t-1} < 0, \end{cases}$$

observe Y_t from LLM j_t , and update $L_t = L_{t-1} + \ell_{j_t}(Y_t)$.

- *Stopping rule and decision:* Stop at the time

$$\tau := \inf\{t \geq 0 : L_t \geq A_\alpha \text{ or } L_t \leq -B_\alpha\},$$

and set

$$D = \begin{cases} A, & \text{if } L_\tau \geq A_\alpha, \\ B, & \text{if } L_\tau \leq -B_\alpha. \end{cases}$$

According to Definition 2, the choice of LLM at each step depends on the sign of the posterior log-odds $L_t + \delta$ (equivalent to a posterior probability $1/2$), so queries are directed toward the specialist aligned with the current direction of evidence. Policy 2LLM has a simple interpretation. When the accumulated evidence favors A , the policy queries the LLM that is most efficient at gathering information in favor of A ; otherwise, it queries the LLM that is most efficient for certifying B . When the posterior is close to $1/2$, the LLR is close to $-\delta$, and the policy may switch between the two specialists as the sign changes. As the evidence becomes more decisive, the policy increasingly queries the specialist associated with the favored hypothesis while retaining the same sign-based rule throughout. The policy uses mean waiting times to select the specialists but does not use realized waiting times in online routing.

In the following, we refer to a 2LLM policy that only queries j_A^* and j_B^* as $\pi(j_A^*, j_B^*)$ whenever we want to make the choice of LLMs explicit.

4.3. Asymptotic Optimality of Policy 2LLM

We next establish that this simple routing rule attains the deterministic lower bound up to a lower-order term. This is stronger than showing that the policy performs well within a restricted two-model class: the comparison is with every admissible policy, including those that use all available models and react to realized delays.

THEOREM 2 (Asymptotic optimality). *Let $c > 0$. As $\alpha \rightarrow 0$, we have $\Psi_\alpha = \Theta((\log(1/\alpha))^\rho)$. Moreover, a pair $(j_A^*(\alpha), j_B^*(\alpha))$ selected as in Proposition 4 yields a policy $\pi := \pi(j_A^*(\alpha), j_B^*(\alpha))$ satisfying, as $\alpha \rightarrow 0$,*

$$\mathcal{R}(\pi; \alpha) = \Psi_\alpha + O\left((\log(1/\alpha))^{\rho-1}\right). \quad (11)$$

Let $\mathcal{R}^*(\alpha) := \inf_{\pi \in \Pi(\alpha)} \mathcal{R}(\pi; \alpha)$. Since $\Psi_\alpha \leq \mathcal{R}^*(\alpha) \leq \mathcal{R}(\pi; \alpha)$, Theorem 2 implies

$$\begin{aligned} 0 \leq \mathcal{R}(\pi; \alpha) - \mathcal{R}^*(\alpha) &= O\left((\log(1/\alpha))^{\rho-1}\right), \\ \frac{\mathcal{R}(\pi; \alpha)}{\mathcal{R}^*(\alpha)} &= 1 + O\left(\frac{1}{\log(1/\alpha)}\right). \end{aligned}$$

With linear waiting cost, the additive gap is bounded. With a superlinear penalty, the additive bound can grow as confidence increases, but it remains small relative to the optimal total cost.

Theorem 2 establishes that Policy 2LLM is asymptotically optimal as the target error level $\alpha \rightarrow 0$. This conclusion has both theoretical and operational implications. Theoretically, it reveals a simple structure in the high confidence regime. Although the decision-maker may have access to many heterogeneous LLMs, near-optimal performance can be achieved by using at most two specialists. The complexity of sequential hypothesis experimentation, therefore, needs *not* to grow with the number of available LLMs. Moreover, we note that Policy 2LLM does not use the realized waiting time but achieves asymptotic optimality among all latency-adaptive policies.

Operationally, the theorem provides direct guidance for policy design. To implement a near optimal strategy, the decision-maker only needs to identify one strong model for certifying A and another for certifying B , and then switch between them as the posterior evolves. Adding more LLMs expands the set of candidate specialists and may potentially reduce the cost if a more efficient specialist is introduced, but it does not require more complex orchestration. As a result, the policy remains simple and scalable under stringent reliability requirements.

Our proof follows a matching-bounds approach centered around the deterministic universal lower bound Ψ_α discussed in Section 4.1. We show that, as $\alpha \rightarrow 0$, the expected query cost $\mathbb{E}[C_\pi]$ and the expected waiting time penalty $\mathbb{E}[g(W_\pi)]$ induced by $\pi = \pi(j_A^*, j_B^*)$ match their counterpart in the lower bound Ψ_α up to lower-order terms, respectively. To do this, we decompose the cumulative

log likelihood ratio (LLR) into a drift component that matches the counterpart in the lower bound and a martingale component that captures the stochasticity. Under $\theta \in \{A, B\}$, let

$$M_t^{(A)} = L_t - \sum_{s=1}^t I_{j_s, A} \text{ and } M_t^{(B)} = -L_t - \sum_{s=1}^t I_{j_s, B}. \quad (12)$$

We show that (see Lemma C.8) $M_t^{(\theta)}$ is a martingale under $\mathbb{P}_\theta$ with bounded increments. This decomposition isolates the predictable evidence accumulation term, which directly relates to the deterministic benchmark, from a martingale fluctuation term that captures overshoot and deviation effects at the stopping time. Under A , for example, every chosen model contributes a strictly positive expected increment to L_t . Remaining on the wrong side of the routing boundary after many queries therefore requires a large unfavorable martingale fluctuation. Using a concentration argument, we show that the expected number of rounds spent on the wrong side is $O(1)$. The same reasoning applies under B . The final query may also supply slightly more evidence than is needed to stop. Accounting for these effects, the policy's expected query cost and mean waiting time remain within a constant of their respective benchmark values under each hypothesis.

The second part of the proof controls the fluctuations of the waiting times accumulated over those queries. When the waiting penalty is nonlinear, finishing early on some paths does not fully offset the additional cost of finishing late on others. Matching the benchmark's mean delay is therefore not sufficient. The proof must also control two sources of variation: how many queries are needed before the evidence is decisive, and how long each query takes. The concentration bounds for accumulated evidence control the first, while the sub-Gaussian waiting-time assumption controls the second. Together, they ensure that unusually long query sequences do not dominate the expected waiting cost.

4.4. Tightness of the Asymptotic Order

The preceding theorem leaves a natural question: is its remainder caused by the simplicity of 2LLM, or by an unavoidable difference between the lower bound program and the optimal policy? The next result shows that, for strictly convex power penalties, even an exact optimal policy remains separated from Ψ_α by a term of this order.

THEOREM 3 (Tightness of the convergence rate). *Suppose $c > 0$ and $\rho > 1$. For every fixed problem instance satisfying Assumptions 1–3, there exist constants $K > 0$ and $\underline{\alpha} > 0$ such that, for all $\alpha \in (0, \underline{\alpha})$,*

$$\inf_{\pi \in \Pi(\alpha)} \mathcal{R}(\pi; \alpha) \geq \Psi_\alpha + K(\log(1/\alpha))^{\rho-1}. \quad (13)$$

Theorem 3 shows that the remaining gap to the benchmark is not simply a consequence of restricting the policy to two LLMs, but a fundamental gap between an optimal policy and the

deterministic benchmark because the latter neglects uncertainty in its formulation. To understand this difference, decompose the gap for an admissible policy π as

$$\mathcal{R}(\pi; \alpha) - \Psi_\alpha = \underbrace{F(\mathbf{n}_A^\pi, \mathbf{n}_B^\pi) - \Psi_\alpha}_{\text{allocation gap}} + \underbrace{\sum_{\theta \in \{A, B\}} \xi_\theta [\mathbb{E}_\theta[g(W_\pi)] - g(\mathbb{E}_\theta[W_\pi])]}_{\text{fluctuation gap}}.$$

The allocation gap measures the gap between policy π 's expected query allocation costs and that under the optimal deterministic allocation. It reflects both which models are queried and how much evidence is collected before stopping. The fluctuation gap is the additional waiting penalty caused by variation in completion time. With a strictly convex penalty, finishing earlier on some paths does not fully offset the extra cost of finishing later on others.

Given our construction, minimizing the total weighted cost requires striking the right balance between these two gaps. For example, a policy that always queries the optimal deterministic LLMs has no allocation gap, but its waiting time is essentially a random sum over $\Theta(\log(1/\alpha))$ queries, and its fluctuation gap is of order $(\log(1/\alpha))^{\rho-1}$. On the other hand, if we want to reduce the variability of the waiting time, any policy will need to query another LLM with different delay profiles. Any such query is not time-efficient, and since the total waiting time grows in the order of $\log(1/\alpha)$, the marginal delay penalty grows in the order of $(\log(1/\alpha))^{\rho-1}$.

5. Value of Querying Two LLMs

The previous section establishes that two LLMs are sufficient for asymptotic optimality, but sufficiency does not imply that both are necessary. We now ask when a single model can attain the same leading-order performance and how much is lost when it cannot. A single-model policy is still sequential: it can stop as soon as the posterior meets the confidence requirement. Thus, this section isolates the value of adapting *which* model is queried, rather than the value of adapting *when* to stop.

Let $\Pi^{(1)}(\alpha)$ be the class of policies that select one LLM before observing any outputs, use that model for every query, and apply the posterior stopping rule in Section 2.1. The model can be chosen using the prior, error tolerance, and all known model characteristics, but cannot depend on the unknown ground truth. Motivated by the lower bound program (LB), we define the following lower bound.

$$\Psi_\alpha^{(1)} := \min_{1 \leq j \leq M} \left\{ \xi_A S_A(\alpha) \kappa_{j,A} + \xi_B S_B(\alpha) \kappa_{j,B} + \xi_A g(S_A(\alpha) \eta_{j,A}) + \xi_B g(S_B(\alpha) \eta_{j,B}) \right\}. \quad (14)$$

This is the lowest cost of supplying both hypothesis-specific information budgets with the same model. It allows the required number of queries to differ under A and B , but not the queried LLM.

Equivalently, it restricts (10) to the diagonal pairs $(\mathbf{e}_j, \mathbf{e}_j)$. Therefore, the lower-bound argument in Proposition 3 gives

$$\inf_{\pi \in \Pi^{(1)}(\alpha)} \mathcal{R}(\pi; \alpha) \geq \Psi_\alpha^{(1)} \geq \Psi_\alpha.$$

The single-LLM case of the proof of Theorem 2 also shows that repeatedly querying a minimizer in (14) attains $\Psi_\alpha^{(1)}$ up to $O((\log(1/\alpha))^{\rho-1})$. Comparing the two deterministic benchmarks therefore reveals the leading-order value of allowing hypothesis-specific specialization, rather than merely comparing two potentially loose bounds.

A single LLM suffices when the two hypothesis-specific benchmark problems have a common optimal model. In that case, for some $j^\sharp$ and all sufficiently small α ,

$$\Psi_\alpha^{(1)} = G_\alpha(\mathbf{e}_{j^\sharp}, \mathbf{e}_{j^\sharp}) = \Psi_\alpha. \quad (15)$$

A sufficient condition is that this model weakly dominates every other model in both cost-efficiency and time-efficiency under both hypotheses:

$$\kappa_{j^\sharp, A} \leq \kappa_{j, A}, \quad \kappa_{j^\sharp, B} \leq \kappa_{j, B}, \quad \eta_{j^\sharp, A} \leq \eta_{j, A}, \quad \eta_{j^\sharp, B} \leq \eta_{j, B}, \quad j \in [M].$$

The two specialists may then be chosen as $j_A^\star = j_B^\star = j^\sharp$, and, for $c > 0$, the resulting single-model policy is asymptotically optimal by Theorem 2. This conclusion is useful because it identifies when routing has no leading-order benefit: one model supplies evidence efficiently regardless of which hypothesis turns out to be true.

In many of the other scenarios, however, there does not exist a single LLM that achieves the lowest per-unit information cost and waiting time under both hypotheses simultaneously. For instance, suppose there are two LLMs such that $\kappa_{1,A} \ll \kappa_{2,A}$ and $\eta_{1,A} \ll \eta_{2,A}$, but $\kappa_{2,B} \ll \kappa_{1,B}$ and $\eta_{2,B} \ll \eta_{1,B}$. In other words, LLM 1 is much more efficient under hypothesis A (since it provides information faster and cheaper) while LLM 2 is much more efficient under hypothesis B . Any single-LLM design must compromise between these conflicting efficiencies and will incur a larger value that is at least $\Psi_\alpha^{(1)}$. In fact, we can construct a family of examples where the gap between $\Psi_\alpha^{(1)}$ and Ψ_α is very large.

EXAMPLE 1 (QUANTIFYING THE VALUE OF QUERYING TWO LLMs). Let $c = 1$ such that $g(x) = x^\rho$. Consider a symmetric prior $\xi_A = \xi_B = 1/2$ and two LLMs defined in Table 2

Table 2 Parameters for Example 1					
LLM j	$I_{j,A}$	$I_{j,B}$	c_j	μ_j	Specialization
1	5	1	10	10	A -specialist
2	1	5	10	10	B -specialist

Let $S(\alpha) := S_A(\alpha) = S_B(\alpha) = \log((1 - \alpha)/\alpha) - C_0$, which is positive for sufficiently small α . Selecting either model as the sole source gives

$$\Psi_\alpha^{(1)} = 6S(\alpha) + \frac{1}{2}g(2S(\alpha)) + \frac{1}{2}g(10S(\alpha)).$$

Under its favorable hypothesis, the selected model requires benchmark mean delay $2S(\alpha)$; under the other hypothesis, it requires $10S(\alpha)$. By contrast, model 1 dominates under A and model 2 dominates under B , so the two-specialist benchmark is

$$\Psi_\alpha = 2S(\alpha) + g(2S(\alpha)).$$

Consequently,

$$\frac{\Psi_\alpha^{(1)}}{\Psi_\alpha} = \frac{6S(\alpha) + \frac{1}{2}(2S(\alpha))^\rho + \frac{1}{2}(10S(\alpha))^\rho}{2S(\alpha) + (2S(\alpha))^\rho} \rightarrow \frac{1 + 5^\rho}{2}.$$

The best single-model policy therefore has asymptotic cost three times as large when $\rho = 1$, and thirteen times as large when $\rho = 2$. The fivefold disadvantage under the unfavorable hypothesis is paid on half the instances. A convex waiting penalty amplifies this disadvantage, which explains why the ratio is much larger under quadratic than linear waiting cost. Since the models have identical per-query costs and mean latencies, the gain comes entirely from obtaining more useful evidence per query under the relevant hypothesis.

The ratios can be made arbitrarily large by increasing the contrast between the two directional information rates. Such contrasts are compatible with the accuracy model: keeping one accuracy strictly between $1/2$ and 1 while letting the other approach 1 makes one directional KL divergence grow without bound while the reverse divergence stays bounded; swapping the accuracies produces a complementary model. From a managerial perspective, this example shows us that the value of a second LLM depends on complementarity, not diversity alone. Access to another model is valuable when it improves information acquisition in a direction that the first model handles inefficiently. $\square$

6. Value of Adaptivity

The single-model policies in Section 5 still adapt their stopping time to the observed evidence. We now compare sequential querying with a static design that fixes every model's query count before observing any outputs. Throughout this section, we retain the special case of zero waiting cost, $g(w) \equiv 0$, or equivalently $c = 0$ in Assumption 3. The objective becomes

$$\mathcal{R}(\pi; \alpha) = \mathbb{E}[C_\pi] = \mathbb{E}\left[\sum_{t=1}^{\tau} c_{j_t}\right].$$

Waiting times may still be positive, but their penalty is zero. This specialization isolates the monetary value of responding to the evidence without combining it with the curvature of the delay

penalty. The static benchmark is allowed to choose both the total query count and its allocation across all available models. Any advantage of sequential querying, therefore, reflects the ability to react to outputs, rather than a comparison with an arbitrarily chosen fixed budget.

Formally, a static design is a vector $\mathbf{n} = (n_1, \dots, n_M) \in \{0, 1, \dots\}^M$, where n_j is the number of queries planned for LLM j . All planned queries are completed before the terminal decision is made. Under Assumption 1, their order does not affect the distribution of accumulated evidence; with $g \equiv 0$, the cost is simply $\sum_j c_j n_j$. Let $\omega := \sum_j n_j$, and define the positive, hypothesis-aligned mean evidence and its variance by

$$\begin{aligned}\mu_A(\mathbf{n}) &:= \mathbb{E}_A[L_\omega] = \sum_{j=1}^M n_j I_{j,A}, & \mu_B(\mathbf{n}) &:= \mathbb{E}_B[-L_\omega] = \sum_{j=1}^M n_j I_{j,B}, \\ \sigma_\theta^2(\mathbf{n}) &:= \text{Var}_\theta(L_\omega) = \sum_{j=1}^M n_j V_{j,\theta}, & V_{j,\theta} &:= \text{Var}_\theta(\ell_j(Y)) > 0.\end{aligned}$$

After the ω queries, the design selects the hypothesis with the larger posterior probability:

$$D := \begin{cases} A, & L_\omega + \delta \geq 0, \\ B, & L_\omega + \delta < 0. \end{cases} \quad (16)$$

We require $\mathbb{P}(D \neq \theta) \leq \alpha$ and denote the feasible static designs by $\Pi^{\text{st}}(\alpha)$. This unconditional Bayes-error requirement is weaker than requiring posterior error at most α on every stopped history, as in the sequential class. In other words, the comparison does not impose the sequential class' stronger sample-path level requirement on static designs. The static policy class $\Pi^{\text{st}}(\alpha)$ is, therefore, not a subclass of the admissible sequential policy $\Pi(\alpha)$.

Let $\mathcal{R}^{\text{st},*}(\alpha) := \inf_{\pi \in \Pi^{\text{st}}(\alpha)} \mathcal{R}(\pi; \alpha)$ denote the optimal static cost. Write $\xi_{\min} := \min\{\xi_A, \xi_B\}$ and take $\alpha < \xi_{\min}$, so a feasible design must make at least one query. Since the Bayes error is $\xi_A \mathbb{P}_A(D = B) + \xi_B \mathbb{P}_B(D = A)$, feasibility implies

$$\mathbb{P}_A(L_\omega + \delta < 0) \leq \frac{\alpha}{\xi_A}, \quad \mathbb{P}_B(L_\omega + \delta \geq 0) \leq \frac{\alpha}{\xi_B}. \quad (17)$$

These are necessary conditions, not a sufficient pair of conditions for Bayes error at most α .

To understand what a fixed query budget must achieve, consider its terminal evidence relative to its fluctuations. Under a static design, the LLR is a sum of independent, though generally nonidentically distributed, increments. A normal approximation suggests

$$\begin{aligned}\mathbb{P}_A(L_\omega + \delta < 0) &\approx \Phi\left(-\frac{\mu_A(\mathbf{n}) + \delta}{\sigma_A(\mathbf{n})}\right), \\ \mathbb{P}_B(L_\omega + \delta \geq 0) &\approx \Phi\left(-\frac{\mu_B(\mathbf{n}) - \delta}{\sigma_B(\mathbf{n})}\right),\end{aligned}$$

where Φ is the standard normal distribution function. Thus, a large expected amount of evidence is not sufficient by itself: it must also be large relative to its standard deviation. This distinction

is important because a fixed horizon must control the probability of an unfavorable evidence realization at that horizon. The following proposition retains the Berry–Esseen approximation error, making the necessary conditions rigorous rather than treating the normal approximation as exact.

PROPOSITION 5. *Consider a static design with at least one query and terminal decision (16). Define*

$$\mathcal{T}_\theta(\mathbf{n}) := \sum_{j=1}^M n_j \mathbb{E}_\theta \left[|\ell_j(Y) - \mathbb{E}_\theta[\ell_j(Y)]|^3 \right].$$

If $\mathbb{P}(D \neq \theta) \leq \alpha$, then

$$\begin{aligned} \frac{\mu_A(\mathbf{n}) + \delta}{\sigma_A(\mathbf{n})} &\geq \nu \left(\frac{\alpha}{\xi_A} + \frac{\mathcal{T}_A(\mathbf{n})}{\sigma_A(\mathbf{n})^3} \right), \\ \frac{\mu_B(\mathbf{n}) - \delta}{\sigma_B(\mathbf{n})} &\geq \nu \left(\frac{\alpha}{\xi_B} + \frac{\mathcal{T}_B(\mathbf{n})}{\sigma_B(\mathbf{n})^3} \right), \end{aligned} \quad (18)$$

where

$$\nu(x) := \begin{cases} \infty, & x \leq 0, \\ \Phi^{-1}(1-x), & 0 < x < 1, \\ -\infty, & x \geq 1. \end{cases}$$

The correction $\mathcal{T}_\theta(\mathbf{n})/\sigma_\theta(\mathbf{n})^3$ measures the possible error of the normal approximation. With very few queries, this correction can be large enough that the constraints are uninformative. We therefore also retain a necessary information condition as follows: For every feasible design,

$$\mathbb{P}_A(D = B) + \mathbb{P}_B(D = A) \leq \frac{\alpha}{\xi_{\min}}.$$

The distributions of the complete output vector under A and B must consequently have total variation distance at least $1 - \alpha/\xi_{\min}$. Pinsker's inequality in either direction gives

$$\mu_\theta(\mathbf{n}) \geq b_\alpha := 2 \left(1 - \frac{\alpha}{\xi_{\min}} \right)^2 > 0, \quad \theta \in \{A, B\}.$$

This requirement rules out designs whose cumulative signal is too small to distinguish the hypotheses, even when the normal-approximation constraints happen to be vacuous. Combining these necessary conditions and relaxing integrality yields

$$\begin{aligned} \Psi_\alpha^{\text{st}} &:= \min_{\mathbf{n} \in \mathbb{R}_+^M} \sum_{j=1}^M c_j n_j \\ \text{s.t.} \quad &\frac{\mu_A(\mathbf{n}) + \delta}{\sigma_A(\mathbf{n})} \geq \nu \left(\frac{\alpha}{\xi_A} + \frac{\mathcal{T}_A(\mathbf{n})}{\sigma_A(\mathbf{n})^3} \right), \\ &\frac{\mu_B(\mathbf{n}) - \delta}{\sigma_B(\mathbf{n})} \geq \nu \left(\frac{\alpha}{\xi_B} + \frac{\mathcal{T}_B(\mathbf{n})}{\sigma_B(\mathbf{n})^3} \right), \\ &\mu_\theta(\mathbf{n}) \geq b_\alpha, \quad \theta \in \{A, B\}. \end{aligned} \quad (19)$$

Every feasible static design is feasible for this program, so $\Psi_\alpha^{\text{st}} \leq \mathcal{R}^{\text{st},*}(\alpha)$. The positive information constraints also exclude the zero-query vector, for which the standardized expressions are undefined.

We next study a weak-signal regime in which each individual output is only slightly more informative about the truth. This regime makes the distinction between fixed-budget and sequential querying especially clear: many observations are needed, and the cost of continuing after a favorable evidence realization can be compared on the same scale as the total query budget.

DEFINITION 3 (WEAK-SIGNAL SCALING). Fix M , positive constants a_j, b_j for $j \in [M]$, the prior, monetary query costs, and waiting-time distributions. For

$$0 < \epsilon < \frac{1}{2 \max_{j \in [M]} \{a_j, b_j\}},$$

set

$$\gamma_{j,A} = \frac{1}{2} + a_j \epsilon, \quad \gamma_{j,B} = \frac{1}{2} + b_j \epsilon, \quad j = 1, \dots, M.$$

Weak-signal scaling means taking $\epsilon \rightarrow 0$ while these other primitives remain fixed. As ϵ decreases, the output distributions under the two hypotheses approach each other. The following theorem further characterizes the value of adaptivity explicitly in a limiting regime.

THEOREM 4 (Value of adaptivity with zero waiting cost). Suppose $g(w) \equiv 0$, and let $\mathcal{R}^*(\alpha) := \inf_{\pi \in \Pi(\alpha)} \mathcal{R}(\pi; \alpha)$. Under weak-signal scaling,

$$\lim_{\alpha \rightarrow 0} \liminf_{\epsilon \rightarrow 0} \frac{\mathcal{R}^{\text{st},*}(\alpha)}{\mathcal{R}^*(\alpha)} = \lim_{\alpha \rightarrow 0} \limsup_{\epsilon \rightarrow 0} \frac{\mathcal{R}^{\text{st},*}(\alpha)}{\mathcal{R}^*(\alpha)} = 4.$$

The theorem establishes a factor-of-four cost difference when outputs first become weakly informative and the error tolerance then approaches zero. To see where the constant four comes from, write $s_j := a_j + b_j$. The expansions in the appendix give

$$I_{j,\theta} = 2s_j^2 \epsilon^2 + O(\epsilon^4), \quad V_{j,\theta} = 4s_j^2 \epsilon^2 + O(\epsilon^4), \quad \theta \in \{A, B\}.$$

For a static design that uses model j , the standardized mean evidence is approximately $s_j \epsilon \sqrt{n}$. To make the terminal error small, this quantity must reach a normal quantile of order $\sqrt{2 \log(1/\alpha)}$. A sequential policy instead stops when the realized evidence reaches a boundary of order $\log(1/\alpha)$; dividing that requirement by the drift gives its leading expected query count. The two scales are therefore

$$\begin{aligned} \text{Static query count} &\approx \frac{2 \log(1/\alpha)}{s_j^2 \epsilon^2}, \\ \text{Sequential expected query count} &\approx \frac{\log(1/\alpha)}{2s_j^2 \epsilon^2}. \end{aligned}$$

These are leading-scale interpretations in the ordered limits, not finite-parameter formulas. Their ratio explains the factor four. A fixed horizon must be long enough to control unfavorable evidence fluctuations, and that entire budget is paid even on paths that become decisive much earlier.

Sequential stopping uses more queries on difficult paths and fewer on favorable ones, avoiding the cost of continuing after the required confidence has been reached.

Together, Sections 5 and 6 identify two distinct benefits of responding to evidence. Model selection exploits its direction, assigning queries to a specialist for the more plausible hypothesis. Sequential stopping exploits its amount, ending the process once enough evidence has accumulated. Efficient orchestration can benefit from both, even though the two sections isolate them in different comparisons.

7. Numerical Experiments

In this section, we empirically evaluate the performance of Policy 2LLM using a pool of 18 distinct LLMs. We consider a binary classification task based on the μ -MATH (Meta U-MATH) benchmark (Chernyshev et al. 2025), which contains 1,084 labeled tasks, each containing a university-level math question, one candidate solution, and ground truth labels for the correctness of the solution. The ground-truth correctness labels were obtained through a combination of automatic verification and manual annotation by mathematical experts. LLMs are asked to judge whether the candidate solution is correct or not without knowing the ground truth. The selected 18 LLMs exhibit heterogeneous verification accuracy, waiting time distribution, and inference cost. We measure these model-specific quantities based on their performance and API usage in our experiments.

We identify hypothesis A with a correct candidate solution and hypothesis B with an incorrect candidate solution. Correct and incorrect instances are sampled in equal proportions, so we use the symmetric prior $\xi_A = \xi_B = 1/2$. Hence $\delta = 0$ and the two stopping boundaries are symmetric:

$$A_\alpha = B_\alpha = H_\alpha := \log \frac{1 - \alpha}{\alpha}.$$

The waiting-time penalty is $g(w) = C_w w^\rho$, with $C_w > 0$ and $\rho \geq 1$. Under this task mapping, the true-positive rate of LLM j is $\gamma_{j,A}$ and its true-negative rate is $\gamma_{j,B}$; together they determine the two information rates $I_{j,A}$ and $I_{j,B}$. The remaining inputs are the per-question cost c_j and mean response time μ_j .

In Figure 3, the horizontal axis reports the true-positive rate $\gamma_{j,A}$, the vertical axis reports the true-negative rate $\gamma_{j,B}$, the color represents the per-question cost, and the numbers identify the corresponding models. The figure shows substantial heterogeneity along both accuracy dimensions, while monetary cost is not monotonically related to either one. This is precisely the setting in which a single accuracy ranking is insufficient: a model can be informative for certifying one hypothesis without being as informative for the other. At the same time, Figure 3 alone does not determine the specialists, because specialist selection depends on information-normalized monetary cost and latency, not on raw accuracy or price in isolation.

Proposition 4 motivates selecting one specialist for each hypothesis by balancing monetary cost and mean latency per unit of information. Under the symmetric prior, the lower-bound budgets in Section 4 are $S_A(\alpha) = S_B(\alpha) = H_\alpha - C_0$. Recall that the constant C_0 is a conservative slack introduced by the proof of the universal lower bound. For the numerical implementation, we simply set $C_0 = 0$, which preserves the leading high-confidence scale because $S_\theta(\alpha) = H_\alpha + O(1)$ while avoiding calibration of a proof-specific slack constant.

Specifically, for $\theta \in \{A, B\}$ we select

$$j_\theta^*(\alpha, C_w, \rho) \in \arg \min_{j \in [M]} \left\{ \frac{H_\alpha c_j}{I_{j,\theta}} + C_w \left(\frac{H_\alpha \mu_j}{I_{j,\theta}} \right)^\rho \right\}.$$

We refer to these as the A - and B -specialists. Once the pair is selected, online routing follows Definition 2: because $\delta = 0$, the policy queries the A -specialist when the current LLR is nonnegative and the B -specialist otherwise. Realized waiting times affect the realized cost, but not the online routing decision.

We evaluate policy performance using pathwise Monte Carlo simulation. For each simulated path, we first sample the true hypothesis according to the equal prior. At each query, the policy selects the A - or B -specialist according to the sign of the current LLR, samples the selected model’s binary response using its true-positive and true-negative rates, samples a model-specific waiting time, updates the LLR, and continues until one of the stopping boundaries is crossed.

For a path with a stopping time τ , the realized total cost is

$$\sum_{t=1}^{\tau} c_{j_t} + C_w \left(\sum_{t=1}^{\tau} T_{j_t, t} \right)^\rho.$$

Averaging this quantity over simulated paths directly estimates the expected nonlinear waiting-time cost of Policy 2LLM. For the experiments with $\rho = 1$ and varying C_w , we use 200,000 simulated paths per parameter setting. For the experiments with varying ρ , we use 1,000,000 simulated paths

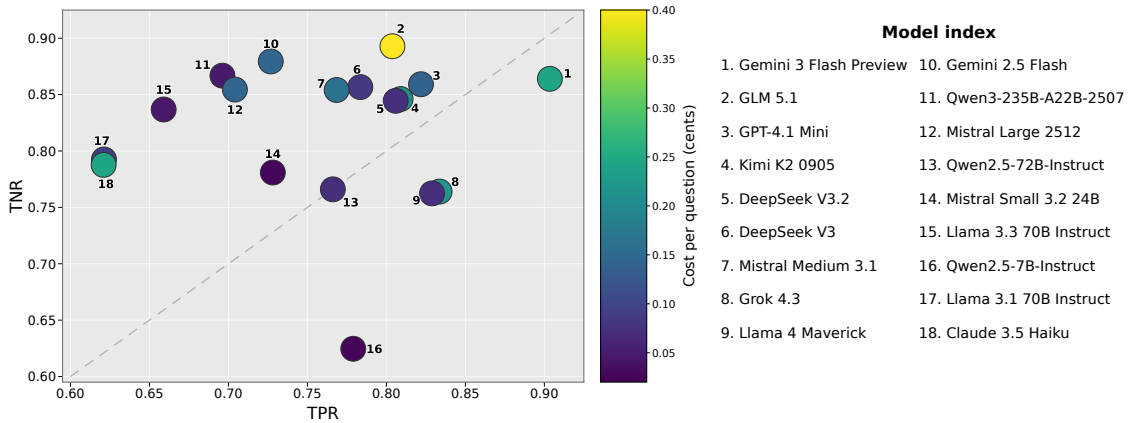


Figure 3 Model characteristics for the 18 LLMs used in the numerical experiments.

per parameter setting. The sample size is larger for the latter because the convex waiting-time penalty for $\rho > 1$ increases the variability of the simulated path costs. Further implementation details, including how we construct the waiting-time distributions, are provided in Appendix E.1.

To assess the performance of 2LLM under finite α , we compute a benchmark based on solving approximate Bellman equations, where we replace the random waiting time with its mean, i.e., $T_{j,t} \equiv \mu_j$ for all j and t . This deterministic-latency model allows the decision maker to use all 18 LLMs and preserve the nonlinear dependence on accumulated delay, but it removes the model-specific waiting-time fluctuations. In particular, when $\rho = 1$, the expected waiting penalty is linear, so replacing the random waiting time by its mean does not change the expected objective. When $\rho > 1$, the waiting penalty depends on the accumulated waiting time. Because the original stochastic policy class may also condition future choices on realized delays, we use this value as a deterministic-latency benchmark rather than as a formally exact lower bound for the full stochastic control problem. We further approximately solve the Bellman equations by discretizing the state space; the precise Bellman recursion and computation details are reported in Appendix E.2.

7.1. Numerical Results

We report the percentage relative difference between the pathwise Monte Carlo estimate of the two-LLM policy and the corresponding Bellman benchmark:

$$\text{Gap}(\alpha, \rho) = 100\% \times \frac{\widehat{\mathcal{R}}_{2\text{LLM}}^{\text{MC}}(\alpha, \rho) - \mathcal{R}_{\text{Bellman}}^{\text{det}}(\alpha, \rho)}{\mathcal{R}_{\text{Bellman}}^{\text{det}}(\alpha, \rho)}.$$

Here, $\widehat{\mathcal{R}}_{2\text{LLM}}^{\text{MC}}$ is the pathwise Monte Carlo estimate for the sign-based two-LLM policy and $\mathcal{R}_{\text{Bellman}}^{\text{det}}$ is the value of the Bellman problem evaluated with model-specific mean latencies. The interpretation of this quantity follows the benchmark discussion above. When $\rho = 1$, expected waiting time is preserved when using the average value, so the gap directly compares Policy 2LLM with a lower bound on the optimal policy (up to Bellman-grid discretization and Monte Carlo error). When $\rho > 1$, the numerator compares the stochastic pathwise evaluation of Policy 2LLM with a deterministic Bellman benchmark.

We first vary the waiting-cost weight C_w while fixing $\rho = 1$. This experiment changes the relative importance of waiting time while retaining the single-dimensional Bellman formulation. Figure 4(a) shows that the largest differences occur at $x = 1$, when the stopping boundaries are relatively close and local routing details matter most. The curves drop sharply once the confidence requirement becomes more stringent and, at larger x , cluster at only a few percentage points across a broad range of waiting-cost weights. The decline is not perfectly monotone for every C_w , which is unsurprising at finite α : changing C_w can change the selected specialists, and the reported values also reflect numerical and Monte Carlo approximation. The overall pattern is nevertheless consistent with the

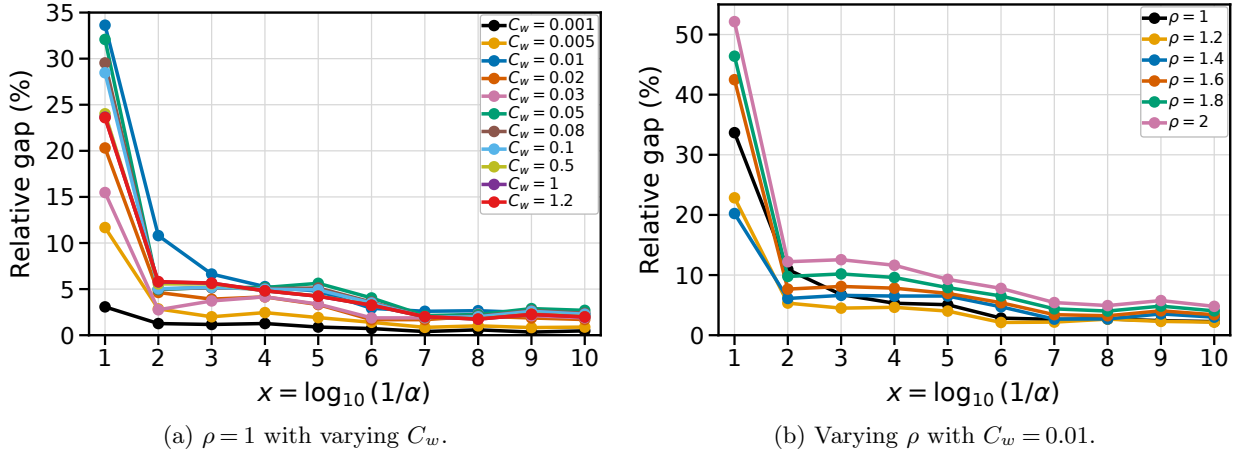


Figure 4 Relative gap between the pathwise Monte Carlo estimate of Policy 2LLM and the corresponding Bellman benchmark.

high-confidence result in Theorem 2: for fixed model characteristics and waiting-cost parameters, the cost of restricting to two specialists becomes increasingly inexpensive as α decreases.

We next fix $C_w = 0.01$ and vary the curvature parameter ρ . Figure 4(b) shows that the reported differences are generally larger when $\rho > 1$ than when $\rho = 1$. This comparison should be interpreted with care. Under Monte Carlo evaluation, Policy 2LLM pays the convex penalty on the realized total waiting time, so paths with higher waiting time receive disproportionate weight. The deterministic-latency Bellman benchmark removes precisely these fluctuations before applying the penalty. Therefore, part of the larger reported gap for $\rho > 1$ reflects this fluctuation mechanism, rather than the limitation of using only two LLMs; see the discussion following Theorem 3. As a result, the curves are not strictly ordered monotonically in ρ . Despite these computational effects, the reported gap declines substantially as the confidence requirement becomes more stringent for every tested value of ρ .

Taken together, the two panels of Figure 4 provide complementary finite- α evidence about the proposed policy. In both cases, the diminishing gaps at higher confidence levels are qualitatively consistent with the asymptotic prediction that Policy 2LLM becomes increasingly effective as $\alpha \rightarrow 0$.

8. Conclusion

This paper studies a sequential decision problem motivated by modern AI systems, where a decision-maker must orchestrate multiple heterogeneous information sources under cost, waiting time, and accuracy trade-offs. We formulate a Bayesian sequential hypothesis testing model that captures key operational features of LLM deployment, including asymmetric diagnostic power, stochastic waiting times, and convex delay penalties.

Our main result is a sharp asymptotic characterization of optimal performance in the high-confidence regime. We establish a universal lower bound through a deterministic convex program and show that an asymptotically optimal policy needs at most two information sources. Motivated by this structure, we propose a simple sign-based two-LLM policy and prove that it achieves the lower bound up to a tight remainder.

These results yield a simple design insight: near-optimal sequential information acquisition does not require coordinating many models, but rather identifying two complementary specialists and switching between them as evidence evolves. The numerical study further illustrates the finite- α performance of this two-LLM policy relative to the Bellman benchmarks. Taken together, our findings show that effective AI decision systems can remain structurally simple even as the number of available models grows.

References

- Abdallah A, Maarof MA, Zainal A (2016) Fraud detection system: A survey. *Journal of Network and Computer Applications* 68:90–113.
- Aggarwal P, Madaan A, Yang Y, Mausam (2023) Let’s sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. *Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing*, 12375–12396.
- Ao R, Luo G, Simchi-Levi D, Wang X (2025) Optimizing LLM inference: Fluid-guided online scheduling with memory constraints. *arXiv preprint arXiv:2504.11320* .
- Artificial Analysis (2026) The artificial analysis intelligence index. URL <https://artificialanalysis.ai/>.
- Bari A, Hegde P, de Veciana G (2025) Optimal scheduling algorithms for LLM inference: Theory and practice. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 9(3).
- Belloni A, Chen Y, Wei Y (2026) Online pandora’s box for contextual LLM cascading. *arXiv preprint arXiv:2606.07392* .
- Bingham NH, Goldie CM, Teugels JL (1989) *Regular Variation*, volume 27 (Cambridge University Press).
- Chen L, Davis J, Hanin B, Bailis P, Stoica I, Zaharia M, Zou J (2024a) Are more LLM calls all you need? towards the scaling properties of compound AI systems. *Proceedings of the 38th Conference on Neural Information Processing Systems*, 45767–45790.
- Chen L, Zaharia M, Zou J (2024b) FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research* .
- Chen Z, Ye Y, Zhou Z (2025) Adaptively robust LLM inference optimization under prediction uncertainty. *arXiv preprint arXiv:2508.14544* .
- Chernoff H (1959) Sequential design of experiments. *The Annals of Mathematical Statistics* 30(3):755–770.

-
- Chernyshev K, Polshkov V, Stepanov V, Myasnikov A, Artemova E, Miasnikov A, Tilga S (2025) U-MATH: A university-level benchmark for evaluating mathematical skills in large language models. *Proceedings of the 4th Workshop on Generation, Evaluation and Metrics*, 974–1001.
- Dayanik S, Sezer SO (2012) Multisource Bayesian sequential binary hypothesis testing problem. *Annals of Operations Research* 201(1):99–130.
- Ding D, Mallick A, Wang C, Sim R, Mukherjee S, Rühle V, Lakshmanan LVS, Awadallah AH (2024) Hybrid LLM: Cost-efficient and quality-aware query routing. *Proceedings of the 12th International Conference on Learning Representations*.
- Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I (2024) Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning*.
- Durrett R (2019) *Probability: Theory and Examples*, volume 49 (Cambridge University Press).
- Freedman D (1975) On tail probabilities for martingales. *The Annals of Probability* 3(1):100–118.
- Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1).
- Gou Z, Shao Z, Gong Y, Yang Y, Duan N, Chen W (2024) CRITIC: Large language models can self-correct with tool-interactive critiquing. *Proceedings of the 12th International Conference on Learning Representations*.
- Guo Y, Jasin S, Lin CAA (2026) Congestion-aware static LLM cascades: Analysis of a steady-state framework. *Available at SSRN 6759238* .
- Hall P, Heyde CC (1980) *Martingale Limit Theory and Its Application* (Academic Press).
- Huang B, Li S, Wu T, Yang Y, Talwalkar A, Ramchandran K, Jordan MI, Jiao J (2026a) Sample complexity and representation ability of test-time scaling paradigms. *Proceedings of the 14th International Conference on Learning Representations*.
- Huang J, Ma W, Zhou Z (2026b) Optimal Bayesian stopping for efficient inference of consistent LLM answers. *Proceedings of the 43rd International Conference on Machine Learning*.
- Jaillet P, Jiang J, Mellou K, Molinaro M, Podimata C, Zhou Z (2025) Online scheduling for LLM inference with KV cache constraints. *arXiv preprint arXiv:2502.07115* .
- Jung J, Brahman F, Choi Y (2025) Trust or escalate: LLM judges with provable guarantees for human agreement. *Proceedings of the 13th International Conference on Learning Representations*.
- Khraisat A, Gondal I, Vamplew P, Kamruzzaman J (2019) Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity* 2(20).
- Kim A, Kim K, Yoon S (2024) DEBATE: Devil’s advocate-based assessment and text evaluation. *Findings of the Association for Computational Linguistics: ACL 2024*, 1885–1897.

- Lambez T, Cohen K (2022) Anomaly search with multiple plays under delay and switching costs. *IEEE Transactions on Signal Processing* 70:174–189.
- Li Y, Dai J, Peng T (2025) Throughput-optimal scheduling algorithms for LLM inference and AI agents. *arXiv preprint arXiv:2504.07347* .
- Li Y, Yuan P, Feng S, Pan B, Wang X, Sun B, Wang H, Li K (2024) Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *Proceedings of the 12th International Conference on Learning Representations*, 14751–14768.
- Mills JP (1926) Table of the ratio: Area to bounding ordinate, for any portion of normal curve. *Biometrika* 18(3/4):395–400.
- Naghshvar M, Javidi T (2013) Active sequential hypothesis testing. *The Annals of Statistics* 41(6):2703–2738.
- Ong I, Almahairi A, Wu V, Chiang WL, Wu T, Gonzalez JE, Kadous MW, Stoica I (2025) RouteLLM: Learning to route LLMs from preference data. *Proceedings of the 13th International Conference on Learning Representations*.
- OpenAI (2025) Introducing GPT-5. URL <https://openai.com/index/introducing-gpt-5/>.
- Setlur A, Nagpal C, Fisch A, Geng X, Eisenstein J, Agarwal R, Agarwal A, Berant J, Kumar A (2025) Rewarding progress: Scaling automated process verifiers for LLM reasoning. *Proceedings of the 13th International Conference on Learning Representations*.
- Shevtsova I (2010) An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, volume 82, 862–864.
- Singhi N, Bansal H, Hosseini A, Grover A, Chang KW, Rohrbach M, Rohrbach A (2025) When to solve, when to verify: Compute-optimal problem solving and generative verification for LLM reasoning. *Proceedings of the 2nd Conference on Language Modeling*.
- Snell C, Lee J, Xu K, Kumar A (2025) Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. *Proceedings of the 13th International Conference on Learning Representations*.
- Song Y, Fellouris G (2017) Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals. *Electronic Journal of Statistics* 11(1):338–363.
- Tsopelakos A, Fellouris G (2023) Sequential anomaly detection under sampling constraints. *IEEE Transactions on Information Theory* 69(12):8126–8146.
- Vaidhiyan NK, Sundaresan R (2015) Active search with a cost for switching actions. *arXiv preprint arXiv:1505.02358* .
- Verga P, Hofstatter S, Althammer S, Su Y, Piktus A, Arkhangorodsky A, Xu M, White N, Lewis P (2024) Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796* .

-
- Viola PA, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Wald A (1945) Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* 16(2):117–186.
- Wang J, Wang J, Athiwaratkun B, Zhang C, Zou JY (2025a) Mixture-of-agents enhances large language model capabilities. *Proceedings of the 13th International Conference on Learning Representations*.
- Wang M, Ye Y, Zhou Z (2025b) LLM serving optimization with variable prefill and decode lengths. *arXiv preprint arXiv:2508.06133* .
- Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D (2022) Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* .
- Wu Y, Sun Z, Li S, Welleck S, Yang Y (2025) Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. *Proceedings of the 13th International Conference on Learning Representations*.
- Xu Z, Kusner M, Weinberger K, Chen M (2013) Cost-sensitive tree of classifiers. *Proceedings of the 30th International Conference on Machine Learning*.
- Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, Narasimhan K (2023a) Tree of thoughts: Deliberate problem solving with large language models. *Proceedings of the 37th Conference on Neural Information Processing Systems*.
- Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y (2023b) ReAct: Synergizing reasoning and acting in language models. *Proceedings of the 11th International Conference on Learning Representations*.
- Zhang L, Hosseini A, Bansal H, Kazemi SM, Kumar A, Agarwal R (2025) Generative verifiers: Reward modeling as next-token prediction. *Proceedings of the 13th International Conference on Learning Representations*.

Appendix

The appendix provides omitted proofs in the main text, and is organized by sections.

Appendix A: Omitted Proofs in Section 2

A.1. Proof of Lemma 1

We use the likelihood of the *joint sequential history*, rather than conditioning on the completed sequence of model choices. The latter sequence is adaptive and can reveal earlier observations. Conditional on each pre-query history, the same selection rule is used under both hypotheses. The selected model's waiting-time law is also the same under both hypotheses and is independent of the fresh output. Thus, the selection, randomization, and waiting-time factors cancel from the likelihood ratio. Iterating the remaining output factors gives

$$\frac{d\mathbb{P}_A|_{\mathcal{G}_t}}{d\mathbb{P}_B|_{\mathcal{G}_t}} = \prod_{s=1}^t \frac{P_{j_s,A}(Y_s)}{P_{j_s,B}(Y_s)} = \exp\left(\sum_{s=1}^t \ell_{j_s}(Y_s)\right) = e^{L_t}.$$

All output probabilities are strictly positive, so these finite-history measures are mutually absolutely continuous. The independent seed has the same law under the two hypotheses and contributes a likelihood ratio of one, including at $t = 0$.

Bayes' rule now yields

$$\mathbb{P}(\theta = A | \mathcal{G}_t) = \frac{\xi_A e^{L_t}}{\xi_A e^{L_t} + \xi_B} = \frac{e^{\delta+L_t}}{1 + e^{\delta+L_t}}, \quad \mathbb{P}(\theta = B | \mathcal{G}_t) = \frac{1}{1 + e^{\delta+L_t}}.$$

The right-hand sides are $\mathcal{H}_t$ -measurable. Taking conditional expectations given $\mathcal{H}_t \subseteq \mathcal{G}_t$ therefore gives the same posterior formulas on $\mathcal{H}_t$, proving (4). This also explains why past response times may affect routing without supplying additional evidence about the ground truth.

Finally, with $H_\alpha := \log((1 - \alpha)/\alpha)$,

$$\begin{aligned} \frac{e^{\delta+L_t}}{1 + e^{\delta+L_t}} &\geq 1 - \alpha \iff L_t \geq H_\alpha - \delta = A_\alpha, \\ \frac{1}{1 + e^{\delta+L_t}} &\geq 1 - \alpha \iff L_t \leq -H_\alpha - \delta = -B_\alpha. \end{aligned}$$

Applying these equivalences at the first crossing proves the stopping and decision rules. They also cover $t = 0$ when the prior already meets the confidence requirement. $\square$

Appendix B: Omitted Proofs in Section 3

Define

$$\bar{\alpha} := \frac{1}{1 + e^4}, \quad H_\alpha := \log \frac{1 - \alpha}{\alpha}, \quad M_\alpha := \lfloor \sqrt{H_\alpha} \rfloor, \quad \lambda_\alpha := \frac{H_\alpha}{M_\alpha}.$$

In the following, we fix $\alpha \in (0, \bar{\alpha}]$. Because $H_\alpha \geq 4$, we have $M_\alpha \geq 2$ and $\lambda_\alpha \geq \sqrt{H_\alpha} \geq 2$. We take a symmetric prior $\xi_A = \xi_B = 1/2$, so that $\delta = 0$ and $A_\alpha = B_\alpha = H_\alpha$. We also set $c_j = 1$ and $T_{j,t} \equiv 1$ for all $j \in [M_\alpha]$, and $g(w) = w$. Hence, for every admissible policy $\pi \in \Pi(\alpha)$,

$$\mathcal{R}(\pi; \alpha) = \mathbb{E}[C_\pi] + \mathbb{E}[W_\pi] = 2\mathbb{E}[\tau].$$

Thus minimizing $\mathcal{R}(\pi; \alpha)$ is equivalent to minimizing $\mathbb{E}[\tau]$.

For each $j \in [M_\alpha]$, define

$$\gamma_{j,A} := \frac{1 - e^{-(M_\alpha+j-1)\lambda_\alpha}}{1 - e^{-(M_\alpha+j)\lambda_\alpha}}, \quad \gamma_{j,B} := \frac{1 - e^{-\lambda_\alpha}}{1 - e^{-(M_\alpha+j)\lambda_\alpha}}.$$

We also define the chain states

$$x_i := i\lambda_\alpha, \quad i = 0, 1, \dots, M_\alpha.$$

Then $x_0 = 0$ and $x_{M_\alpha} = H_\alpha = A_\alpha$.

For a continuation state $L \in (-H_\alpha, H_\alpha)$, define

$$p(L) := \frac{e^L}{1 + e^L},$$

the posterior probability of A under the symmetric prior. Let $V(L)$ be the minimum expected remaining number of queries when the current LLR state is L , and set $V(L) := 0$ whenever $|L| \geq H_\alpha$. Since each query has unit cost and $g(w) = w$, the Bellman equation is

$$V(L) = \min_{j \in [M_\alpha]} Q(L, j), \quad |L| < H_\alpha,$$

where

$$Q(L, j) := 1 + q_j(L)V(L + \ell_j(A)) + (1 - q_j(L))V(L + \ell_j(B)),$$

and

$$q_j(L) := p(L)\gamma_{j,A} + (1 - p(L))(1 - \gamma_{j,B}).$$

B.1. Supporting Results

LEMMA B.1. *For every $j \in [M_\alpha]$, the primitive parameters satisfy*

$$\gamma_{j,A}, \gamma_{j,B} \in (1/2, 1),$$

and the induced LLR increments satisfy

$$\ell_j(A) = \lambda_\alpha, \quad \ell_j(B) = -(M_\alpha + j - 1)\lambda_\alpha.$$

Proof of Lemma B.1 Fix $j \in [M_\alpha]$ and write

$$D_j := (M_\alpha + j - 1)\lambda_\alpha, \quad U_j := \lambda_\alpha.$$

Because $\lambda_\alpha \geq 2$, we have $U_j \geq 2$ and $D_j \geq 2$. Since $0 < 1 - e^{-(U_j + D_j)} < 1$,

$$\gamma_{j,A} = \frac{1 - e^{-D_j}}{1 - e^{-(U_j + D_j)}} > 1 - e^{-D_j} \geq 1 - e^{-2} > \frac{1}{2}.$$

Similarly,

$$\gamma_{j,B} = \frac{1 - e^{-U_j}}{1 - e^{-(U_j + D_j)}} > 1 - e^{-U_j} \geq 1 - e^{-2} > \frac{1}{2}.$$

Also, $D_j < U_j + D_j$ and $U_j < U_j + D_j$, so

$$1 - e^{-D_j} < 1 - e^{-(U_j + D_j)}, \quad 1 - e^{-U_j} < 1 - e^{-(U_j + D_j)},$$

which implies $\gamma_{j,A} < 1$ and $\gamma_{j,B} < 1$.

Next, we have

$$\frac{\gamma_{j,A}}{1 - \gamma_{j,B}} = \frac{\frac{1 - e^{-D_j}}{1 - e^{-(U_j + D_j)}}}{\frac{e^{-U_j}(1 - e^{-D_j})}{1 - e^{-(U_j + D_j)}}} = e^{U_j}, \quad \frac{1 - \gamma_{j,A}}{\gamma_{j,B}} = \frac{\frac{e^{-D_j}(1 - e^{-U_j})}{1 - e^{-(U_j + D_j)}}}{\frac{1 - e^{-U_j}}{1 - e^{-(U_j + D_j)}}} = e^{-D_j},$$

which imply $\ell_j(A) = U_j = \lambda_\alpha$ and $\ell_j(B) = -D_j = -(M_\alpha + j - 1)\lambda_\alpha$. □

LEMMA B.2. For every $L \in (-H_\alpha, H_\alpha)$ and every $j \in [M_\alpha]$, we have $1 - \gamma_{j,B} = e^{-\lambda_\alpha} \gamma_{j,A}$ and hence

$$q_j(L) = (p(L) + e^{-\lambda_\alpha}(1 - p(L)))\gamma_{j,A}.$$

In particular, for each fixed L , the map $j \mapsto q_j(L)$ is strictly increasing.

Proof of Lemma B.2 The identity $1 - \gamma_{j,B} = e^{-\lambda_\alpha} \gamma_{j,A}$ follows directly from the definitions. Substituting into the formula for $q_j(L)$ gives

$$q_j(L) = p(L)\gamma_{j,A} + (1 - p(L))e^{-\lambda_\alpha}\gamma_{j,A} = (p(L) + e^{-\lambda_\alpha}(1 - p(L)))\gamma_{j,A}.$$

Thus it remains to show that $\gamma_{j,A}$ is strictly increasing in j . Define

$$h(t) := \frac{1 - e^{-t}}{1 - e^{-(t+\lambda_\alpha)}}, \quad t > 0.$$

Then $\gamma_{j,A} = h((M_\alpha + j - 1)\lambda_\alpha)$. Let $a := e^{-\lambda_\alpha} \in (0, 1)$ and $u := e^{-t} \in (0, 1)$. Then, we have

$$h(t) = \frac{1 - u}{1 - au}, \quad \text{and} \quad \frac{d}{du} \left(\frac{1 - u}{1 - au} \right) = -\frac{1 - a}{(1 - au)^2} < 0.$$

Since $u = e^{-t}$ is strictly decreasing in t , $h(t)$ is strictly increasing in t . Therefore $\gamma_{j,A}$, and hence $q_j(L)$, is strictly increasing in j . $\square$

LEMMA B.3. Fix $i \in \{0, 1, \dots, M_\alpha - 1\}$ and $j \in [M_\alpha]$.

(i) The good successor state from x_i under action j is

$$x_i + \ell_j(A) = x_{i+1}.$$

(ii) If $j \geq i + 1$, then the bad successor state exits the continuation region:

$$x_i + \ell_j(B) \leq -H_\alpha.$$

(iii) If $j \leq i$, then the bad successor state remains in the continuation region and is negative:

$$x_i + \ell_j(B) \in (-H_\alpha, 0).$$

Proof of Lemma B.3 Part (i) follows from $\ell_j(A) = \lambda_\alpha$. If $j \geq i + 1$, then

$$x_i + \ell_j(B) = i\lambda_\alpha - (M_\alpha + j - 1)\lambda_\alpha = -(M_\alpha + j - 1 - i)\lambda_\alpha \leq -M_\alpha\lambda_\alpha = -H_\alpha,$$

which proves (ii). If $j \leq i$, then

$$x_i + \ell_j(B) = -(M_\alpha + j - 1 - i)\lambda_\alpha < 0.$$

Moreover, because $j \geq 1$ and $i \leq M_\alpha - 1$,

$$M_\alpha + j - 1 - i \leq M_\alpha - 1,$$

so

$$x_i + \ell_j(B) = -(M_\alpha + j - 1 - i)\lambda_\alpha \geq -(M_\alpha - 1)\lambda_\alpha > -M_\alpha\lambda_\alpha = -H_\alpha.$$

This proves (iii). $\square$

LEMMA B.4. *For each $i \in \{0, 1, \dots, M_\alpha\}$, we have $V(x_i) \leq M_\alpha - i$.*

Proof of Lemma B.4 The claim is immediate for $i = M_\alpha$ because $x_{M_\alpha} = H_\alpha$ is a stopping state and $V(x_{M_\alpha}) = 0$. Now fix $i \in \{0, 1, \dots, M_\alpha - 1\}$. Consider the policy that queries LLM $i + 1$ first, then LLM $i + 2$ if the first outcome is A , then LLM $i + 3$ if the first two outcomes are both A , and so on.

By Lemma B.3(i), each A outcome moves the state from x_r to x_{r+1} . By Lemma B.3(ii), a B outcome from state x_r under action $r + 1$ exits immediately. Therefore this policy uses at most one query at each of the states $x_i, x_{i+1}, \dots, x_{M_\alpha-1}$, so it stops after at most $M_\alpha - i$ queries pathwise. Hence, we have $V(x_i) \leq M_\alpha - i$.

□

PROPOSITION B.1. *If $\alpha \in (0, \bar{\alpha}]$, then for every $i \in \{0, 1, \dots, M_\alpha - 1\}$, the unique Bellman-optimal action at state x_i is LLM $i + 1$, i.e., $Q(x_i, i + 1) < Q(x_i, j)$ for every $j \neq i + 1$.*

Proof of Proposition B.1 Fix $i \in \{0, 1, \dots, M_\alpha - 1\}$. We consider two cases.

Case 1: comparison with later actions $j > i + 1$.

Fix such a j . By Lemma B.3(i) and (ii), both actions $i + 1$ and j have the same good successor x_{i+1} and both make the bad successor stop immediately. Hence, we can derive

$$Q(x_i, i + 1) = 1 + q_{i+1}(x_i)V(x_{i+1}), \quad Q(x_i, j) = 1 + q_j(x_i)V(x_{i+1}).$$

Because $j \geq i + 2$, necessarily $i \leq M_\alpha - 2$, so $x_{i+1} < H_\alpha$ and therefore $V(x_{i+1}) \geq 1$. By Lemma B.2, $q_j(x_i) > q_{i+1}(x_i)$, and thus

$$Q(x_i, j) - Q(x_i, i + 1) = (q_j(x_i) - q_{i+1}(x_i))V(x_{i+1}) > 0.$$

Case 2: comparison with earlier actions $j \leq i$.

Fix such a j . By Lemma B.3(i) and (iii), we have $x_i + \ell_j(A) = x_{i+1}$ and $x_i + \ell_j(B) \in (-H_\alpha, 0)$, which implies

$$Q(x_i, j) = 1 + q_j(x_i)V(x_{i+1}) + (1 - q_j(x_i))V(x_i + \ell_j(B)), \quad Q(x_i, i + 1) = 1 + q_{i+1}(x_i)V(x_{i+1}).$$

Subtracting the above equations yields

$$Q(x_i, j) - Q(x_i, i + 1) = (q_j(x_i) - q_{i+1}(x_i))V(x_{i+1}) + (1 - q_j(x_i))V(x_i + \ell_j(B)).$$

Since $x_i + \ell_j(B)$ is a continuation state, at least one further query is needed before stopping, i.e., $V(x_i + \ell_j(B)) \geq 1$. Hence, we have

$$Q(x_i, j) - Q(x_i, i + 1) \geq -(q_{i+1}(x_i) - q_j(x_i))V(x_{i+1}) + (1 - q_j(x_i)). \quad (\text{B.1})$$

By Lemma B.4, we have $V(x_{i+1}) \leq M_\alpha - i - 1 \leq M_\alpha - 1$.

We now bound the two terms on the right-hand side.

First, by Lemma B.2, we have

$$q_{i+1}(x_i) - q_j(x_i) = (p(x_i) + e^{-\lambda\alpha}(1 - p(x_i))) (\gamma_{i+1,A} - \gamma_{j,A}) \leq \gamma_{i+1,A} - \gamma_{j,A}.$$

Because $\gamma_{j,A} \leq \gamma_{i+1,A} \leq 1$, we have

$$\gamma_{i+1,A} - \gamma_{j,A} \leq 1 - \gamma_{j,A} = \frac{e^{-(M_\alpha+j-1)\lambda_\alpha}(1 - e^{-\lambda_\alpha})}{1 - e^{-(M_\alpha+j)\lambda_\alpha}} \leq e^{-(M_\alpha+j-1)\lambda_\alpha} \leq e^{-M_\alpha\lambda_\alpha} = e^{-H_\alpha}.$$

which implies that $q_{i+1}(x_i) - q_j(x_i) \leq e^{-H_\alpha}$.

Second, again by Lemma B.2, we have

$$1 - q_j(x_i) > (1 - p(x_i))(1 - e^{-\lambda_\alpha}).$$

Since $x_i = i\lambda_\alpha \leq (M_\alpha - 1)\lambda_\alpha$, we can derive

$$1 - p(x_i) = \frac{1}{1 + e^{x_i}} \geq \frac{1}{1 + e^{(M_\alpha-1)\lambda_\alpha}}.$$

Because $\lambda_\alpha \geq 1$ and $1 + e^{(M_\alpha-1)\lambda_\alpha} \leq 2e^{(M_\alpha-1)\lambda_\alpha}$, it follows that

$$1 - q_j(x_i) > \frac{1 - e^{-\lambda_\alpha}}{1 + e^{(M_\alpha-1)\lambda_\alpha}} \geq \frac{1 - e^{-1}}{2} e^{-(M_\alpha-1)\lambda_\alpha}.$$

Substituting these estimates into (B.1) gives

$$Q(x_i, j) - Q(x_i, i+1) > -(M_\alpha - 1)e^{-H_\alpha} + \frac{1 - e^{-1}}{2} e^{-(M_\alpha-1)\lambda_\alpha}.$$

Factoring out $e^{-(M_\alpha-1)\lambda_\alpha}$ and using $H_\alpha = M_\alpha\lambda_\alpha$, we obtain

$$Q(x_i, j) - Q(x_i, i+1) > e^{-(M_\alpha-1)\lambda_\alpha} \left(\frac{1 - e^{-1}}{2} - (M_\alpha - 1)e^{-\lambda_\alpha} \right).$$

The quantity in parentheses is strictly positive when $\alpha \in (0, \bar{\alpha}]$, so $Q(x_i, j) > Q(x_i, i+1)$.

Combining Cases 1 and 2 shows that action $i+1$ is the unique Bellman-optimal action at state x_i . $\square$

B.2. Proof of Theorem 1

By Proposition B.1, for each $i = 0, 1, \dots, M_\alpha - 1$, the unique Bellman-optimal action at state x_i is LLM $i+1$.

Now consider the event

$$E := \{Y_1 = A, Y_2 = A, \dots, Y_{M_\alpha} = A\}.$$

On E , the LLR evolves as

$$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_{M_\alpha} = H_\alpha = A_\alpha,$$

so every exact optimal policy must query LLMs $1, 2, \dots, M_\alpha$ in order before stopping. Therefore, we have

$$E \subseteq \{N_j(\tau) \geq 1 \text{ for all } j \in [M_\alpha]\},$$

which implies

$$\mathbb{P}_A(N_j(\tau) \geq 1 \text{ for all } j \in [M_\alpha]) \geq \mathbb{P}_A(E).$$

Under $\mathbb{P}_A$, whenever LLM j is queried, the probability of outcome A is $\gamma_{j,A}$, so

$$\mathbb{P}_A(E) = \prod_{j=1}^{M_\alpha} \gamma_{j,A}.$$

Moreover, we can derive

$$1 - \gamma_{j,A} = \frac{e^{-(M_\alpha+j-1)\lambda_\alpha}(1 - e^{-\lambda_\alpha})}{1 - e^{-(M_\alpha+j)\lambda_\alpha}} \leq e^{-(M_\alpha+j-1)\lambda_\alpha} \leq e^{-M_\alpha\lambda_\alpha} = e^{-H_\alpha} = \frac{\alpha}{1 - \alpha}.$$

Using the union bound, we have

$$1 - \prod_{j=1}^{M_\alpha} \gamma_{j,A} \leq \sum_{j=1}^{M_\alpha} (1 - \gamma_{j,A}) \leq M_\alpha \frac{\alpha}{1 - \alpha}.$$

Therefore, we can derive

$$\mathbb{P}_A(N_j(\tau) \geq 1 \text{ for all } j \in [M_\alpha]) \geq \mathbb{P}_A(E) \geq 1 - M_\alpha \frac{\alpha}{1 - \alpha}.$$

Finally, we have $\alpha M_\alpha / (1 - \alpha) \rightarrow 0$ as $\alpha \rightarrow 0$ because $M_\alpha = \lfloor \sqrt{\log((1 - \alpha)/\alpha)} \rfloor$ and $\sqrt{\log(1/\alpha)} \cdot \alpha \rightarrow 0$. $\square$

Appendix C: Omitted Proofs in Section 4

C.1. Supporting Results

Write

$$h_\alpha := \log(1/\alpha), \quad H_\alpha := \log((1 - \alpha)/\alpha), \quad I_{\min, \theta} := \min_{j \in [M]} I_{j, \theta} > 0.$$

We work in the nontrivial regime $\alpha < \min\{\xi_A, \xi_B\}$, so both stopping boundaries are positive. Define

$$C_\ell := \max_{j \in [M]} \max\{|\ell_j(A)|, |\ell_j(B)|\} < \infty, \quad (\text{C.2})$$

$$t_\alpha := \frac{\min\{A_\alpha, B_\alpha\}}{C_\ell} < \infty, \quad (\text{C.3})$$

$$v_\ell^2 := \max_{j \in [M]} \max\{\text{Var}_A(\ell_j(Y)), \text{Var}_B(\ell_j(Y))\} < \infty. \quad (\text{C.4})$$

Finiteness follows from $M < \infty$ and strictly interior accuracies. When a deterministic-time process is used after the policy stops, extend its LLM-selection rule nonanticipatively beyond τ ; this extension has no effect on stopped quantities.

C.1.1. Likelihood Ratio Martingale and Change of Measure

Let $\Lambda_t := e^{L_t}$, with $\Lambda_0 = 1$. In this subsection, we provide some preliminary properties of Λ_t .

LEMMA C.5 (Likelihood ratio martingale). *Under $\mathbb{P}_B$, $(\Lambda_t)_{t \geq 0}$ is a nonnegative $(\mathcal{G}_t)$ -martingale with*

$$\mathbb{E}_B[\Lambda_t | \mathcal{G}_{t-1}] = \Lambda_{t-1} \quad (t \geq 1), \quad \mathbb{E}_B[\Lambda_t] = 1 \quad (t \geq 0).$$

Under $\mathbb{P}_A$, $(\Lambda_t^{-1})_{t \geq 0}$ is a nonnegative martingale with mean one.

Proof of Lemma C.5. Fix $t \geq 1$ and work under $\mathbb{P}_B$. Conditional on $\theta = B$ and $\mathcal{G}_{t-1}$, we have

$$\begin{aligned} \mathbb{E}_B[\Lambda_t | \mathcal{G}_{t-1}] &= \mathbb{E}_B[e^{L_{t-1} + \ell_{j_t}(Y_t)} | \mathcal{G}_{t-1}] = e^{L_{t-1}} \sum_{y \in \{A, B\}} \mathbb{P}_B(Y_t = y | \mathcal{G}_{t-1}) e^{\ell_{j_t}(y)} \\ &= e^{L_{t-1}} \sum_{y \in \{A, B\}} \mathbb{P}(Y = y | \theta = A, j_t) = e^{L_{t-1}} = \Lambda_{t-1}, \end{aligned}$$

where the third equality is because $e^{\ell_{j_t}(y)} = \mathbb{P}(Y = y | \theta = A, j_t) / \mathbb{P}(Y = y | \theta = B, j_t)$. Thus $(\Lambda_t)_{t \geq 0}$ is a martingale under $\mathbb{P}_B$ with $\mathbb{E}_B[\Lambda_0] = 1$. By the optional stopping theorem (see Theorem 4.8.6 in Durrett 2019), we have $\mathbb{E}_B[\Lambda_t] = 1$ for all t . The argument under $\mathbb{P}_A$ is analogous: for $Z_t := \Lambda_t^{-1} = e^{-L_t}$ we have

$$\mathbb{E}_A[Z_t | \mathcal{G}_{t-1}] = Z_{t-1}, \quad \mathbb{E}_A[Z_t] = 1. \quad \square$$

LEMMA C.6 (**Change-of-measure identity at stopping times**). *Let τ be any almost surely finite stopping time with respect to $(\mathcal{G}_t)_{t \geq 0}$. Then for every bounded $\mathcal{G}_\tau$ -measurable random variable X ,*

$$\mathbb{E}_A[X] = \mathbb{E}_B[\Lambda_\tau X], \quad \mathbb{E}_B[X] = \mathbb{E}_A[\Lambda_\tau^{-1} X].$$

Furthermore, for every event $F \in \mathcal{G}_\tau$,

$$\mathbb{P}_A(F) = \mathbb{E}_A[\mathbf{1}_F] = \mathbb{E}_B[\Lambda_\tau \mathbf{1}_F], \quad \mathbb{P}_B(F) = \mathbb{E}_B[\mathbf{1}_F] = \mathbb{E}_A[\Lambda_\tau^{-1} \mathbf{1}_F].$$

Proof of Lemma C.6. For every deterministic t , $\Lambda_t = e^{L_t}$ is the likelihood ratio of $\mathbb{P}_A$ with respect to $\mathbb{P}_B$ on $\mathcal{G}_t$. Hence, for every bounded $\mathcal{G}_t$ -measurable random variable H , $\mathbb{E}_A[H] = \mathbb{E}_B[\Lambda_t H]$. Now let X be bounded and $\mathcal{G}_\tau$ -measurable. Since $X \mathbf{1}_{\{\tau=t\}}$ is $\mathcal{G}_t$ -measurable, the deterministic-time identity gives

$$\mathbb{E}_A[X \mathbf{1}_{\{\tau=t\}}] = \mathbb{E}_B[\Lambda_t X \mathbf{1}_{\{\tau=t\}}].$$

Summing over $t \geq 0$ and using the almost-sure finiteness of τ ,

$$\mathbb{E}_A[X] = \sum_{t=0}^{\infty} \mathbb{E}_A[X \mathbf{1}_{\{\tau=t\}}] = \sum_{t=0}^{\infty} \mathbb{E}_B[\Lambda_t X \mathbf{1}_{\{\tau=t\}}] = \mathbb{E}_B[\Lambda_\tau X].$$

The interchange is justified since X is bounded; equivalently, one may apply the argument separately to X^+ and X^- . The second identity is proved symmetrically using Λ_t^{-1} under $\mathbb{P}_A$. Taking $X = \mathbf{1}_F$ gives the event identities. $\square$

C.1.2. Error Probability Bounds

In this subsection, we show that the probability of concluding a wrong answer is upper bounded.

LEMMA C.7 (**Exponential error bounds**). *For the posterior-threshold rule,*

$$\mathbb{P}_A(D = B) \leq e^{-B_\alpha} \leq c_{\text{err}} \alpha, \quad \mathbb{P}_B(D = A) \leq e^{-A_\alpha} \leq c_{\text{err}} \alpha,$$

where $c_{\text{err}} := 2 \max\{\xi_A/\xi_B, \xi_B/\xi_A\}$ is independent of α and the policy.

Proof of Lemma C.7. We prove the first inequality; the second is analogous. Under $\theta = A$, the error event $\{D = B\}$ occurs if and only if $L_\tau \leq -B_\alpha$. By the change-of-measure identity (Lemma C.6),

$$\mathbb{P}_A(D = B) = \mathbb{P}_A(L_\tau \leq -B_\alpha) = \mathbb{E}_B[\Lambda_\tau \mathbf{1}_{\{L_\tau \leq -B_\alpha\}}].$$

On the event $\{L_\tau \leq -B_\alpha\}$, we have $\Lambda_\tau = e^{L_\tau} \leq e^{-B_\alpha}$, so

$$\mathbb{P}_A(D = B) \leq e^{-B_\alpha} \mathbb{E}_B[\mathbf{1}_{\{L_\tau \leq -B_\alpha\}}] \leq e^{-B_\alpha} \mathbb{E}_B[1] = e^{-B_\alpha} = \frac{\alpha}{1 - \alpha} \frac{\xi_B}{\xi_A} \leq c_{\text{err}} \cdot \alpha,$$

where $c_{\text{err}} = 2 \cdot \max\{\xi_A/\xi_B, \xi_B/\xi_A\}$ is a constant independent of α . Similarly, we can deduce that $\mathbb{P}_B(D = A) \leq e^{-A_\alpha} \leq c_{\text{err}} \cdot \alpha$. $\square$

C.1.3. Martingale Decomposition of Log-Likelihood Ratio

In this subsection, we show that $M_t^{(\theta)}$ in (12) is a martingale, which will be used to establish concentration results. Recall that, by (5), $N_j(t)$ denotes the total number of queries sent to LLM j until period t , $N_j(t) := \sum_{\ell=1}^t \mathbf{1}_{\{j_\ell=j\}}$.

LEMMA C.8 (Drift-martingale decomposition). *For each $t \geq 0$,*

$$L_t = \sum_{s=1}^t I_{j_s, A} + M_t^{(A)} = \sum_{j=1}^M I_{j, A} N_j(t) + M_t^{(A)},$$

where $(M_t^{(A)})_{t \geq 0}$ is a $(\mathcal{G}_t)$ -martingale under $\mathbb{P}_A$ with $\mathbb{E}_A[M_t^{(A)}] = 0$ and bounded increments $|M_t^{(A)} - M_{t-1}^{(A)}| \leq 2C_\ell$. Similarly,

$$-L_t = \sum_{s=1}^t I_{j_s, B} + M_t^{(B)} = \sum_{j=1}^M I_{j, B} N_j(t) + M_t^{(B)},$$

where $(M_t^{(B)})_{t \geq 0}$ is a $(\mathcal{G}_t)$ -martingale under $\mathbb{P}_B$ with $\mathbb{E}_B[M_t^{(B)}] = 0$ and bounded increments $|M_t^{(B)} - M_{t-1}^{(B)}| \leq 2C_\ell$.

Proof of Lemma C.8. We prove the claim under $\mathbb{P}_A$; the argument under $\mathbb{P}_B$ is identical with signs adjusted. Define

$$D_s^{(A)} := \ell_{j_s}(Y_s) - I_{j_s, A}, \quad M_t^{(A)} := \sum_{s=1}^t D_s^{(A)}, \quad t \geq 0,$$

with $M_0^{(A)} := 0$. Then, we have

$$L_t = \sum_{s=1}^t \ell_{j_s}(Y_s) = \sum_{s=1}^t I_{j_s, A} + M_t^{(A)}.$$

Under $\mathbb{P}_A$, we have $\mathbb{E}_A[\ell_{j_s}(Y_s) \mid \mathcal{G}_{s-1}] = I_{j_s, A}$, so

$$\mathbb{E}_A[D_s^{(A)} \mid \mathcal{G}_{s-1}] = \mathbb{E}_A[\ell_{j_s}(Y_s) - I_{j_s, A} \mid \mathcal{G}_{s-1}] = 0.$$

Hence $(M_t^{(A)})_{t \geq 0}$ is a $(\mathcal{G}_t)$ -martingale under $\mathbb{P}_A$. The martingale $(M_t^{(A)})$ has bounded increments $|M_t^{(A)} - M_{t-1}^{(A)}|$ because

$$|M_t^{(A)} - M_{t-1}^{(A)}| \leq |L_t - L_{t-1}| + |I_{j_t, A}| \leq 2C_\ell \quad \text{almost surely for all } t \geq 1.$$

The identities for expectations follow immediately by taking $\mathbb{E}_A[\cdot]$ on both sides and using $\mathbb{E}_A[M_t^{(A)}] = 0$. As for the equality, it suffices to note the following equation:

$$\sum_{s=1}^t I_{j_s, A} = \sum_{j=1}^M I_{j, A} \sum_{s=1}^t \mathbf{1}_{\{j_s=j\}} = \sum_{j=1}^M I_{j, A} N_j(t).$$

□

C.1.4. Concentration of Log-Likelihood Ratio

In this subsection, we present concentration results for the cumulative log-likelihood ratio L_t . We first state the version of Freedman's inequality that we will use.

LEMMA C.9 (Freedman's inequality; Theorem 1.6 in Freedman 1975). *Let $(M_t)_{t \geq 0}$ be a real-valued martingale with respect to a filtration $(\mathcal{F}_t)_{t \geq 0}$, with $M_0 = 0$, and define the martingale differences $X_t := M_t - M_{t-1}$ for all $t \geq 1$. Assume that there exists a constant $b > 0$ such that $|X_t| \leq b$ almost surely for all $t \geq 1$. Define the predictable quadratic variation $V_t := \sum_{s=1}^t \mathbb{E}[X_s^2 | \mathcal{F}_{s-1}]$ with the convention $V_0 := 0$. Then for every $t \geq 1$, every $x \geq 0$, and every deterministic constant $v \geq 0$ such that $V_t \leq v$, we have*

$$\mathbb{P}(|M_t| \geq x) \leq 2 \exp\left(-\frac{x^2}{2(v + bx)}\right), \quad x \geq 0.$$

Then, we prove the concentration inequalities.

LEMMA C.10 (Freedman-type concentration for the LLR martingales). *Under Assumptions 1-3, for each $\theta \in \{A, B\}$, each integer $t \geq 1$, and each $x > 0$,*

$$\mathbb{P}_\theta(|M_t^{(\theta)}| \geq x) \leq 2 \exp\left(-\frac{x^2}{2(v_\ell^2 t + 2C_\ell x)}\right).$$

Consequently, under $\mathbb{P}_A$,

$$\mathbb{P}_A(|L_t - \sum_{s=1}^t I_{j_s, A}| \geq x) \leq 2 \exp\left(-\frac{x^2}{2(v_\ell^2 t + 2C_\ell x)}\right),$$

and under $\mathbb{P}_B$,

$$\mathbb{P}_B(|-L_t - \sum_{s=1}^t I_{j_s, B}| \geq x) \leq 2 \exp\left(-\frac{x^2}{2(v_\ell^2 t + 2C_\ell x)}\right).$$

Proof of Lemma C.10. Fix $\theta \in \{A, B\}$ and $t \geq 1$. From Lemma C.8, we know that $(M_s^{(\theta)})_{s \geq 0}$ is a $(\mathcal{G}_s)$ -martingale under $\mathbb{P}_\theta$ with $M_0^{(\theta)} = 0$. Then, we apply Lemma C.9 to the martingale $(M_s^{(\theta)})_{s=0}^t$ with respect to the filtration $(\mathcal{G}_s)_{s=0}^t$ under the probability measure $\mathbb{P}_\theta$. The martingale difference $D_s^{(\theta)} = M_s^{(\theta)} - M_{s-1}^{(\theta)}$ is defined in Lemma C.8. From Lemma C.8, we have $|D_s^{(\theta)}| \leq 2C_\ell$ almost surely for all $s \geq 1$. Thus the increment bound in Lemma C.9 holds with $b = 2C_\ell$.

The predictable quadratic variation is, by (C.4),

$$V_t^{(\theta)} = \sum_{s=1}^t \mathbb{E}_\theta[(D_s^{(\theta)})^2 | \mathcal{G}_{s-1}] \leq v_\ell^2 t \quad \text{almost surely.}$$

By Lemma C.9 with the choices $b = 2C_\ell$ and $v = v_\ell^2 t$, for each $x \geq 0$,

$$\mathbb{P}_\theta(|M_t^{(\theta)}| \geq x) \leq 2 \exp\left(-\frac{x^2}{2(v_\ell^2 t + 2C_\ell x)}\right).$$

Under $\mathbb{P}_A$, the decomposition (12) gives

$$\mathbb{P}_A(|L_t - \sum_{s=1}^t I_{j_s, A}| \geq x) = \mathbb{P}_A(|M_t^{(A)}| \geq x) \leq 2 \exp\left(-\frac{x^2}{2(v_\ell^2 t + 2C_\ell x)}\right). \quad \square$$

C.1.5. Overshoot and Expected LLR at Stopping

In this subsection, we investigate the relationship between L_τ and (A_α, B_α) . When the stopping time τ is reached, the LLR L_τ typically overshoots the boundary by a small amount. The following lemma provides a uniform bound on this overshoot.

LEMMA C.11 (**Uniform overshoot bound**). *For the posterior-threshold stopping rule, the overshoot is uniformly bounded:*

$$|L_\tau - A_\alpha| \leq C_\ell \quad \text{conditioning on } \{L_\tau \geq A_\alpha\},$$

and

$$|L_\tau + B_\alpha| \leq C_\ell \quad \text{conditioning on } \{L_\tau \leq -B_\alpha\},$$

where C_ℓ is the constant in (C.2).

Proof of Lemma C.11. By definition of τ , we have $|L_t| < \max\{A_\alpha, B_\alpha\}$ for all $t < \tau$, and at time τ the process L_τ first crosses one of the boundaries. On the event $\{L_\tau \geq A_\alpha\}$, we must have $L_{\tau-1} < A_\alpha$ (otherwise τ would have occurred earlier), and

$$L_\tau = L_{\tau-1} + \ell_{j_\tau}(Y_\tau) < A_\alpha + C_\ell,$$

by (C.2). Since $L_\tau \geq A_\alpha$ by assumption, we have

$$A_\alpha \leq L_\tau < A_\alpha + C_\ell,$$

which gives $|L_\tau - A_\alpha| < C_\ell$. The bound on the other event is proved identically. $\square$

The next lemma gives upper and lower bounds on the expected cumulative LLR at stopping under each hypothesis.

LEMMA C.12 (**Expected LLR at stopping**). *Suppose the policy π satisfies $\mathbb{E}_A[\tau] < \infty$ and $\mathbb{E}_B[\tau] < \infty$. Then for all sufficiently small $\alpha > 0$,*

$$A_\alpha - c_{\text{err}}\alpha \cdot (A_\alpha + B_\alpha + C_\ell) \leq \mathbb{E}_A[L_\tau] \leq A_\alpha + C_\ell, \quad (\text{C.5})$$

$$B_\alpha - c_{\text{err}}\alpha \cdot (A_\alpha + B_\alpha + C_\ell) \leq \mathbb{E}_B[-L_\tau] \leq B_\alpha + C_\ell. \quad (\text{C.6})$$

In particular, $\mathbb{E}_A[L_\tau] = A_\alpha + O(1)$ and $\mathbb{E}_B[-L_\tau] = B_\alpha + O(1)$ as $\alpha \rightarrow 0$.

Proof of Lemma C.12. We prove (C.5); the proof of (C.6) is analogous. When $\theta = A$, we can decompose the expectation by the decision outcome:

$$\mathbb{E}_A[L_\tau] = \mathbb{E}_A[L_\tau \mathbb{1}_{\{D=A\}}] + \mathbb{E}_A[L_\tau \mathbb{1}_{\{D=B\}}] = \mathbb{E}_A[L_\tau \mathbb{1}_{\{L_\tau \geq A_\alpha\}}] + \mathbb{E}_A[L_\tau \mathbb{1}_{\{L_\tau \leq -B_\alpha\}}].$$

On the event $\{L_\tau \geq A_\alpha\}$, Lemma C.11 gives $A_\alpha \leq L_\tau \leq A_\alpha + C_\ell$. On the event $\{L_\tau \leq -B_\alpha\}$, we have $-B_\alpha - C_\ell \leq L_\tau \leq -B_\alpha < 0$. Therefore,

$$\begin{aligned} \mathbb{E}_A[L_\tau] &\geq A_\alpha \mathbb{P}_A(L_\tau \geq A_\alpha) + (-B_\alpha - C_\ell) \mathbb{P}_A(L_\tau \leq -B_\alpha) \\ &\geq A_\alpha \mathbb{P}_A(D = A) - (B_\alpha + C_\ell) \mathbb{P}_A(D = B) \\ &= A_\alpha - (A_\alpha + B_\alpha + C_\ell) \mathbb{P}_A(D = B) \end{aligned}$$

By Lemma C.7, we have $\mathbb{P}_A(D = B) \leq c_{\text{err}}\alpha$ and hence

$$\mathbb{E}_A[L_\tau] \geq A_\alpha - c_{\text{err}}\alpha \cdot (A_\alpha + B_\alpha + C_\ell),$$

which establishes the lower bound in (C.5). For the upper bound, note that on $\{L_\tau \geq A_\alpha\}$ we have $L_\tau \leq A_\alpha + C_\ell$ and on $\{L_\tau \leq -B_\alpha\}$ we have $L_\tau \leq 0$. Therefore,

$$\begin{aligned}\mathbb{E}_A[L_\tau] &\leq (A_\alpha + C_\ell) \mathbb{P}_A(L_\tau \geq A_\alpha) + 0 \cdot \mathbb{P}_A(L_\tau \leq -B_\alpha) \\ &\leq A_\alpha + C_\ell.\end{aligned}$$

Combining the bounds gives (C.5). Moreover, since $x \log(1/x) \leq 1/e$, we can deduce that $c_{\text{err}}\alpha(A_\alpha + B_\alpha + C_\ell)$ is upper bounded by a constant. Thus, we have $\mathbb{E}_A[L_\tau] = A_\alpha + O(1)$.

The proof of (C.6) is identical with $-L_\tau$ and B_α in place of L_τ and A_α . $\square$

C.1.6. Properties of Waiting Cost Function

Define $\phi(s) := s^\rho - 1 - \rho(s-1)$ for $s \geq 0$ and $q := \max\{2, \rho\}$. In the following lemma, we provide bounds for the function $\phi(s)$.

LEMMA C.13. *For $\rho \geq 1$, $\phi(s) \geq 0$. There is a finite $c_1 > 0$ such that*

$$\phi(s) \leq c_1((s-1)^2 + |s-1|^q), \quad s \geq 0.$$

If $\rho > 1$, there is $c_2 > 0$ such that

$$\phi(s) \geq c_2 \min\{|s-1|, 1/2\}^2, \quad s \geq 0.$$

Proof of Lemma C.13. We start with the first statement. Since $\phi(s)$ is convex and $\phi'(1) = 0$, we have $\phi(s)$ decreases when $s \in [0, 1)$ and increases when $s \in [1, \infty)$. Thus, we can deduce that $\phi(s) \geq \phi(1) = 0$ for $s \geq 0$.

Then we prove the second statement. The inequality holds for any constant when $s = 1$. For $s \in (0, \infty) \setminus \{1\}$, we define the ratio

$$R(s) = \frac{\phi(s)}{(s-1)^2 + |s-1|^q}.$$

Then, we consider several limits as follows:

$$\begin{aligned}\lim_{s \rightarrow 0^+} R(s) &= \frac{\rho-1}{2} \\ \lim_{s \rightarrow 1} R(s) &= \lim_{s \rightarrow 1} \frac{\rho s^{\rho-1} - \rho}{2(s-1) + q|s-1|^{q-1}} = \lim_{s \rightarrow 1} \frac{\rho(\rho-1)s^{\rho-2}}{2 + q(q-1)|s-1|^{q-2}} = \begin{cases} \frac{\rho(\rho-1)}{2} & q > 2 \\ \frac{\rho(\rho-1)}{4} & q = 2 \end{cases} \\ \lim_{s \rightarrow \infty} R(s) &= \begin{cases} 0 & \rho < 2 \\ 1/2 & \rho = 2 \\ 1 & \rho > 2 \end{cases},\end{aligned}$$

which implies that $R(s) < \infty$ for $s \in (0, \infty) \setminus \{1\}$. Since $R(s)$ is continuous on $(0, 1)$ and $(1, \infty)$, there exists a constant c_1 such that the $\phi(s) \leq c_1((s-1)^2 + |s-1|^q)$.

Lastly, if $\rho > 1$, we have $\phi''(s) = \rho(\rho-1)s^{\rho-2}$. By Taylor expansion, for $s \in [1/2, 3/2]$, we have

$$\phi(s) = \phi(1) + \phi'(1)(s-1) + \frac{1}{2}\phi''(s_0)(s-1)^2 = \frac{1}{2}\rho(\rho-1)s_0^{\rho-2}(s-1)^2,$$

where $s_0 \in [1/2, 3/2]$. Thus, there exists a constant $c_2 = \min_{x \in [1/2, 3/2]} \rho(\rho-1)x^{\rho-2}/2$ such that $\phi(s) \geq c_2(s-1)^2$. When $|s-1| > 1/2$, we have $\phi(s) \geq \min\{\phi(1/2), \phi(3/2)\} \geq c_2 \cdot (1/2)^2$. $\square$

C.2. Proof of Proposition 1

We proceed in two steps.

Step 1: Drift identities under $\theta = A$ and $\theta = B$.

By Lemma C.8, under $\mathbb{P}_A$ we can write

$$L_t = \sum_{s=1}^t I_{j_s, A} + M_t^{(A)},$$

where $(M_t^{(A)})_{t \geq 0}$ is a martingale under $\mathbb{P}_A$ with respect to the natural filtration $(\mathcal{G}_t)_{t \geq 0}$, and $I_{j_s, A} = \mathbb{E}_A[\ell_{j_s}(Y)]$ is the conditional expectation of the LLR increment given j_s .

First, we have the following conditions:

- (i) $\mathbb{E}_A[\tau] < \infty$,
- (ii) The increments of $M_t^{(A)}$ are bounded: $|M_t^{(A)} - M_{t-1}^{(A)}| \leq |L_t - L_{t-1}| + |I_{j_t, A}| \leq 2C_\ell$ almost surely for all $t \geq 1$.

Given the above conditions, by the optional stopping theorem (see Theorem 4.8.6 in Durrett 2019),

$$\mathbb{E}_A[M_\tau^{(A)}] = \mathbb{E}_A[M_0^{(A)}] = 0.$$

Taking expectations at the stopping time τ ,

$$\mathbb{E}_A[L_\tau] = \mathbb{E}_A\left[\sum_{s=1}^{\tau} I_{j_s, A} + M_\tau^{(A)}\right] = \mathbb{E}_A\left[\sum_{s=1}^{\tau} I_{j_s, A}\right] + \mathbb{E}_A[M_\tau^{(A)}] = \mathbb{E}_A\left[\sum_{s=1}^{\tau} I_{j_s, A}\right].$$

Rearranging the sum by grouping terms according to which LLM is queried,

$$\sum_{s=1}^{\tau} I_{j_s, A} = \sum_{j=1}^M I_{j, A} \sum_{s=1}^{\tau} \mathbf{1}_{\{j_s=j\}} = \sum_{j=1}^M I_{j, A} N_j(\tau),$$

where $N_j(\tau)$ denotes the total number of queries sent to LLM j before stopping.

To interchange expectation and summation, we verify the conditions for Fubini's theorem. Since M is finite, $I_{j, A} \in (0, C_\ell]$ for all j , and $\mathbb{E}_A[\tau] < \infty$, we have

$$\mathbb{E}_A\left[\sum_{j=1}^M I_{j, A} N_j(\tau)\right] \leq \mathbb{E}_A[C_\ell \cdot \tau] = C_\ell \cdot \mathbb{E}_A[\tau] < \infty.$$

Therefore, by Fubini's theorem, we can interchange expectation and the finite sum:

$$\mathbb{E}_A[L_\tau] = \sum_{j=1}^M I_{j, A} \mathbb{E}_A[N_j(\tau)] = \sum_{j=1}^M I_{j, A} n_{j, A}.$$

An analogous argument under $\mathbb{P}_B$ uses the fact that $-L_t$ has drift $I_{j_s, B}$ under $\theta = B$. Specifically, by Lemma C.8, we can write

$$-L_t = \sum_{s=1}^t I_{j_s, B} + M_t^{(B)},$$

where $(M_t^{(B)})$ is a martingale under $\mathbb{P}_B$ with $M_0^{(B)} = 0$. The same verification as above shows that $\mathbb{E}_B[-L_\tau] = \sum_{j=1}^M I_{j, B} n_{j, B}$.

Step 2: Lower bounds.

According to Lemma C.12, we obtain

$$\mathbb{E}_A[L_\tau] \geq A_\alpha - c_{\text{err}}\alpha(A_\alpha + B_\alpha + C_\ell),$$

Since $x \log \frac{1}{x} \leq 1/e$, we can also deduce that

$$c_{\text{err}}\alpha(A_\alpha + B_\alpha + C_\ell) \leq 2c_{\text{err}}\alpha \log \frac{1}{\alpha} + c_{\text{err}}\alpha C_\ell \leq \frac{2c_{\text{err}}}{e} + c_{\text{err}}C_\ell.$$

Thus, we can set $C_0 = (2c_{\text{err}}/e) + c_{\text{err}}C_\ell$.

By Step 1, $\mathbb{E}_A[L_\tau] = \sum_j I_{j,A} n_{j,A}$, so we have established the result under $\mathbb{P}_A$. A similar argument can induce the second inequality in (7). $\square$

C.3. Proof of Proposition 2

By definition, the total cost is

$$C_\pi = \sum_{t=1}^{\tau} c_{j_t} = \sum_{j=1}^M c_j \sum_{t=1}^{\tau} \mathbb{1}_{\{j_t=j\}} = \sum_{j=1}^M c_j N_j(\tau).$$

Taking expectations under $\mathbb{P}_\theta$ gives

$$\mathbb{E}_\theta[C_\pi] = \sum_{j=1}^M c_j \mathbb{E}_\theta[N_j(\tau)] = \sum_{j=1}^M c_j n_{j,\theta}.$$

For the waiting time, by Assumption 1 and the law of iterated expectations, we have

$$\mathbb{E}_\theta[W_\pi] = \sum_{t \geq 1} \mathbb{E}_\theta[\mathbb{1}\{t \leq \tau\} T_{j_t, t}] = \sum_{t \geq 1} \mathbb{E}_\theta[\mathbb{1}\{t \leq \tau\} \mu_{j_t}] = \sum_{j=1}^M \mu_j n_{j,\theta}. \quad \square$$

C.4. Proof of Proposition 3

We start by proving the following inequality: Under any admissible policy π , we have

$$\mathbb{E}[C_\pi + g(W_\pi)] \geq F(\mathbf{n}_A, \mathbf{n}_B) \quad (\text{C.7})$$

We first verify that $\mathbb{E}[W_\pi] < \infty$. By Proposition 2, we have

$$\mathbb{E}[W_\pi] = \sum_{\theta \in \{A, B\}} \xi_\theta \sum_{j=1}^M \mu_j n_{j,\theta} \leq \mu_{\max} \mathbb{E}[\tau] < \infty.$$

Since $g : [0, \infty) \rightarrow [0, \infty)$ is convex by Assumption 3, and $W_\pi \geq 0$ almost surely (since waiting times are nonnegative), Jensen's inequality for convex functions states that

$$\mathbb{E}[g(W_\pi)] = \xi_A \mathbb{E}_A[g(W_\pi)] + \xi_B \mathbb{E}_B[g(W_\pi)] \geq \xi_A g(\mathbb{E}[W_\pi^{(A)}]) + \xi_B g(\mathbb{E}[W_\pi^{(B)}]).$$

Substituting the expression of $\mathbb{E}[W_\pi]$ in (8) into the above inequality, we have:

$$\mathbb{E}[g(W_\pi)] \geq \xi_A g\left(\sum_{j=1}^M \mu_j n_{j,A}\right) + \xi_B g\left(\sum_{j=1}^M \mu_j n_{j,B}\right).$$

Inequality (C.7) then follows from the above inequality and Proposition 2.

We now construct the following lower bound problem:

$$\begin{aligned}
& \min_{\mathbf{n}_A, \mathbf{n}_B} F(\mathbf{n}_A, \mathbf{n}_B) & (\text{LB}') \\
& \text{s.t.} \quad \sum_{j=1}^M I_{j,A} n_{j,A} \geq S_A(\alpha), \\
& \quad \sum_{j=1}^M I_{j,B} n_{j,B} \geq S_B(\alpha), \\
& \quad n_{j,A}, n_{j,B} \geq 0 \quad \forall j \in [M].
\end{aligned}$$

By Propositions 1 and inequality (C.7), under any feasible policy, the expected numbers of queries $(\mathbf{n}_A, \mathbf{n}_B)$ are feasible to (LB') and its expected cost is greater than that in (LB'). Thus, we can derive that (LB') is a lower bound.

Then, we show that any minimizer of (LB') must satisfy both constraints with equality. Suppose $(\mathbf{n}_A^*, \mathbf{n}_B^*)$ is a minimizer of (LB') subject to (7), but

$$\sum_{j=1}^M I_{j,A} n_{j,A}^* > S_A(\alpha).$$

We will construct a feasible point with strictly smaller objective value, contradicting optimality.

Define the scaling factor

$$\lambda := \frac{S_A(\alpha)}{\sum_{\ell=1}^M I_{\ell,A} n_{\ell,A}^*}.$$

Since $\sum_{\ell=1}^M I_{\ell,A} n_{\ell,A}^* > S_A(\alpha)$ by assumption, we have $\lambda < 1$. Consider the scaled vector $\tilde{n}_{j,A} := \lambda n_{j,A}^*$ for $j = 1, \dots, M$, and keep $\tilde{\mathbf{n}}_B := \mathbf{n}_B^*$ unchanged. Since the objective is non-decreasing in $n_{j,A}$'s, it can be verified that $(\tilde{\mathbf{n}}_A, \mathbf{n}_B^*)$ is feasible and $F(\tilde{\mathbf{n}}_A, \mathbf{n}_B^*) < F(\mathbf{n}_A^*, \mathbf{n}_B^*)$. Therefore, any minimizer must satisfy the first constraint with equality. The same argument applies to the second constraint.

Lastly due to the convexity of g , $F(\mathbf{n}_A, \mathbf{n}_B)$ is also convex in $(\mathbf{n}_A, \mathbf{n}_B)$. □

C.5. Proof of Proposition 4

For (ALO), since $\Delta_M \times \Delta_M$ is compact, there exists a minimizer $(\mathbf{w}_A^*, \mathbf{w}_B^*)$. Define

$$z_A^* := S_A(\alpha) \sum_{j=1}^M \eta_{j,A} \cdot w_{j,A}^*, \quad z_B^* := S_B(\alpha) \sum_{j=1}^M \eta_{j,B} \cdot w_{j,B}^*.$$

KKT conditions rule out multiple actives generically: The problem is convex with affine constraints, so the KKT conditions are necessary and sufficient. Thus, there exist $\lambda_A \in \mathbb{R}$ and $\lambda_B \in \mathbb{R}$ such that for all j , we have

$$\begin{aligned}
& \xi_A S_A(\alpha) \kappa_{j,A} + \xi_A S_A(\alpha) \eta_{j,A} g'(z_A^*) \begin{cases} = \lambda_A, & \text{if } w_{j,A}^* > 0, \\ \geq \lambda_A, & \text{if } w_{j,A}^* = 0, \end{cases} \\
& \xi_B S_B(\alpha) \kappa_{j,B} + \xi_B S_B(\alpha) \eta_{j,B} g'(z_B^*) \begin{cases} = \lambda_B, & \text{if } w_{j,B}^* > 0, \\ \geq \lambda_B, & \text{if } w_{j,B}^* = 0. \end{cases}
\end{aligned} \tag{C.8}$$

In the following, we prove that there exists an optimal solution with $\mathbf{w}_A^* = \mathbf{e}_{j^*}$ and $\mathbf{w}_B^* = \mathbf{e}_{k^*}$ for some j^* and k^* . We present the proof for $\mathbf{w}_A^* = \mathbf{e}_{j^*}$ in the following and the proof for $\mathbf{w}_B^* = \mathbf{e}_{k^*}$ is similar.

Suppose two distinct indices $j \neq k$ are active ($w_{j,A}^*, w_{k,A}^* > 0$). Subtracting the equalities in (C.8) gives

$$(\kappa_{j,A} - \kappa_{k,A}) + g'(z_A^*)(\eta_{j,A} - \eta_{k,A}) = 0. \quad (\text{C.9})$$

Case 1: $\eta_{j,A} = \eta_{k,A}$. Then (C.9) becomes:

$$\kappa_{j,A} - \kappa_{k,A} = 0,$$

which implies $\kappa_{j,A} = \kappa_{k,A}$. In this case, LLMs j and k are equivalent for the optimization: they have the same cost-efficiency $\kappa_{j,A} = \kappa_{k,A}$ and the same time-efficiency $\eta_{j,A} = \eta_{k,A}$. Therefore, for any allocation $(w_{j,A}, w_{k,A})$ with $w_{j,A}, w_{k,A} > 0$, we can transfer all mass to a single LLM (say, set $\tilde{w}_{j,A} = w_{j,A} + w_{k,A}$ and $\tilde{w}_{k,A} = 0$) without changing the objective value, i.e., $G_\alpha(\mathbf{w}_A, \mathbf{w}_B) = G_\alpha(\tilde{\mathbf{w}}_A, \mathbf{w}_B)$. Thus, even when two such equivalent LLMs are active, there exists an equally optimal solution that is concentrated. If multiple LLMs are active, we can also find an equally optimal concentrated solution by moving all mass to one LLM.

Case 2: $\eta_{j,A} \neq \eta_{k,A}$. Equation (C.9) forces

$$g'(z_A^*) = K_{jk} := \frac{\kappa_{k,A} - \kappa_{j,A}}{\eta_{j,A} - \eta_{k,A}}. \quad (\text{C.10})$$

We now show that for large $S_A(\alpha)$, $g'(z_A^*)$ avoids all K_{jk} , so generically there is at most one active coordinate. Note that

$$z_A^* = S_A(\alpha) \sum_{j=1}^M \eta_{j,A} \cdot w_{j,A}^* \geq \eta_{\min} \cdot S_A(\alpha),$$

where $\eta_{\min} := \min_j \eta_{j,A} > 0$. Therefore, as $\alpha \rightarrow 0$, we have $z_A^* \rightarrow \infty$. Recall that by Assumption 3, we have $g(x) = c \cdot x^\rho$. Then, we analyze two cases:

Subcase (a): $\rho > 1$. In this case, we have $g(x) = c \cdot x^\rho$ and $g'(x) = c\rho \cdot x^{\rho-1}$. Thus, as $\alpha \rightarrow 0$, we have $z_A^* \rightarrow \infty$ and hence $g'(z_A^*)$ avoids all K_{jk} 's.

Subcase (b): $\rho = 1$. Since $g(x) = c \cdot x$, we have $g'(x) = c$ for any x . If $\min_{\eta_j \neq \eta_k} |c - K_{jk}| > 0$, then $g'(x)$ avoids all K_{jk} .

Suppose $c = K_{jk}$ for some pair with $\eta_{j,A} \neq \eta_{k,A}$. In this case, the KKT conditions allow both coordinates j and k to be active, so we must check whether an interior split can actually be optimal.

Fix such a pair (j, k) and hold all other coordinates constant. Let $s := w_{j,A} + w_{k,A} > 0$. We only need to consider the reduced one-dimensional objective

$$\phi_s(\theta) := \xi_A S_A(\alpha) \kappa_{j,A} \theta s + \xi_A S_A(\alpha) \kappa_{k,A} (1 - \theta) s + c \cdot S_A(\alpha) \cdot (\xi_A \eta_{j,A} \theta s + \xi_A \eta_{k,A} (1 - \theta) s), \quad \theta \in [0, 1].$$

Since $\phi_s(\theta)$ is linear in θ , the minimum of ϕ_s lies at an endpoint, and the allocation can be concentrated on a single coordinate.

Conclusion: For all sufficiently large $S_A(\alpha)$, there exists a minimizer that has at most one positive coordinate for each hypothesis. Moreover, the optimal coordinates (j^*, k^*) can be computed as

$$j^* \in \arg \min_{j \in [M]} \{S_A(\alpha) \kappa_{j,A} + g(S_A(\alpha) \eta_{j,A})\} \quad k^* \in \arg \min_{k \in [M]} \{S_B(\alpha) \kappa_{k,B} + g(S_B(\alpha) \eta_{k,B})\}.$$

□

C.6. Proof of Theorem 2

As $\alpha \rightarrow 0$, the order for Ψ_α is straightforward to verify.

Step 1: Decomposing and bounding the query cost and waiting time.

Let $\pi = \pi(j_A^*, j_B^*)$. We start by decomposing the expected query cost and expected waiting time. Note that the total query cost can be written as

$$C_\pi = \sum_{t=1}^{\tau} c_{j_t} = c_{j_A^*} N_{j_A^*}(\tau) + c_{j_B^*} N_{j_B^*}(\tau).$$

According to Proposition 2 and the definition of $n_{j,\theta}$, if we take expectations under the prior (ξ_A, ξ_B) using the law of total expectation, we have

$$\begin{aligned} \mathbb{E}[C_\pi] &= \xi_A \mathbb{E}_A[C_\pi] + \xi_B \mathbb{E}_B[C_\pi] \\ &= \xi_A (c_{j_A^*} n_{j_A^*,A} + c_{j_B^*} n_{j_B^*,A}) + \xi_B (c_{j_A^*} n_{j_A^*,B} + c_{j_B^*} n_{j_B^*,B}) \\ &= c_{j_A^*} \cdot \xi_A \cdot n_{j_A^*,A} + c_{j_B^*} \cdot \xi_B \cdot n_{j_B^*,B} + (c_{j_A^*} \cdot \xi_B \cdot n_{j_A^*,B} + c_{j_B^*} \cdot \xi_A \cdot n_{j_B^*,A}) \\ &= \kappa_{j_A^*,A} \cdot \xi_A \cdot S_A(\alpha) \cdot \frac{n_{j_A^*,A} \cdot I_{j_A^*,A}}{S_A(\alpha)} + \kappa_{j_B^*,B} \cdot \xi_B \cdot S_B(\alpha) \cdot \frac{n_{j_B^*,B} \cdot I_{j_B^*,B}}{S_B(\alpha)} \\ &\quad + (c_{j_A^*} \cdot \xi_B \cdot n_{j_A^*,B} + c_{j_B^*} \cdot \xi_A \cdot n_{j_B^*,A}). \end{aligned} \tag{C.11}$$

Following a similar logic, we can derive

$$\mathbb{E}_\theta[W_\pi] = \eta_{j_\theta^*,\theta} \cdot S_\theta(\alpha) \cdot \frac{n_{j_\theta^*,\theta} \cdot I_{j_\theta^*,\theta}}{S_\theta(\alpha)} + \mu_{j_{-\theta}^*} \cdot n_{j_{-\theta}^*,\theta}, \tag{C.12}$$

where $-\theta$ is the alternative hypothesis rather than θ .

In the following proposition, we bound the “wrong-side” query counts $n_{j_B^*,A}$ and $n_{j_A^*,B}$.

PROPOSITION C.2. *Under policy $\pi(j_A^*, j_B^*)$, we can bound the wrong-side query counts $n_{j_B^*,A}$ and $n_{j_A^*,B}$ by some constant independent of α , i.e., $n_{j_B^*,A} = O(1)$ and $n_{j_A^*,B} = O(1)$.*

Proof of Proposition C.2. We prove the bound on $n_{j_B^*,A}$; the argument for $n_{j_A^*,B}$ is symmetric.

By Lemma C.8, under $\mathbb{P}_A$ the cumulative LLR admits the drift-martingale decomposition

$$L_t = \sum_{s=1}^t I_{j_s,A} + M_t^{(A)}, \quad |M_t^{(A)} - M_{t-1}^{(A)}| \leq 2C_\ell, \quad \mathbb{E}_A[M_t^{(A)}] = 0.$$

Set $m_A := \min\{I_{j_A^*,A}, I_{j_B^*,A}\} > 0$. Since each conditional drift $I_{j_s,A} \geq m_A$, we have the *pathwise* bound

$$\sum_{s=1}^{t-1} I_{j_s,A} \geq (t-1)m_A \quad \text{for every realization.}$$

Thus on $\{L_{t-1} < -\delta\}$, we have

$$M_{t-1}^{(A)} = L_{t-1} - \sum_{s=1}^{t-1} I_{j_s,A} < -\delta - (t-1)m_A.$$

Define $t_0 := 1 + \lceil 2|\delta|/m_A \rceil$. For $t \geq t_0$, $\delta + (t-1)m_A \geq (t-1)m_A/2 > 0$, so

$$\mathbb{P}_A(L_{t-1} < -\delta) \leq \mathbb{P}_A\left(M_{t-1}^{(A)} < -(t-1)m_A/2\right).$$

Apply Freedman's inequality (Lemma C.9) to the martingale $(M_s^{(A)})$ with bounded differences $|M_t^{(A)} - M_{t-1}^{(A)}| \leq 2C_\ell$ and predictable variation $V_{t-1}^{(A)} \leq v_\ell^2(t-1)$:

$$\begin{aligned} \mathbb{P}_A\left(M_{t-1}^{(A)} < -(t-1)m_A/2\right) &\leq \exp\left(-\frac{((t-1)m_A/2)^2}{2(v_\ell^2(t-1) + 2C_\ell \cdot (t-1)m_A/2)}\right) \\ &= \exp\left(-\frac{(t-1)m_A^2}{8(v_\ell^2 + C_\ell m_A)}\right) = e^{-c_0(t-1)}, \end{aligned}$$

where $c_0 := m_A^2/[8(v_\ell^2 + C_\ell m_A)] > 0$ depends only on m_A, v_ℓ^2, C_ℓ (not on α).

By the sign-based policy, $j_t = j_B^*$ if and only if $L_{t-1} < -\delta$. Therefore

$$n_{j_B^*, A} = \sum_{t=1}^{\infty} \mathbb{P}_A(j_t = j_B^*, t \leq \tau) \leq \sum_{t=1}^{\infty} \mathbb{P}_A(L_{t-1} < -\delta) \leq t_0 + \sum_{t \geq t_0} e^{-c_0(t-1)} =: C_{\text{neg}}^A,$$

which is finite and depends only on $m_A, v_\ell^2, C_\ell, \delta$. The proof for the case when $\theta = B$ is symmetric, with $m_B := \min\{I_{j_A^*, B}, I_{j_B^*, B}\}$. $\square$

Then, we prove that the “correct-side” query counts $I_{j_\theta^*, \theta} n_{j_\theta^*, \theta}$ are constant order away from $S_\theta(\alpha)$.

PROPOSITION C.3. *Under policy $\pi(j_A^*, j_B^*)$, for all sufficiently small α , we have*

$$I_{j_A^*, A} n_{j_A^*, A} = S_A(\alpha) + O(1), \quad I_{j_B^*, B} n_{j_B^*, B} = S_B(\alpha) + O(1).$$

Proof of Proposition C.3. We prove the statements under hypothesis A ; the case B is analogous.

First, according to (6), under policy $\pi(j_A^*, j_B^*)$, we obtain $\mathbb{E}_A[L_\tau] = I_{j_A^*, A} n_{j_A^*, A} + I_{j_B^*, A} n_{j_B^*, A}$. Second, according to Lemma C.12 and Proposition 1, we have $A_\alpha - C_0 \leq \mathbb{E}_A[L_\tau] \leq A_\alpha + C_\ell$. Combining the above two results, we have:

$$A_\alpha - C_0 \leq I_{j_A^*, A} n_{j_A^*, A} + I_{j_B^*, A} n_{j_B^*, A} \leq A_\alpha + C_\ell. \quad (\text{C.13})$$

First, since $I_{j_B^*, A} > 0$, we have

$$I_{j_A^*, A} n_{j_A^*, A} \leq A_\alpha + C_\ell - I_{j_B^*, A} n_{j_B^*, A} \leq A_\alpha + C_\ell = S_A(\alpha) + C_0 + C_\ell.$$

Second, since $I_{j_B^*, A} > 0$ and $n_{j_B^*, A} \leq C_{\text{neg}}^A$ (Proposition C.2), we have

$$I_{j_A^*, A} n_{j_A^*, A} \geq A_\alpha - C_0 - I_{j_B^*, A} n_{j_B^*, A} \geq A_\alpha - C_0 - I_{j_B^*, A} C_{\text{neg}}^A$$

Therefore, we have $I_{j_A^*, A} n_{j_A^*, A} = S_A(\alpha) + O(1)$. $\square$

Combining Propositions C.2 and C.3, we can capture the asymptotic behaviors of the expectations (C.11) and (C.12) as follows:

$$\mathbb{E}[C_\pi] = \xi_A S_A(\alpha) \kappa_{j_A^*, A} + \xi_B S_B(\alpha) \kappa_{j_B^*, B} + O(1), \quad \mathbb{E}_\theta[W_\pi] = S_\theta(\alpha) \eta_{j_\theta^*, \theta} + O(1). \quad (\text{C.14})$$

Step 2: Bounding the waiting time penalty.

We now address the challenge of matching the expected waiting time penalty incurred by π to its deterministic benchmark $\xi_{Ag}\left(\sum_{j=1}^M \mu_j n_{j,A}\right) + \xi_{Bg}\left(\sum_{j=1}^M \mu_j n_{j,B}\right)$ underpinning the lower bound Ψ_α . Let $W_\pi^{(\theta)}$

denote the total waiting time, conditioning on the hypothesis θ . Let $N_\theta := N_{j_\theta^*}(\tau)$ denote the number of queries of LLM j_θ^* . To establish the results, the key step is to anchor

$$W_\pi = \sum_{t=1}^{\tau} T_{j_t, t} = \sum_{\ell=1}^{N_A} T_{j_A^*, \ell} + \sum_{\ell'=1}^{N_B} T_{j_B^*, \ell'} \text{ under each hypothesis } \theta,$$

to its conditional mean

$$m_\pi^\theta = \mathbb{E}_\theta[W_\pi] = \mu_{j_A^*} \cdot n_{j_A^*, \theta} + \mu_{j_B^*} \cdot n_{j_B^*, \theta}.$$

According to Proposition C.2, under each hypothesis $\theta \in \{A, B\}$, when $\alpha \rightarrow 0$, the query will gradually concentrate on j_θ^* so that $\mathbb{E}_\theta[N_\theta] = \mathbb{E}_\theta[\tau] + O(1)$. Motivated by this insight, we are going to match W_π to its mean by establishing the convergence result $N_\theta \rightarrow \mathbb{E}_\theta[\tau]$ for each θ and further control the deviation of the cumulative waiting time to its mean conditioned on the realized τ .

Define $S_\pi := \sum_{t=1}^{\tau} \mu_{j_t}$. In the following proposition, we provide properties of W_π and S_π .

PROPOSITION C.4. *For each $\theta \in \{A, B\}$, given $\pi = \pi(j_A^*, j_B^*)$, we have*

1. *Conditioning on $(\theta, \mathcal{H}_\tau)$, $W_\pi - S_\pi$ is sub-Gaussian with proxy $\psi^2 \tau$, i.e., $\mathbb{E}[e^{\lambda(W_\pi - S_\pi)} \mid \mathcal{H}_\tau, \theta] \leq e^{\psi^2 \lambda^2 \tau / 2}$.*
2. *For each $\rho \geq 1$, we have $\mathbb{E}_\theta[\tau^\rho] = O((\log(1/\alpha))^\rho)$.*
3. *For each $p \geq 2$, we have $\mathbb{E}_\theta|\tau - \mathbb{E}_\theta[\tau]|^p = O((\log(1/\alpha))^{p/2})$ and $\mathbb{E}_\theta|W_\pi - m_\pi^\theta|^p = O((\log(1/\alpha))^{p/2})$.*

Proof of Proposition C.4. We start with the first statement. Fix $\theta \in \{A, B\}$. We will establish the moment generating function (MGF) bound conditional on θ . By definition, we have $W_\pi - S_\pi = \sum_{t=1}^{\tau} (T_{j_t, t} - \mu_{j_t})$. By Assumption 1, the array $(T_{j_t, t})_{j_t, t}$ is independent of $(\theta, (Y_t)_{t \geq 1})$ and of the policy's internal randomization. Consequently, under the sign-based policy, conditional on $(\theta, \mathcal{H}_\tau)$:

- The LLM indices $(j_t)_{t \leq \tau}$ and the stopping time τ are $\mathcal{H}_\tau$ -measurable, hence they are deterministic given $(\theta, \mathcal{H}_\tau)$.
- Given $j_{1:\tau}$, the waiting times $(T_{j_t, t})_{t \leq \tau}$ are independent of $(\theta, \mathcal{H}_\tau)$ and mutually independent across t (since $(T_{j_t, t})_{t \geq 1}$ is i.i.d. in t for each fixed j by Assumption 1).

Then, we can compute the conditional moment generating function as follows:

$$\begin{aligned} \mathbb{E}_\theta[e^{\lambda(W_\pi - S_\pi)} \mid \mathcal{H}_\tau, \theta] &= \mathbb{E}_\theta\left[\prod_{t=1}^{\tau} e^{\lambda(T_{j_t, t} - \mu_{j_t})} \mid \mathcal{H}_\tau, \theta\right] = \prod_{t=1}^{\tau} \mathbb{E}_\theta[e^{\lambda(T_{j_t, t} - \mu_{j_t})} \mid \mathcal{H}_\tau, \theta] \\ &= \prod_{t=1}^{\tau} \mathbb{E}[e^{\lambda(T_{j_t, t} - \mu_{j_t})} \mid j_t] \leq \exp\left(\frac{\psi^2 \lambda^2}{2} \tau\right). \end{aligned}$$

Then, we prove the second statement. Similar to the proof of Proposition C.2, for $t \geq t_1 := 2A_\alpha/I_{\min}$, we have

$$\mathbb{P}_A(\tau > t) \leq \mathbb{P}_A(L_t < A_\alpha) \leq \mathbb{P}_A(L_t < I_{\min} t / 2) \leq e^{-c_3 t}.$$

Thus, we have

$$\mathbb{E}_A[\tau^\rho] = \int_0^\infty \rho t^{\rho-1} \mathbb{P}_A(\tau > t) dt \leq t_1^\rho + \int_{t_1}^\infty \rho t^{\rho-1} e^{-c_3 t} dt = O((\log(1/\alpha))^\rho).$$

Lastly, we prove the third statement. First, we prove that $\mathbb{E}_\theta |\tau - \mathbb{E}_\theta[\tau]|^p = O((\log(1/\alpha))^{p/2})$ for each $p \geq 2$. We fix $\theta = A$ and the proof for the case with $\theta = B$ is symmetric.

By Lemma C.8, we have

$$L_\tau = I_{j_A^*, A} N_A + I_{j_B^*, A} N_B + M_\tau^{(A)},$$

where $(M_t^{(A)})_{t \geq 0}$ is a martingale under $\mathbb{P}_A$ with $\mathbb{E}_A[M_A^{(A)}] = 0$ and bounded increments. Since $\tau = N_A + N_B$, we have

$$I_{j_A^*, A}(\tau - \mathbb{E}_A[\tau]) = (L_\tau - \mathbb{E}_A[L_\tau]) - (I_{j_B^*, A} - I_{j_A^*, A})(N_B - \mathbb{E}_A[N_B]) - M_\tau^{(A)}.$$

By the convexity of $|x|^p$, we have $|x + y + z|^p \leq 3^{p-1}(|x|^p + |y|^p + |z|^p)$. Thus, we can derive that

$$I_{j_A^*, A}^p \mathbb{E}_A |\tau - \mathbb{E}_A[\tau]|^p \leq 3^{p-1} (\mathbb{E}_A |L_\tau - \mathbb{E}_A[L_\tau]|^p + |I_{j_B^*, A} - I_{j_A^*, A}|^p \cdot \mathbb{E}_A |N_B - \mathbb{E}_A[N_B]|^p + \mathbb{E}_A |M_\tau^{(A)}|^p).$$

For the first term, by Lemmas C.7, C.11 and C.12, we have

$$\begin{aligned} \mathbb{E}_A |L_\tau - \mathbb{E}_A[L_\tau]|^p &= \mathbb{P}_A(D = A) \cdot \mathbb{E}_A[|L_\tau - \mathbb{E}_A[L_\tau]|^p \mid D = A] + \mathbb{P}_A(D = B) \cdot \mathbb{E}_A[|L_\tau - \mathbb{E}_A[L_\tau]|^p \mid D = B] \\ &\leq \mathbb{E}_A[(|L_\tau - A_\alpha| + |A_\alpha - \mathbb{E}_A[L_\tau]|)^p \mid D = A] + c_{err} \alpha \mathbb{E}_A[(|L_\tau - A_\alpha| + |A_\alpha - \mathbb{E}_A[L_\tau]|)^p \mid D = B] \\ &= O(1) + O(\alpha \cdot (\log(1/\alpha))^p) = O(1). \end{aligned}$$

For the second term, since $|x + y|^p \leq 2^{p-1}(|x|^p + |y|^p)$, we have

$$\mathbb{E}_A |N_B - \mathbb{E}_A[N_B]|^p \leq 2^{p-1} \cdot (\mathbb{E}_A[N_B^p] + (\mathbb{E}_A[N_B])^p).$$

By Proposition C.2, we have $\mathbb{E}_A[N_B] = O(1)$. According to the proof of Proposition C.2, for $k \geq t_0$, we have

$$\mathbb{P}_A(N_B \geq k) \leq \sum_{t \geq k} \mathbb{P}_A(L_{t-1} < -\delta) \leq \sum_{t \geq k} e^{-c_0(t-1)} = \frac{e^{-c_0(k-1)}}{1 - e^{-c_0}} \leq C_1 e^{-c_0 k},$$

where C_1 is a constant independent of α . Then, we can deduce that

$$\mathbb{E}_A[N_B^p] = \sum_{k=1}^{\infty} p k^{p-1} \mathbb{P}_A(N_B \geq k) \leq \sum_{k=1}^{t_0} p k^{p-1} + \sum_{k=t_0+1}^{\infty} p k^{p-1} C_1 e^{-c_0 k} = O(1).$$

Thus, we have $\mathbb{E}_A |N_B - \mathbb{E}_A[N_B]|^p = O(1)$.

For the third term, according to Burkholder's inequality (Theorem 2.10 in Hall and Heyde 1980), for $p > 1$, there exists a constant C_3 such that

$$\mathbb{E}_A |M_\tau^{(A)}|^p \leq C_3 \mathbb{E}_A \left| \sum_{t=1}^{\tau} (M_t^{(A)} - M_{t-1}^{(A)})^2 \right|^{p/2} \leq C_3 \cdot (2C_\ell)^p \cdot \mathbb{E}_A |\tau|^{p/2} = O((\log(1/\alpha))^{p/2}).$$

Therefore, we can derive that $\mathbb{E}_A |\tau - \mathbb{E}_A[\tau]|^p = O((\log(1/\alpha))^{p/2})$. Similarly, we can prove that $\mathbb{E}_B |\tau - \mathbb{E}_B[\tau]|^p = O((\log(1/\alpha))^{p/2})$.

Since $|x + y|^p \leq 2^{p-1}(|x|^p + |y|^p)$, we have

$$\mathbb{E}_\theta |W_\pi - m_\pi^\theta|^p \leq 2^{p-1} \cdot (\mathbb{E}_\theta |W_\pi - S_\pi|^p + \mathbb{E}_\theta |S_\pi - m_\pi^\theta|^p).$$

For the first term, since $W_\pi - S_\pi \mid \mathcal{H}_\tau, \theta$ is sub-Gaussian, we have

$$\begin{aligned}\mathbb{E}_\theta |W_\pi - S_\pi|^p &= \mathbb{E}_\theta [\mathbb{E}[|W_\pi - S_\pi|^p \mid \mathcal{H}_\tau, \theta]] = \mathbb{E}_\theta \left[\int_0^\infty p t^{p-1} \mathbb{P}(|W_\pi - S_\pi| > t \mid \mathcal{H}_\tau, \theta) dt \right] \\ &= \mathbb{E}_\theta \left[\int_0^\infty p t^{p-1} 2e^{-t^2/(2\psi^2\tau)} dt \right] \leq O(\mathbb{E}_\theta[\tau^{p/2}]) = O((\log(1/\alpha))^{p/2}).\end{aligned}$$

For the second term, we have

$$\begin{aligned}\mathbb{E}_A |S_\pi - m_\pi^A|^p &= \mathbb{E}_A \left| \mu_{j_A^*}(\tau - \mathbb{E}_A[\tau]) + (\mu_{j_B^*} - \mu_{j_A^*})(N_B - \mathbb{E}_A[N_B]) \right|^p \\ &\leq 2^{p-1} \left(\mu_{j_A^*}^p \mathbb{E}_A |\tau - \mathbb{E}_A[\tau]|^p + |\mu_{j_B^*} - \mu_{j_A^*}|^p \cdot \mathbb{E}_A |N_B - \mathbb{E}_A[N_B]|^p \right) \\ &= O((\log(1/\alpha))^{p/2}) + O(1) = O((\log(1/\alpha))^{p/2}).\end{aligned}$$

Therefore, we have $\mathbb{E}_\theta |W_\pi - m_\pi^\theta| \leq O((\log(1/\alpha))^{p/2})$. $\square$

We exploit the polynomial assumption imposed on g to control the fluctuation of $g(W_\pi)$ around $g(\mathbb{E}[W_\pi])$. Although g is convex and increasing, its variation along the sample path remains regulated by the exponential tail of $M_t^{(\theta)}$. In particular, since $\mathbb{E}_\theta[g'(m_\pi^\theta)(W_\pi - m_\pi^\theta)] = g'(m_\pi^\theta) \cdot \mathbb{E}_\theta[W_\pi - m_\pi^\theta] = 0$, we have

$$\begin{aligned}\mathbb{E}_\theta[g(W_\pi)] - g(m_\pi^\theta) &= \mathbb{E}_\theta [g(W_\pi) - g(m_\pi^\theta) - g'(m_\pi^\theta)(W_\pi - m_\pi^\theta)] \\ &= c \cdot (m_\pi^\theta)^\rho \cdot \mathbb{E}_\theta[\phi(W_\pi/m_\pi^\theta)],\end{aligned}$$

where $\phi(\cdot)$ is defined in Appendix C.1.6. By Lemma C.13 and Proposition C.4, we have

$$\begin{aligned}\mathbb{E}_\theta[g(W_\pi)] - g(m_\pi^\theta) &= c \cdot (m_\pi^\theta)^\rho \mathbb{E}_\theta[\phi(W_\pi/m_\pi^\theta)] \leq c \cdot c_1 \cdot ((m_\pi^\theta)^{\rho-2} \mathbb{E}_\theta[(W_\pi - m_\pi^\theta)^2] + (m_\pi^\theta)^{\rho-q} \mathbb{E}_\theta[(W_\pi - m_\pi^\theta)^q]) \\ &\leq O((\log(1/\alpha))^{\rho-1}) + O((\log(1/\alpha))^{\rho-q/2}) \\ &= O((\log(1/\alpha))^{\rho-1})\end{aligned}\tag{C.15}$$

where the last equality follows since $q \geq 2$.

Moreover, by Propositions C.2 and C.3, we have $\mathbb{E}_A[\tau] = n_{j_A^*, A} + n_{j_B^*, A} = \Theta(\log(1/\alpha))$, which implies that

$$g(m_\pi^\theta) = \Theta((\log(1/\alpha))^\rho).\tag{C.16}$$

Step 3: Combining the bounds.

By equation C.14, (C.15) and (C.16), we have

$$\begin{aligned}\mathcal{R}(\pi; \alpha) &= \mathbb{E}[C_\pi] + \mathbb{E}[g(W_\pi)] = \mathbb{E}[C_\pi] + \xi_A g(\mathbb{E}_A[W_\pi]) + \xi_B g(\mathbb{E}_B[W_\pi]) + O((\log(1/\alpha))^{\rho-1}) \\ &= \xi_A S_A(\alpha) \kappa_{j_A^*, A} + \xi_B S_B(\alpha) \kappa_{j_B^*, B} + O(1) \\ &\quad + \xi_A g(S_A(\alpha) \eta_{j_A^*, A} + O(1)) + \xi_B g(S_B(\alpha) \eta_{j_B^*, B} + O(1)) + O((\log(1/\alpha))^{\rho-1}) \\ &= \xi_A S_A(\alpha) \kappa_{j_A^*, A} + \xi_B S_B(\alpha) \kappa_{j_B^*, B} + \xi_A g(S_A(\alpha) \eta_{j_A^*, A}) + \xi_B g(S_B(\alpha) \eta_{j_B^*, B}) + O((\log(1/\alpha))^{\rho-1}) \\ &= \Psi_\alpha + O((\log(1/\alpha))^{\rho-1}).\end{aligned}$$

which establishes (11). $\square$

C.7. Proof of Theorem 3

Assume $c > 0$ and $\rho > 1$. We work first under A and temporarily suppress the hypothesis subscript on I_j, κ_j, η_j . Define

$$\eta^* := \min_j \eta_j > 0, \quad \kappa^* := \min_{j: \eta_j = \eta^*} \kappa_j, \quad v_{\min} := \min_j \text{Var}_A(\ell_j(Y)) > 0.$$

For the chosen policy, write $n_j = \mathbb{E}_A[N_j(\tau)]$, $m = \mathbb{E}_A[W_\pi]$, and

$$Q(\mathbf{n}) := \sum_j I_j n_j, \quad F_A(\mathbf{n}) := \sum_j c_j n_j + g\left(\sum_j \mu_j n_j\right), \quad \Psi_{\alpha, A} := \min_{\mathbf{x} \geq 0: Q(\mathbf{x}) \geq S_A(\alpha)} F_A(\mathbf{x}).$$

The definition of F_A agrees with the conditional objective in the proof of Proposition 3.

Set $d_j := \mu_j - \eta^* I_j = I_j(\eta_j - \eta^*) \geq 0$. Define stopped processes, with zero initial values, by

$$V_t := \sum_{s=1}^t \mathbf{1}\{s \leq \tau\} (T_{j_s, s} - \eta^* \ell_{j_s}(Y_s)),$$

$$\Gamma_t := \sum_{s=1}^t \mathbf{1}\{s \leq \tau\} d_{j_s}, \quad Z_t := V_t - \Gamma_t.$$

Let $\Gamma := \Gamma_\tau$ and $Z := Z_\tau$. Then

$$W_\pi - \eta^* L_\tau = \Gamma + Z. \tag{C.17}$$

The nonnegative term Γ measures excess expected delay relative to the best time per unit of evidence. The martingale term Z captures fresh delay noise and evidence noise. Sources tied in time efficiency have $d_j = 0$ and cause no difficulty in this decomposition.

C.7.1. Supporting Results We give the moment and deviation properties of (C.17).

LEMMA C.14. *There are constants $\underline{\sigma}^2 = (\eta^*)^2 v_{\min} > 0$, $C_Z < \infty$, and $K_0 < \infty$, independent of the policy and α , such that*

$$\Gamma_t \text{ is nondecreasing,} \quad \mathbb{E}_A[\Gamma] = \sum_j d_j n_j,$$

$$\mathbb{E}_A[Z] = 0, \quad \mathbb{E}_A[Z^2] \geq \underline{\sigma}^2 t_\alpha, \quad \mathbb{E}_A[Z^4] \leq C_Z \log(1/\alpha)^2,$$

and, on $\{D = A\}$,

$$|W_\pi - m| \geq |Z| - \Gamma - \mathbb{E}_A[\Gamma] - K_0.$$

Moreover, $\mathbb{E}_A[Z^2] = O(\log(1/\alpha))$ uniformly over policies.

Proof. The claims about Γ follow from its nonnegative predictable increments and Tonelli's theorem. The increments of Z_t are

$$\Delta Z_t = \mathbf{1}\{t \leq \tau\} [(T_{j_t, t} - \mu_{j_t}) - \eta^* (\ell_{j_t}(Y_t) - I_{j_t})].$$

They have conditional mean zero. Given the pre-query history, the two centered terms are independent, by Assumption 1. Their conditional MGFs and variances therefore give

$$\mathbb{E}_A[e^{\lambda \Delta Z_t} \mid \mathcal{G}_{t-1}] \leq \exp\left(\frac{\tilde{\psi}^2 \lambda^2}{2} \mathbf{1}\{t \leq \tau\}\right), \quad \tilde{\psi}^2 := \psi^2 + 4(\eta^*)^2 C_\ell^2,$$

$$\mathbb{E}_A[(\Delta Z_t)^2 \mid \mathcal{G}_{t-1}] = \mathbf{1}\{t \leq \tau\} (\sigma_{j_t}^2 + (\eta^*)^2 \text{Var}_A(\ell_{j_t}(Y_t))) \geq \underline{\sigma}^2 \mathbf{1}\{t \leq \tau\}.$$

The MGF bound also bounds the conditional variance above by $\tilde{\psi}^2 \mathbf{1}\{t \leq \tau\}$. Orthogonality of martingale increments consequently gives

$$\mathbb{E}_A[Z_t^2] \leq \tilde{\psi}^2 \mathbb{E}_A[t \wedge \tau].$$

The martingale is Cauchy in L^2 , because the mean square of its increment after time t is at most $\tilde{\psi}^2 \mathbb{E}_A[(\tau - t)^+] \rightarrow 0$. Its almost-sure terminal value Z thus has mean zero, and

$$\sigma^2 \mathbb{E}_A[\tau] \leq \mathbb{E}_A[Z^2] \leq \tilde{\psi}^2 \mathbb{E}_A[\tau] = O(\log(1/\alpha)).$$

Use $\tau \geq t_\alpha$ for the lower bound.

Moreover, the conditional sub-Gaussian bound on ΔZ_t implies that, for some constant $C_4 < \infty$ independent of α and the policy,

$$\mathbb{E}_A[|\Delta Z_t|^4 \mid \mathcal{G}_{t-1}] \leq C_4 \mathbf{1}\{t \leq \tau\}.$$

By the standard fourth-moment inequality for martingales,

$$\mathbb{E}_A[|Z_{n \wedge \tau}|^4] \leq C \left(\mathbb{E}_A[(n \wedge \tau)^2] + \mathbb{E}_A[n \wedge \tau] \right).$$

The same concentration argument as in the proof of Proposition C.4, using $I_{\min, A} := \min_j I_{j, A} > 0$, yields uniformly over admissible policies $\mathbb{E}_A[\tau^2] = O(\log^2(1/\alpha))$. Hence,

$$\sup_n \mathbb{E}_A[|Z_{n \wedge \tau}|^4] = O(\log^2(1/\alpha)).$$

Letting $n \rightarrow \infty$ and applying Fatou's lemma gives

$$\mathbb{E}_A[Z^4] = O(\log^2(1/\alpha)).$$

Finally, centering (C.17) gives

$$W_\pi - m = Z + (\Gamma - \mathbb{E}_A \Gamma) + \eta^*(L_\tau - \mathbb{E}_A L_\tau).$$

On $\{D = A\}$, the last term has magnitude at most $\eta^*(C_0 + C_\ell)$ by the overshoot and expected-evidence bounds. Taking this as K_0 and applying the triangle inequality proves the deviation claim. $\square$

LEMMA C.15. *For all sufficiently small α , the conditional benchmark is*

$$\Psi_{\alpha, A} = S_A(\alpha) \kappa^* + g(S_A(\alpha) \eta^*).$$

There is a policy-independent $k_1 > 0$ such that

$$F_A(\mathbf{n}) - \Psi_{\alpha, A} \geq \frac{\iota_\alpha}{2} \mathbb{E}_A[\Gamma] \geq k_1 \log(1/\alpha)^{\rho-1} \mathbb{E}_A[\Gamma], \quad \iota_\alpha := g'(S_A(\alpha) \eta^*).$$

Proof. Let $S = S_A(\alpha)$ and $D_j := (\kappa_j - \kappa^*) + \iota_\alpha(\eta_j - \eta^*)$. When $\eta_j = \eta^*$, $D_j = \kappa_j - \kappa^* \geq 0$. When $\eta_j > \eta^*$, the divergence of $\iota_\alpha = c\rho(S\eta^*)^{\rho-1}$ implies

$$D_j \geq \frac{\iota_\alpha}{2} (\eta_j - \eta^*) \geq 0$$

for all sufficiently small α , uniformly across the finite source pool.

For any feasible allocation $\mathbf{x}$, the supporting-line inequality for g at $S\eta^*$ yields

$$F_A(\mathbf{x}) - [S\kappa^* + g(S\eta^*)] \geq \sum_j I_j x_j D_j + (\kappa^* + \iota_\alpha \eta^*)[Q(\mathbf{x}) - S] \geq 0.$$

Equality is attained by allocating S/I_j queries to a source with $\eta_j = \eta^*$ and $\kappa_j = \kappa^*$. This proves the benchmark formula. For the expected allocation induced by a policy, Proposition 1 gives $Q(\mathbf{n}) \geq S$. Retaining the first term of the supporting-line bound gives

$$F_A(\mathbf{n}) - \Psi_{\alpha,A} \geq \frac{\iota_\alpha}{2} \sum_j I_j n_j (\eta_j - \eta^*) = \frac{\iota_\alpha}{2} \mathbb{E}_A[\Gamma].$$

Since $\iota_\alpha = \Theta(\log(1/\alpha)^{\rho-1})$, the result follows. This argument retains the nonnegative contribution of excess evidence, rather than identifying the allocation gap solely with the use of a different source. $\square$

LEMMA C.16. *There are constants $k_2 > 0$ and $\alpha_2 > 0$, independent of the policy, such that every admissible policy with $\mathbb{E}_A[\Gamma] \leq 1$ satisfies*

$$\mathbb{E}_A[g(W_\pi)] - g(\mathbb{E}_A[W_\pi]) \geq k_2 \log(1/\alpha)^{\rho-1}, \quad 0 < \alpha < \alpha_2.$$

Proof. Proposition 2 implies

$$\mu_{\min} t_\alpha \leq m \leq \mu_{\max} \mathbb{E}_A[\tau] = O(\log(1/\alpha)),$$

so $m = \Theta(\log(1/\alpha))$ uniformly over policies. For sufficiently small α , $t_\alpha \geq b_0 \log(1/\alpha)$ for some fixed $b_0 > 0$, for example $b_0 = 1/(2C_\ell)$. Define

$$u_\pi := \frac{1}{2} \sqrt{\mathbb{E}_A[Z^2]}.$$

Lemma C.14 gives both $u_\pi \geq \frac{1}{2} \sqrt{\sigma^2 b_0 \log(1/\alpha)}$ and $u_\pi = O(\sqrt{\log(1/\alpha)})$. Applying the Paley–Zygmund inequality to Z^2 yields

$$\mathbb{P}_A(|Z| \geq u_\pi) \geq \frac{9(\mathbb{E}_A Z^2)^2}{16\mathbb{E}_A Z^4} \geq \frac{9\sigma^4 b_0^2}{16C_Z}.$$

Choose a fixed positive p_0 no larger than this last constant and $1/2$. Unlike a bound retaining t_α without its logarithmic denominator, this choice is independent of α and the policy.

Markov's inequality and Lemma C.7 give

$$\mathbb{P}_A(\Gamma > u_\pi/4) \leq 4/u_\pi, \quad \mathbb{P}_A(D = B) \leq c_{\text{err}} \alpha.$$

On $\{D = A, \Gamma \leq u_\pi/4, |Z| \geq u_\pi\}$, Lemma C.14 therefore implies

$$|W_\pi - m| \geq u_\pi - u_\pi/4 - 1 - K_0 \geq u_\pi/2$$

for sufficiently small α , uniformly over the policies in question. Consequently,

$$\mathbb{P}_A(|W_\pi - m| \geq u_\pi/2) \geq p_0 - 4/u_\pi - c_{\text{err}} \alpha \geq p_0/2. \quad (\text{C.18})$$

Since $\mathbb{E}_A[W_\pi - m] = 0$, the normalized convexity remainder gives

$$\mathbb{E}_A[g(W_\pi)] - g(m) = cm^\rho \mathbb{E}_A[\phi(W_\pi/m)].$$

For sufficiently small α , $u_\pi/(2m) \leq 1/2$. On the event in (C.18), Lemma C.13 bounds the integrand below by $c_2(u_\pi/(2m))^2$; elsewhere it is nonnegative. Hence

$$\mathbb{E}_A[g(W_\pi)] - g(m) \geq \frac{cc_2 p_0}{8} m^{\rho-2} u_\pi^2 \geq k_2 \log(1/\alpha)^{\rho-1}.$$

Both upper and lower bounds on $m/\log(1/\alpha)$ were established, so the last step holds also when $1 < \rho < 2$.

$\square$

C.7.2. Proof of the Lower Bound (13) For each hypothesis, decompose the conditional gap as

$$\Xi_\theta := \mathbb{E}_\theta[C_\pi + g(W_\pi)] - \Psi_{\alpha,\theta} = [F_\theta(\mathbf{n}_\theta) - \Psi_{\alpha,\theta}] + [\mathbb{E}_\theta g(W_\pi) - g(\mathbb{E}_\theta W_\pi)]. \quad (\text{C.19})$$

Both terms are nonnegative. Under A , if $\mathbb{E}_A[\Gamma] \geq 1$, Lemma C.15 bounds the allocation term below by $k_1 \log(1/\alpha)^{\rho-1}$. If $\mathbb{E}_A[\Gamma] < 1$, Lemma C.16 gives a lower bound of this order on the fluctuation term. Thus $\Xi_A \geq K_A \log(1/\alpha)^{\rho-1}$ with a fixed $K_A > 0$, uniformly over all policies.

Under B , repeat the argument with $-L_t$ as evidence in favor of the truth and with the B -specific information ratios. This gives a fixed $K_B > 0$. The relaxed problem separates by hypothesis, so $\Psi_\alpha = \xi_A \Psi_{\alpha,A} + \xi_B \Psi_{\alpha,B}$. Therefore,

$$\mathcal{R}(\pi; \alpha) - \Psi_\alpha = \xi_A \Xi_A + \xi_B \Xi_B \geq (\xi_A K_A + \xi_B K_B) \log(1/\alpha)^{\rho-1}.$$

Taking the infimum over policies proves the theorem. The variance lower bound used only positive evidence variance; deterministic per-query response times and ties in time efficiency are both allowed. $\square$

Appendix D: Omitted Proofs in Section 6

D.1. Proof of Proposition 5

First, we present the Berry-Esseen bound in Shevtsova (2010).

LEMMA D.17 (Berry–Esseen Bound for Independent Summation; Shevtsova 2010). *Let $X_1, \dots, X_t$ be independent with $\mathbb{E}[X_i] = 0$, variances $\text{Var}(X_i) = \sigma_i^2$, and third absolute moments $\mathbb{E}[|X_i|^3] < \infty$. Let $S_t := \sum_{i=1}^t X_i$ and $s_t^2 := \sum_{i=1}^t \sigma_i^2 > 0$. Then, for all $x \in \mathbb{R}$,*

$$\left| \mathbb{P}\left(\frac{S_t}{s_t} \leq x\right) - \Phi(x) \right| \leq C_{\text{BE}} \frac{\sum_{i=1}^t \mathbb{E}[|X_i|^3]}{s_t^3},$$

where Φ is the standard normal CDF and one may take $C_{\text{BE}} = 0.56$.

Since $\ell_j(Y) \in \{\ell_j(A), \ell_j(B)\}$ and $\gamma_{j,\theta} \in (1/2, 1)$, the increments are bounded by $C_\ell < \infty$ and have finite third moments. Under $\theta = A$ write

$$L_\omega - \mu_A(\mathbf{n}) = \sum_{j=1}^M \sum_{i=1}^{n_j} X_{j,i}, \quad X_{j,i} := \ell_j(Y_{j,i}) - \mathbb{E}_A[\ell_j(Y)],$$

where $\{Y_{j,i}\}$ are independent draws from model j (static sampling). The $X_{j,i}$ are independent, mean zero, and satisfy $|X_{j,i}| \leq 2C_\ell$. Let

$$\mathcal{T}_\theta(\mathbf{n}) := \sum_{j=1}^M n_j \mathbb{E}_\theta \left[|\ell_j(Y) - \mathbb{E}_\theta[\ell_j(Y)]|^3 \right].$$

Then Lemma D.17 implies that for all $x \in \mathbb{R}$,

$$\left| \mathbb{P}_A\left(\frac{L_\omega - \mu_A(\mathbf{n})}{\sigma_A(\mathbf{n})} < x\right) - \Phi(x) \right| \leq C_{\text{BE}} \frac{\mathcal{T}_A(\mathbf{n})}{\sigma_A(\mathbf{n})^3} \leq \frac{\mathcal{T}_A(\mathbf{n})}{\sigma_A(\mathbf{n})^3}.$$

As mentioned, a feasible policy should satisfy $\mathbb{P}_A(D = B) \leq \alpha/\xi_A$ and $\mathbb{P}_B(D = A) \leq \alpha/\xi_B$. Under $\theta = A$, by Berry-Esseen bound in Lemma D.17, we have

$$\mathbb{P}_A(D = B) = \mathbb{P}_A\left(\frac{L_n - \mu_A(\mathbf{n})}{\sigma_A(\mathbf{n})} < \frac{-\delta - \mu_A(\mathbf{n})}{\sigma_A(\mathbf{n})}\right) \geq \Phi\left(-\frac{\delta + \mu_A(\mathbf{n})}{\sigma_A(\mathbf{n})}\right) - \frac{\mathcal{T}_A(\mathbf{n})}{\sigma_A(\mathbf{n})^3}.$$

Thus we have

$$\frac{\mu_A(\mathbf{n}) + \delta}{\sigma_A(\mathbf{n})} \geq \nu \left(\frac{\alpha}{\xi_A} + \frac{\mathcal{T}_A(\mathbf{n})}{\sigma_A(\mathbf{n})^3} \right).$$

Similarly, for $\theta = B$, by Berry-Esseen bound in Lemma D.17, we have

$$\frac{\mu_B(\mathbf{n}) - \delta}{\sigma_B(\mathbf{n})} \geq \nu \left(\frac{\alpha}{\xi_B} + \frac{\mathcal{T}_B(\mathbf{n})}{\sigma_B(\mathbf{n})^3} \right).$$

□

D.2. Proof of Theorem 4

For $j \in [M]$, let

$$s_j := a_j + b_j > 0, \quad r_j := \frac{c_j}{s_j^2}, \quad r^* := \min_j r_j > 0, \quad j^* \in \arg \min_j r_j.$$

Let $z_\alpha := \nu(\alpha) = \Phi^{-1}(1 - \alpha)$. The following two propositions state precisely the lower- and upper-limit conclusions needed for the theorem.

PROPOSITION D.5 (Asymptotic static cost). *With zero waiting cost and weak-signal scaling,*

$$\lim_{\alpha \rightarrow 0} \liminf_{\epsilon \rightarrow 0} \frac{\epsilon^2 \mathcal{R}^{\text{st},*}(\alpha, \epsilon)}{z_\alpha^2} = r^*, \quad \lim_{\alpha \rightarrow 0} \limsup_{\epsilon \rightarrow 0} \frac{\epsilon^2 \mathcal{R}^{\text{st},*}(\alpha, \epsilon)}{z_\alpha^2} = r^*.$$

PROPOSITION D.6 (Asymptotic adaptive cost). *Under the same scaling,*

$$\lim_{\alpha \rightarrow 0} \liminf_{\epsilon \rightarrow 0} \frac{\epsilon^2 \mathcal{R}^*(\alpha, \epsilon)}{z_\alpha^2} = \frac{r^*}{4}, \quad \lim_{\alpha \rightarrow 0} \limsup_{\epsilon \rightarrow 0} \frac{\epsilon^2 \mathcal{R}^*(\alpha, \epsilon)}{z_\alpha^2} = \frac{r^*}{4}.$$

The propositions are proved in the next two subsections. To combine them rigorously, define

$$a_{\alpha, \epsilon} := \frac{\epsilon^2 \mathcal{R}^{\text{st},*}(\alpha, \epsilon)}{z_\alpha^2}, \quad b_{\alpha, \epsilon} := \frac{\epsilon^2 \mathcal{R}^*(\alpha, \epsilon)}{z_\alpha^2}.$$

The adaptive lower bound below makes $\liminf_{\epsilon \rightarrow 0} b_{\alpha, \epsilon} > 0$ for sufficiently small fixed α , and both upper limits are finite. Therefore,

$$\begin{aligned} \frac{\liminf_{\epsilon \rightarrow 0} a_{\alpha, \epsilon}}{\limsup_{\epsilon \rightarrow 0} b_{\alpha, \epsilon}} &\leq \liminf_{\epsilon \rightarrow 0} \frac{a_{\alpha, \epsilon}}{b_{\alpha, \epsilon}} \\ &\leq \limsup_{\epsilon \rightarrow 0} \frac{a_{\alpha, \epsilon}}{b_{\alpha, \epsilon}} \leq \frac{\limsup_{\epsilon \rightarrow 0} a_{\alpha, \epsilon}}{\liminf_{\epsilon \rightarrow 0} b_{\alpha, \epsilon}}. \end{aligned}$$

As $\alpha \rightarrow 0$, the first and last expressions both converge to $r^*/(r^*/4) = 4$. This proves both ordered-limit statements in Theorem 4, without assuming that an ordinary inner limit exists for every fixed α . □

D.3. Proof of Proposition D.5

Weak-signal expansions. Fix a sufficiently small upper bound on ϵ that keeps all accuracies strictly inside $(1/2, 1)$. For each LLM, write $p_{j,A} = 1/2 + a_j \epsilon$ and $p_{j,B} = 1/2 - b_j \epsilon$ for the probability of output A under the respective hypotheses. Taylor expansion of the two logarithms gives

$$\begin{aligned} \ell_j(A) &= \log \frac{1 + 2a_j \epsilon}{1 - 2b_j \epsilon}, \quad \ell_j(B) = \log \frac{1 - 2a_j \epsilon}{1 + 2b_j \epsilon}, \\ I_{j,\theta} &= 2s_j^2 \epsilon^2 + O(\epsilon^4), \quad V_{j,\theta} = 4s_j^2 \epsilon^2 + O(\epsilon^4), \\ \beta_{j,\theta} &:= \mathbb{E}_\theta |\ell_j(Y) - \mathbb{E}_\theta \ell_j(Y)|^3 = 8s_j^3 \epsilon^3 + O(\epsilon^5). \end{aligned}$$

For the last two expansions, if $d_j := \ell_j(A) - \ell_j(B)$ and $p = p_{j,\theta}$, the exact Bernoulli formulas are

$$V_{j,\theta} = p(1-p)d_j^2, \quad \beta_{j,\theta} = p(1-p)[p^2 + (1-p)^2]|d_j|^3,$$

with $d_j = 4s_j\epsilon + O(\epsilon^3)$.

Define the total signal budget

$$Q_\epsilon(\mathbf{n}) := \epsilon^2 \sum_j s_j^2 n_j.$$

Since $0 < s_{\min} \leq s_j \leq s_{\max} < \infty$, there is a fixed K such that, for *every* nonnegative allocation and sufficiently small ϵ ,

$$\begin{aligned} \mu_\theta(\mathbf{n}) &= 2Q_\epsilon(\mathbf{n})[1 + O(\epsilon^2)], \\ \sigma_\theta^2(\mathbf{n}) &= 4Q_\epsilon(\mathbf{n})[1 + O(\epsilon^2)], \\ \mathcal{T}_\theta(\mathbf{n}) &\leq K\epsilon Q_\epsilon(\mathbf{n}), \quad \Delta_\theta(\mathbf{n}) \leq \frac{K\epsilon}{\sqrt{Q_\epsilon(\mathbf{n})}} \quad (Q_\epsilon(\mathbf{n}) > 0). \end{aligned}$$

All relative error constants here are independent of the allocation.

Normal-quantile scale. For $u \rightarrow 0$, let $z_u = \Phi^{-1}(1-u)$ and let φ be the standard normal density. The Mills-ratio bounds used in the manuscript (Mills 1926) give

$$\frac{z_u}{1+z_u^2} \varphi(z_u) \leq u \leq \frac{\varphi(z_u)}{z_u}.$$

Taking logarithms yields $z_u^2 \sim 2\log(1/u)$. Consequently, for every fixed $k > 0$,

$$\frac{\nu(k\alpha)}{z_\alpha} \rightarrow 1, \quad \frac{H_\alpha}{z_\alpha^2} \rightarrow \frac{1}{2} \quad \text{as } \alpha \rightarrow 0.$$

Lower bound. Fix $0 < \alpha < \xi_{\min}/2$. The information constraints established above and the uniform mean expansion imply $Q_\epsilon(\mathbf{n}) \geq b_\alpha/3 > 0$ for every feasible design when ϵ is sufficiently small. Thus its total query count diverges at least on the order of ϵ^{-2} , and the Berry–Esseen correction converges to zero *uniformly over feasible allocations*. Proposition 5 gives

$$\frac{\mu_A(\mathbf{n}) + \delta}{\sigma_A(\mathbf{n})} \geq \nu(\alpha/\xi_A + \Delta_A(\mathbf{n})), \quad \frac{\mu_B(\mathbf{n}) - \delta}{\sigma_B(\mathbf{n})} \geq \nu(\alpha/\xi_B + \Delta_B(\mathbf{n})). \quad (\text{D.20})$$

Writing $Q = Q_\epsilon(\mathbf{n})$, the uniform expansions give

$$\frac{\mu_A(\mathbf{n}) + \delta}{\sigma_A(\mathbf{n})} + \frac{\mu_B(\mathbf{n}) - \delta}{\sigma_B(\mathbf{n})} = 2\sqrt{Q} + O(\epsilon^2\sqrt{Q}) + O(\epsilon^2/\sqrt{Q}). \quad (\text{D.21})$$

The leading prior offsets cancel; their residual effect is included in the last term. Since $Q \geq b_\alpha/3$, the last term is $o(1)$. The right-hand sides of (D.20) sum to at least $\nu(\alpha/\xi_A) + \nu(\alpha/\xi_B) + o(1)$. Bounding the middle error term in (D.21) above by $K\epsilon^2\sqrt{Q}$ and rearranging therefore gives

$$\sqrt{Q} \geq \frac{\nu(\alpha/\xi_A) + \nu(\alpha/\xi_B) + o(1)}{2(1 + K\epsilon^2)}.$$

The quantiles are positive for the chosen α , so this inequality can be squared. Also,

$$\epsilon^2 \sum_j c_j n_j \geq r^* \epsilon^2 \sum_j s_j^2 n_j = r^* Q.$$

Taking the infimum over feasible designs and then the lower limit yields

$$\liminf_{\epsilon \rightarrow 0} \epsilon^2 \mathcal{R}^{\text{st},*}(\alpha, \epsilon) \geq \frac{r^*}{4} [\nu(\alpha/\xi_A) + \nu(\alpha/\xi_B)]^2.$$

The uniformity above makes this step valid for allocations chosen optimally as functions of ϵ .

Upper bound. Use the source j^* in every query and select an integer sample size

$$n_{\alpha, \epsilon}^* := \left\lceil \frac{x_\alpha^2}{s_{j^*}^2 \epsilon^2} \right\rceil, \quad x_\alpha := \max_{\theta \in \{A, B\}} \nu\left(\frac{\alpha}{2\xi_\theta}\right) + |\delta| + 1.$$

For the fixed $\alpha < \xi_{\min}/2$, $x_\alpha > 1$, and this allocation has $Q_\epsilon = x_\alpha^2 + O(\epsilon^2)$. Under the two hypotheses, its standardized distances to the decision boundary converge to

$$x_\alpha + \frac{\delta}{2x_\alpha}, \quad x_\alpha - \frac{\delta}{2x_\alpha}.$$

Both are at least $\max_\theta \nu(\alpha/(2\xi_\theta)) + 1$. The limiting normal error probabilities are therefore strictly below $\alpha/(2\xi_\theta)$, and the Berry–Esseen correction tends to zero. For sufficiently small ϵ , the actual errors satisfy $e_\theta \leq \alpha/(2\xi_\theta)$, implying $\xi_A e_A + \xi_B e_B \leq \alpha$. This is an explicit feasible static policy, not merely a feasible point of the relaxation. Its cost gives

$$\limsup_{\epsilon \rightarrow 0} \epsilon^2 \mathcal{R}^{\text{st},*}(\alpha, \epsilon) \leq \lim_{\epsilon \rightarrow 0} \epsilon^2 c_{j^*} n_{\alpha, \epsilon}^* = r^* x_\alpha^2.$$

Dividing both bounds by z_α^2 and then letting $\alpha \rightarrow 0$ proves both claims in Proposition D.5, because $[\nu(\alpha/\xi_A) + \nu(\alpha/\xi_B)]/(2z_\alpha) \rightarrow 1$ and $x_\alpha/z_\alpha \rightarrow 1$. $\square$

D.4. Proof of Proposition D.6

Write

$$\bar{H}_\alpha := \xi_A A_\alpha + \xi_B B_\alpha = H_\alpha + (\xi_B - \xi_A)\delta.$$

We derive bounds that are uniform over adaptive policies before taking either limit. The single-query expansions above imply, for some fixed K and every j, θ ,

$$I_{j, \theta}(\epsilon) \leq 2s_j^2 \epsilon^2 (1 + K\epsilon^2), \quad C_\ell(\epsilon) = O(\epsilon).$$

All stopping-time and drift identities from Appendix C apply for each fixed $\epsilon > 0$. Their fixed-instance remainder constants need not be uniform in ϵ ; the explicit bounds below avoid that requirement.

Lower bound. Let π be any admissible policy. Put $U_{A, \tau} = L_\tau$, $U_{B, \tau} = -L_\tau$, and let $e_\theta = \mathbb{P}_\theta(D \neq \theta)$. From $c_j \geq r^* s_j^2$, the expected-count identities imply

$$\epsilon^2 \mathcal{R}(\pi; \alpha, \epsilon) = \epsilon^2 \sum_\theta \xi_\theta \sum_j c_j n_{j, \theta} \geq \frac{r^*}{2(1 + K\epsilon^2)} \sum_\theta \xi_\theta \sum_j I_{j, \theta} n_{j, \theta} = \frac{r^*}{2(1 + K\epsilon^2)} \sum_\theta \xi_\theta \mathbb{E}_\theta[U_{\theta, \tau}].$$

The overshoot argument, retaining the actual error probabilities, gives

$$\mathbb{E}_A[L_\tau] \geq A_\alpha - (2H_\alpha + C_\ell(\epsilon))e_A, \quad \mathbb{E}_B[-L_\tau] \geq B_\alpha - (2H_\alpha + C_\ell(\epsilon))e_B.$$

Every admissible posterior-threshold policy has Bayes error $\xi_A e_A + \xi_B e_B \leq \alpha$. Therefore,

$$\epsilon^2 \mathcal{R}(\pi; \alpha, \epsilon) \geq \frac{r^*}{2(1 + K\epsilon^2)} [\bar{H}_\alpha - \alpha(2H_\alpha + C_\ell(\epsilon))].$$

The inequality holds for all policies, whether or not an optimal policy is attained. Taking their infimum and then $\liminf_{\epsilon \rightarrow 0}$ yields

$$\liminf_{\epsilon \rightarrow 0} \epsilon^2 \mathcal{R}^*(\alpha, \epsilon) \geq \frac{r^*}{2} (\bar{H}_\alpha - 2\alpha H_\alpha).$$

The right-hand side is positive for all sufficiently small α . In particular, the wrong-boundary contribution is kept throughout this argument; it is not replaced by a vanishing overshoot constant.

Upper bound. Use only source j^* and stop at the original posterior boundaries, with stopping time τ_0 . This is admissible and has finite expected query count for each fixed ϵ . The stopped drift and overshoot identities give

$$I_{j^*,A} \mathbb{E}_A[\tau_0] = \mathbb{E}_A[L_{\tau_0}] \leq A_\alpha + C_\ell(\epsilon), \quad I_{j^*,B} \mathbb{E}_B[\tau_0] = \mathbb{E}_B[-L_{\tau_0}] \leq B_\alpha + C_\ell(\epsilon).$$

Since $I_{j^*,\theta} = 2s_{j^*}^2 \epsilon^2 [1 + O(\epsilon^2)]$ and $C_\ell(\epsilon) \rightarrow 0$,

$$\begin{aligned} \limsup_{\epsilon \rightarrow 0} \epsilon^2 \mathcal{R}^*(\alpha, \epsilon) &\leq \lim_{\epsilon \rightarrow 0} \epsilon^2 c_{j^*} \left[\frac{\xi_A(A_\alpha + C_\ell(\epsilon))}{I_{j^*,A}} + \frac{\xi_B(B_\alpha + C_\ell(\epsilon))}{I_{j^*,B}} \right] \\ &= \frac{r^*}{2} \bar{H}_\alpha. \end{aligned}$$

Finally, $\bar{H}_\alpha/z_\alpha^2 \rightarrow 1/2$ and $2\alpha H_\alpha/z_\alpha^2 \rightarrow 0$ as $\alpha \rightarrow 0$. Dividing the bounds by z_α^2 proves both assertions in Proposition D.6. $\square$

Appendix E: Further Details on the Numerical Experiments

This appendix describes the Bellman benchmarks and waiting time distribution used in Section 7. Hypothesis A means that the candidate solution is correct, whereas B means that it is incorrect. The simulation uses the symmetric prior $\xi_A = \xi_B = 1/2$, so $\delta = 0$ and $A_\alpha = B_\alpha = H_\alpha := \log((1 - \alpha)/\alpha)$. The waiting penalty is $g(w) = C_w w^\rho$, with $C_w > 0$ and $\rho \geq 1$. Here, C_w is the coefficient denoted by c in Assumption 3.

E.1. Reconstruction of Waiting Time Distribution

The raw query-level waiting times are not observable. We therefore reconstructed the waiting-time distribution of each LLM using an equal mixture of ten Gamma distributions, each parameterized to match one of the ten available repeat-specific average waiting times and variances. At each query to a given LLM, one of the ten Gamma components is selected with equal probability and a waiting time is sampled from that component. This construction preserves positivity and incorporates the repeat-specific average waiting times, variances, and between-repeat variability. Because the reconstruction is based on summary statistics rather than raw query-level observations, it should be interpreted as a moment-matched reconstruction rather than an empirical bootstrap.

The Gamma-mixture reconstruction is used only for the numerical evaluation. Because an untruncated Gamma distribution is not sub-Gaussian, this reconstructed waiting time model does not directly satisfy Assumption 2. The simulation should therefore be interpreted as an empirical robustness evaluation under a positive, moment-matched waiting time distribution rather than as a numerical instantiation of the theoretical tail assumption.

E.2. Definition of the Bellman benchmark

At LLR state L , the posterior probability of hypothesis A is

$$p_A(L) = \frac{\exp(\delta + L)}{1 + \exp(\delta + L)}.$$

Conditional on querying LLM j , the predictive probability of observing output A is

$$q_j(L) = p_A(L)\gamma_{j,A} + [1 - p_A(L)](1 - \gamma_{j,B}).$$

The two terms correspond respectively to a correct output of A when $\theta = A$ and an incorrect output of A when $\theta = B$.

The deterministic-latency benchmark sets $T_{j,t} \equiv \mu_j$. At state (L, w) , w is the delay that has already accumulated. Define $V_\alpha(L, w)$ as the minimum expected sum of *future* monetary query costs and the waiting penalty assessed on the *total* delay at stopping. Previously paid query costs are excluded, but previously incurred delay remains relevant through the terminal penalty. Thus, for $-B_\alpha < L < A_\alpha$,

$$V_\alpha(L, w) = \min_{j \in [M]} \left\{ c_j + q_j(L)V_\alpha(L + \ell_j(A), w + \mu_j) + [1 - q_j(L)]V_\alpha(L + \ell_j(B), w + \mu_j) \right\}, \quad (\text{E.22})$$

with a terminal condition

$$V_\alpha(L, w) = g(w) \quad \text{if } L \geq A_\alpha \text{ or } L \leq -B_\alpha.$$

The minimum is over all models. Although each query has a deterministic duration in this benchmark, the outputs, subsequent model choices, stopping round, and total accumulated delay can still be random. The benchmark therefore retains uncertainty in evidence acquisition while removing the additional randomness in each query's duration.

Equivalently, define the transformed value function

$$U_\alpha(L, w) = V_\alpha(L, w) - g(w).$$

Its terminal value is zero, and the incremental waiting cost associated with querying LLM j is

$$g(w + \mu_j) - g(w).$$

For $g(w) = C_w w^\rho$, this increment becomes

$$g(w + \mu_j) - g(w) = C_w [(w + \mu_j)^\rho - w^\rho].$$

These increments telescope to the final waiting penalty minus $g(w)$ at the initial state of the recursion. Thus, the transformation subtracts a state-dependent constant, preserves the minimizing model choices, and uses a zero terminal value without charging the waiting penalty twice.

When $\rho = 1$, the incremental waiting cost reduces to $C_w \mu_j$ and is independent of the accumulated waiting time w . Therefore, w need not be included in the state, and the transformed value function satisfies the single-dimensional Bellman equation

$$U_\alpha(L) = \min_{j \in [M]} \left\{ c_j + C_w \mu_j + q_j(L)U_\alpha(L + \ell_j(A)) + [1 - q_j(L)]U_\alpha(L + \ell_j(B)) \right\}, \quad (\text{E.23})$$

with a terminal condition

$$U_\alpha(L) = 0 \quad \text{whenever } L \geq A_\alpha \text{ or } L \leq -B_\alpha.$$

The initial benchmark values are $U_\alpha(0)$ for the one-dimensional formulation and $V_\alpha(0,0) = U_\alpha(0,0)$ for the augmented-state formulation. The reported quantity $\mathcal{R}_{\text{Bellman}}^{\text{det}}$ in Section 7 is their numerical approximation. For the $\rho = 1$, varying- C_w experiment, the LLR grid has spacing $\Delta L = 0.0025$, and no waiting-time grid or truncation is used. For the varying- ρ experiment with $1 \leq \rho \leq 2$, the augmented-state calculation uses approximately 4,001 LLR grid points, $\Delta w = 0.125$ seconds, and $W_{\max} = 500$ seconds. Its LLR grid includes both stopping boundaries and $L = 0$ exactly; the resulting spacing ranges from approximately 0.00110 to 0.01151 over $x = \log_{10}(1/\alpha) \in \{1, \dots, 10\}$. The truncation level is a Bellman-grid parameter, not a posterior stopping boundary.