Asymptotically Optimal Sequential Testing with Heterogeneous LLMs

Guokai Li
guokaili@link.cuhk.edu.cn

Jiaxin (Alys) Liang
alys.liang@mcgill.ca

Mo Liu
mo*l**iu*@*unc.edu*

Yanzhe (Murray) Lei
yl64@queensu.ca

Stefanus Jasin, Fenghua Yang, Preet Baxi
sjasin, yfenghua, preetb@umich.edu

We study a Bayesian binary sequential hypothesis testing problem with multiple large language models (LLMs). Each LLM j has per-query cost $c_j > 0$, random waiting time with mean $\mu_j > 0$ and sub-Gaussian tails, and *asymmetric* accuracies: the probability of returning the correct label depends on the true hypothesis $\theta \in \{A, B\}$ and needs not be the same under A and B. This asymmetry induces two distinct information rates $(I_{j,A}, I_{j,B})$ per LLM, one under each hypothesis. The decision-maker chooses LLMs sequentially, observes their noisy binary answers, and stops when the posterior probability of one hypothesis exceeds $1-\alpha$. The objective is to minimize the sum of expected query cost and expected waiting cost, $\mathbb{E}[C_\pi] + \mathbb{E}[g(W_\pi)]$, where C_π is the total query cost, W_π is the total waiting time and g is a polynomial function (e.g., $g(x) = x^\rho$ with $\rho \geq 1$). We prove that as the error tolerance $\alpha \to 0$, the optimal policy is asymptotically equivalent to one that uses at most two LLMs. In this case, a single-LLM policy is *not* generically optimal: optimality now requires exploiting a two-dimensional tradeoff between information under A and information under B. Any admissible policy induces an expected information-allocation vector in $\mathbb{R}^2_+$, and we show that the optimal allocation lies at an extreme point of the associated convex set when α is relatively small, and hence uses at most two LLMs. We construct belief-dependent policies that first mix between two LLMs when the posterior is ambiguous, and then switch to a single "specialist" LLM when the posterior is sufficiently close to one of the hypotheses. These policies match the universal lower bound up to a $(1 + o(1))$ factor as $\alpha \to 0$.

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. 2024, Wu et al. 2024). Agentic workflows route tasks through cascades of specialized models and tools, each invocation adding cost and latency (Wang et al.

2024). 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. 2024, Zhang et al. 2024, Huang et al. 2025). 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 concerns such as heterogeneous costs, random latency, and asymmetric diagnostic power. These practical considerations are critical when optimizing system performance. In Figure 1 and 2, we show how different model may have different output speed and Intelligence Index[1] (a score computed based on 10 commonly used benchmarks).

[1] Source: Artificial Analysis, `https://artificialanalysis.ai/`

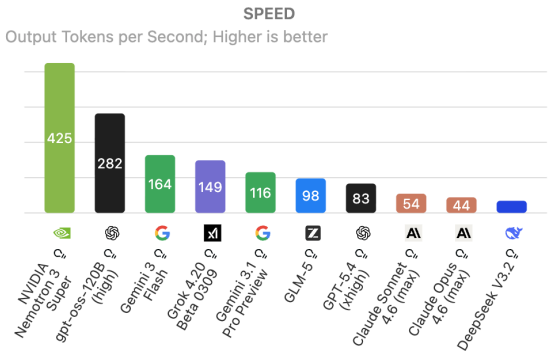

Figure 1. Output Speed of Different LLMs

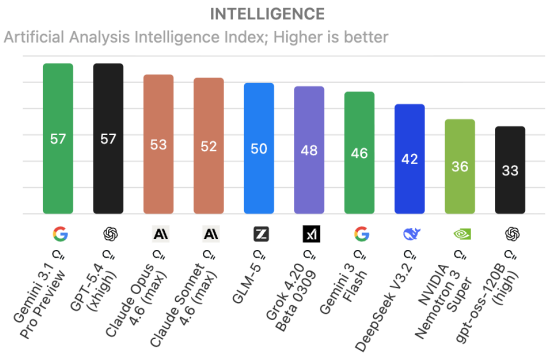

Figure 2. Intelligence Index of Different LLMs

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 response time (i.e., latency) with a mean μ_j, and a pair of asymmetric accuracies $(\gamma_{j,A}, \gamma_{j,B})$ with $\gamma_{j,\theta} \in (0.5, 1)$ 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 cost 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 were able to conduct a sharp asymptotic analysis in the high-confidence regime where error tolerance α approaches 0. 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 model 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-sources policy.

Taken together, these results contribute to the emerging intersection of Operations Research 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 latency, and stopping decisions within one unified objective. In contrast with most existing work on test-time compute in computer science, we were 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 there exists an asymptotically optimal policy that relies on at most two sources, regardless of how many are available. This is a strong and actionable design principle. A near-optimal system does not need to orchestrate a large collection of models in a complicated way; it needs to identify one source that is especially effective for each source, 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 models.

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 (2022) 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 latency 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. 2026). 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. 2024, Huang et al. 2025, 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. 2024, Setlur et al. 2024). 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. 2022, 2023, Gou et al. 2023, Du et al. 2023). 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 guaranties for the proposed policy.

Lastly, our work also lies at the intersection of Operations Research (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. 2025, 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. Among these, Huang et al. (2026) is the closest technical neighbor. They study Bayesian stopping for repeated sampling from a single LLM in order to identify the most consistent answer under informative priors. Our paper, instead, studies a binary hypothesis testing problem with multiple heterogeneous sources, each with its own diagnostic power, monetary cost, and stochastic response time. 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, including the sequential decision framework, cost structure, and key information-theoretic quantities. Section 3 presents the sign-based two-LLM policy and provides an overview of its performance. Section 4 establishes the asymptotic optimality results through a matching lower- and upper-bound analysis. Section 5 concludes with a discussion of the main insights and implications for the design of efficient AI systems.

2. Model Setup

In many applications such as automated content moderation and fraud screening, a decision-maker must make a binary decision $D \in \{A, B\}$ (e.g., escalate vs. resolve) with access to M different decision-support tools or information sources (e.g., multiple LLMs), indexed by $j = 1, 2, \ldots, M$. For consistency, we use the term "LLM" to refer to an information source. Let $[M] := \{1, 2, \ldots, M\}$ be the set of these available LLMs. We assume there is a single ground truth $\theta \in \{A, B\}$ unknown to the decision-maker. Prior to querying any LLM, the decision maker holds a belief represented by a random variable $\theta \in \{A, B\}$ with prior

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

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 each query $t = 1, 2, \ldots,$ an LLM is selected based on past observations and returns a binary output $Y_t \in \{A, B\}$. We denote the index of the selected LLM by $j_t \in [M]$. Naturally, querying LLMs is costly. In particular, each query incurs a per-use cost and a random waiting time $T_{j_t,t}$. We assume each LLM $j \in [M]$ is characterized by three sets of parameters:

1. A per-query monetary cost $c_j > 0$.

2. The accuracies $\gamma_{j,A}, \gamma_{j,B} \in (1/2, 1)$: Under $\theta = A$, the model returns A with probability $\gamma_{j,A}$; under $\theta = B$, it returns B with probability $\gamma_{j,B}$.
3. A nonnegative waiting time $T_{j,t}$ captures the response latency (e.g., server congestion or network delay). We assume $\mathbb{E}[T_{j,t}] = \mu_j \in (0, \infty)$ is known and that $T_{j,t}$ is sub-Gaussian (see Assumption 2).

The parameters $(\gamma_{j,A}, \gamma_{j,B})$ allow for asymmetric accuracy. This captures the scenarios where an LLM may be more reliable when the ground truth is A than when it is B, or vise versa. This asymmetry reflects the fact that some models are conservative and excel at detecting "negative" cases (e.g., fraud or policy violations) while being less precise on "positive" ones. As a consequence, each LLM induces two distinct information rates under each hypothesis. The objective of the decision-maker is to reach a satisfactory decision while accounting for both the query cost and latency. In the next section, we introduce the key components of our model.

2.1. Admissible Policy and Information Gain under Posterior Dynamics

Throughout this paper, we focus on the non-anticipative policy π that stops once the posterior probability of one hypothesis exceeds $1 - \alpha$ for some fixed α. The total query times can be written as

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

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

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

We let $\mathcal{G}_t := \sigma\big(\{j_s, Y_s, T_{j_s,s}\}_{s=1}^{t}\big)$ denote the observable σ-field at time t. Under such policy, the decision-maker achieves the error restriction

$$\mathbb{P}(D \neq \theta \mid \mathcal{G}_\tau) \leq \alpha. \tag{3}$$

This model is analogous to Wald's sequential probability ratio test (SPRT), but with asymmetric information increments and waiting-time costs. To explicitly characterize the evolution of information gain, we start by defining the log-likelihood ratio for a single LLM. We 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 log-likelihood ratio of $\theta = A$ versus $\theta = B$ can be written as

$$\ell_j(y) := \log \frac{\mathbb{P}(Y = y \mid \theta = A, j)}{\mathbb{P}(Y = y \mid \theta = B, j)} = \begin{cases} \log \dfrac{\gamma_{j,A}}{1 - \gamma_{j,B}}, & \text{if } y = A, \\ \log \dfrac{1 - \gamma_{j,A}}{\gamma_{j,B}}, & \text{if } y = B. \end{cases} \tag{4}$$

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 the *cumulative log-likelihood ratio (LLR) up to* t can be written as

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

For ease of exposition, we let $\mathbb{P}_\theta(\,\cdot\,) := \mathbb{P}(\,\cdot \mid \theta)$ and $\mathbb{E}_\theta[\cdot] := \mathbb{E}[\cdot \mid \theta]$ for $\theta \in \{A, B\}$. We also keep the Bayes measure $\mathbb{P}$ and expectation $\mathbb{E}$ with respect to the prior (ξ_A, ξ_B), so that

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

For later use, we introduce some global bounds for the increments of L_t. Define

$$C_\ell := \max_{1 \le j \le M} \max\left\{|\ell_j(A)|, |\ell_j(B)|\right\} < \infty \tag{5}$$

$$v_\ell^2 := \max_{1 \le j \le M} \max\{\mathrm{Var}_A(\ell_j(Y)), \mathrm{Var}_B(\ell_j(Y))\} < \infty. \tag{6}$$

where the finiteness of C_ℓ and v_ℓ^2 follows from the fact that each ℓ_j takes only two finite values (for outputs A and B) and the accuracies $\gamma_{j,A}, \gamma_{j,B} \in (1/2, 1)$ are bounded away from 0 and 1.

Using L_t, we can interpret stopping rule via a threshold structure. We state a lemma.

Lemma 1. *Suppose the output realization only depends on the chosen LLM and the ground truth* θ*, then for each time* $t = 1, 2, ...,$ *we have*

$$\mathbb{P}(\theta = A \mid Y_{1:t}, j_{1:t}) = \frac{e^{\delta + L_t}}{1 + e^{\delta + L_t}}, \qquad \mathbb{P}(\theta = B \mid Y_{1:t}, j_{1:t}) = \frac{1}{1 + e^{\delta + L_t}}. \tag{7}$$

Moreover, the stopping time and the decision rule correspond to an equivalent form bearing a threshold structure

$$\tau := \inf\{t \ge 0 : L_t \ge A_\alpha \text{ or } L_t \le -B_\alpha\}, \qquad D := \begin{cases} A, & \text{if } L_\tau \ge A_\alpha, \\ B, & \text{if } L_\tau \le -B_\alpha, \end{cases} \tag{Threshold Decision Rule}$$

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

The intuition of the threshold policy is that, the query of the LLMs will continue until sufficient evidence has been accumulated towards a hypothesis. We now formally define admissible policies.

Definition 1 (Admissible Policy). For given $\alpha > 0$, a non-anticipative policy $\pi := \{\phi_t\}_{1 \le t \le \tau}$ consists of a sequence of LLM selection rules $\phi_1, \phi_2, \ldots, \phi_\tau$, where ϕ_t is $\mathcal{G}_{t-1}$-measurable and $j_t := \phi_t(\mathcal{G}_{t-1}) \in \{1, \ldots, M\}$, and where τ is the stopping time defined in (Threshold Decision Rule). We let $\Pi(\alpha)$ denote the set of all admissible policies.

2.2. Objective: Query Cost and Latency

The objective of the decision maker is to achieve the accuracy threshold $1-\alpha$ at the lowest expected cost. Suppose one such policy π has total query time τ, then the total query cost and total waiting time can be written as

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

To characterize the waiting time penalty, we let $g : \mathbb{R}_+ \to \mathbb{R}_+$ be convex, increasing, and satisfying $g(0) = 0$. The penalty (i.e., cost) of a policy π at error tolerance α can then be defined as

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

The expectation is taken over four sources of randomness: (i) the prior on θ, (ii) the realized outputs $Y_{1:\tau}$, (iii) the LLM selection path $j_{1:\tau}$, and (iv) the stopping time τ.

Our objective is to minimize $\mathcal{R}(\pi;\alpha)$ by striking a trade-off between information gain and penalty. 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 function $\mathcal{R}(\pi,\alpha)$ for a given α. To this end, we define the key information-theoretic quantities for each LLM. Intuitively, each additional query will provide more information of the ground truth. To quantify the information gain, we associate each LLM $j \in [M]$ with a *pair* of expected information gain under different θ:

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

which can be interpreted as expected per-query information or information rates. Observe that both $I_{j,A}$ and $I_{j,A}$ are strictly positive because $\gamma_{j,A}, \gamma_{j,B} \in (1/2, 1)$. Such definition is consistent with the KL divergence. In fact, if we let $P_{j,A}$ denote the distribution of the output Y under $(\theta = A, j)$, and $P_{j,B}$ the distribution under $(\theta = B, j)$. Then

$$I_{j,A} = \mathrm{KL}(P_{j,A} \,\|\, P_{j,B}) > 0, \qquad I_{j,B} = \mathrm{KL}(P_{j,B} \,\|\, P_{j,A}) > 0,$$

where $\mathrm{KL}(P\|Q)$ is the Kullback–Leibler divergence.

Remark 1. Throughout the paper, we only need to focus on policies satisfying $\mathbb{E}[\tau] < \infty$, as $\mathbb{E}[\tau] = \infty$ cannot happen under optimality. To see this, let $\mathbb{E}_A[\tau]$ and $\mathbb{E}_B[\tau]$ denote the expected number of queries when the ground truth is A and B, respectively. If under a policy π either $\mathbb{E}_A[\tau] = \infty$ or $\mathbb{E}_B[\tau] = \infty$ holds, then

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

which implies that π is strictly suboptimal. Hence, any optimal policy must satisfy $\mathbb{E}_A[\tau] < \infty$ and $\mathbb{E}_B[\tau] < \infty$. In particular, this implies that $\tau < \infty$ almost surely under optimality.

2.3. Assumptions and Discussion

In this paper, we impose the following independence assumptions on outputs and latency.

ASSUMPTION 1 **(Independence)**. *We assume:*

(i) OUTPUT INDEPENDENCE. *Conditional on the true hypothesis θ and the sequence of LLM selections $(j_t)_{t\geq 1}$, the outputs $(Y_t)_{t\geq 1}$ are independent across queries, with*

$$P(Y_t = y \mid \theta, j_t) = P(Y = y \mid \theta, j_t), \qquad y \in \{A, B\},\ t \geq 1.$$

(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 μ_j and finite variance.*

(iii) SEPARATION OF INFORMATION AND LATENCY. *Conditional on the selected LLM, waiting times are independent of both the true hypothesis and observed outputs: $(T_{j_t,t} \mid j_t) \perp\!\!\!\perp (\theta, Y_{1:t})$, $t \geq 1$.*

(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 sub-gaussian property imposed on the waiting-time distributions.

ASSUMPTION 2 **(Sub-Gaussian Waiting Times)**. *For each LLM j, the waiting times $T_{j,t}$ are sub-Gaussian. That is, there exists a constant $\psi > 0$ such that for all $\lambda \in \mathbb{R}$,*

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

Assumption 2 guaranties that the total waiting time concentrates around its mean under any policy. To enable our asymptotic analysis in Section 4, we introduce a mild regularity assumptions on the waiting-time cost function $g(\cdot)$.

ASSUMPTION 3 **(Polynomial Waiting Cost)**. *We assume that $g(x) = c \cdot x^\rho$ is a polynomial function with $c \geq 0$ and $\rho \geq 1$.*

Assumption 3 captures the representative class of convex increasing functions that characterize the waiting time penalty.

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 $\alpha \to 0$. 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.

2.4. Notation Table

The above completes the description of the model and the basic information-theoretic quantities. In Table 1, we provide a summary of the notations.

In the next section we will discuss our policy which is asymptotically optimal within the regime $\alpha \to 0$. We will also highlight the key insights into the policy design.

3. The Two-LLM Policy

According to the decision rule in Section 2.1, the query sequence proceeds until a sufficient amount of evidence has been collected in favor of either A or B and this evidence is measured by the cumulative LLR. Because the LLR adds up across queries, each LLM contributes linearly to it through its information rates $I_{j,A}$ and $I_{j,B}$. Consequently, any policy can be viewed as collecting evidence in favor of either A or B across the available LLMs.

Among all admissible policies, we highlight a particularly simple class of sign-based two-LLM sequential policies that use at most two LLMs, denoted by j_A and j_B. These can be interpreted as an A-specialist and a B-specialist. The A-specialist j_A is better suited for confirming hypothesis A and tends to deliver more reliable support for A when $\theta = A$. Likewise, the B-specialist j_B is better

Variables	Definitions
M	Number of LLMs
θ	Ground-truth hypothesis
ξ_y	Prior belief of hypothesis $y \in \{A, B\}$
$\delta = \log(\xi_A/\xi_B)$	Prior log-odds in favor of A
$\gamma_{j,y}$	Accuracy of LLM $j \in [M]$ when $\theta = y \in \{A, B\}$
$T_{j,t}$	Random waiting time of LLM j in period t
μ_j	Expected waiting time of LLM j
c_j	Per-query cost of LLM j
$\ell_j(y)$	Single-sample log-likelihood ratio of $\theta = A$ versus $\theta = B$
L_t	Cumulative log-likelihood up to round t
$A_\alpha = \log((1-\alpha)/\alpha) - \delta$	Upper stopping threshold of L_t
$B_\alpha = \log((1-\alpha)/\alpha) + \delta$	Lower stopping threshold of L_t
$S_\alpha = \xi_A A_\alpha + \xi_B B_\alpha$	Weighted average threshold
$I_{j,y}$	Expected log-likelihood ratio drift under LLM j when $\theta = y \in \{A, B\}$
$\kappa_{j,y} = c_j/I_{j,y}$	Cost-efficiency index under LLM j when $\theta = y \in \{A, B\}$
$\eta_{j,y} = \mu_j/I_{j,y}$	Waiting-time-efficiency index under LLM j when $\theta = y \in \{A, B\}$

Table 1 Summary of notations

suited for confirming hypothesis B and tends to deliver more reliable support for B when $\theta = B$. This two-LLM structure fits with asymmetric accuracies and allows the policy to route queries toward the model that best matches the current direction of evidence.

We now describe a policy induced by a pair (j_A, j_B) for a given error tolerance threshold α, denoted by $\pi^{(2)}_{j_A,j_B}(\alpha)$. The decision-maker tracks the cumulative log-likelihood ratio LLR and stops once L_t exceeds either the upper threshold A_α or the lower threshold $-B_\alpha$. The key idea of the two-LLM policy is simple: the choice of LLM at each step depends on the sign of the current LLR, so queries are directed toward the specialist aligned with the current direction of evidence.

Definition 2 (Sign-based two-LLM Sequential Policy). Fix an error tolerance α and a pair of LLMs (j_A, j_B). The policy $\pi^{(2)}_{j_A,j_B}(\alpha)$ proceeds as follows.

- *Initialization:* Set $L_0 = 0$.
- *LLM selection and update:* For each $t \geq 1$, choose

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

  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 first time

$$\tau := \inf\{t \geq 1 : L_t \geq A_\alpha \text{ or } L_t \leq -B_\alpha\},$$

  and set

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

Such policy has a simple interpretation. When the accumulated evidence (i.e., L_t) favors A, the policy queries the LLM that is most efficient at gathering information in favor of A; when the evidence favors B, it queries the LLM that is most efficient for certifying B. Early in the process, when the posterior is close to the prior and L_t fluctuates around zero, the policy naturally alternates between the two specialists, effectively mixing between them. intuitively, as the belief becomes more decisive, the policy commits to a single specialist and behaves like a single-LLM sequential test near the boundary. This behavior mirrors how a human would proceed in acquiring evidence: When unsure, consult multiple viewpoints; once a hypothesis becomes more likely, focus on the tool best suited to confirm it.

3.1. Performance Analysis

Despite its simplicity, we are able to show that this policy is asymptotically optimal in the high confidence regime $\alpha \to 0$. To facilitate our analysis, we first establish a universal lower bound for the performance under any admissible policy. We state an informal theorem.

INFORMAL THEOREM 1. *There exists some $\Phi_\alpha > 0$ depending only on the model primitives and α, such that $\mathcal{R}(\pi;\alpha) \geq \Phi_\alpha$ for any $\pi \in \Pi(\alpha)$. Moreover,*

$$\Phi_\alpha = \Theta\Big((\log(1/\alpha))^\rho\Big).$$

The result identifies a universal lower bound, which serves as a natural benchmark for evaluating any sequential policy in the regime $\alpha \to 0$. For $\rho = 1$, the benchmark grows on the information scale $\log(1/\alpha)$, while for $\rho > 1$, the convex delay penalty dominates and leads to the higher-order growth $\log(1/\alpha))^\rho$.

The benchmark Φ_α is constructed by solving a deterministic convex program that can be interpreted as a query allocation problem:

$$\Phi_\alpha := \min_{n^A, n^B \in \mathcal{U}} F_\alpha(n^A, n^B) \tag{8}$$

where

$$F_\alpha(n^A, n^B) := \underbrace{\left(\xi_A \sum_{j=1}^M c_j n_{j,A} + \xi_B \sum_{j=1}^M c_j n_{j,B}\right)}_{\text{query cost}} + \underbrace{\xi_A g\left(\sum_{j=1}^M \mu_j n_{j,A}\right) + \xi_B g\left(\sum_{j=1}^M \mu_j n_{j,B}\right)}_{\text{waiting time penalty}} \tag{9}$$

consists of a *query cost* component and a *waiting time* component. The decision variables are $n^A := (n_{1,A}, \ldots, n_{M,A})$ and $n^B := (n_{1,B}, \ldots, n_{M,B})$ with each pair $(n_{j,A}, n_{j,B})$ representing the expected total query times of LLM j under different hypotheses for any given admissible policy π. The feasible region, denoted as $\mathcal{U}$, characterizes the LLR requirements specified in Lemma 1 under any given admissible policy π.

Intuitively, the convex problem can be interpreted as minimizing a deterministic lower bound of the expected cost over all admissible policies by allocating query time across different LLMs under each hypothesis. Therefore, by solving this convex problem, we obtain a universal lower bound and also a benchmark query allocation over the LLMs by identifying the optimal solution n^A, n^B. (We defer a more detailed discussion to Section 4.) We further show that there exists an optimal allocation rule that involves at most two LLMs, i.e., $(n^A, n^B) = (e_{j_A^\star}, e_{j_B^\star})$ for some $j_A^\star, j_B^\star \in [M]$ where each e_j denotes a vector with a single nonzero element in the j-th position. This sparse structure motivates us to investigate the performance of the Policy $\pi^{(2)}_{j_A^\star, j_B^\star}$. Indeed, we are able to show that this deterministic benchmark Φ_α is asymptotically attainable under $\pi^{(2)}_{j_A^\star, j_B^\star}$ in the high-confidence regime $\alpha \to 0$. The following theorem, which is the most important result of the paper, characterizes the best possible convergence rate to Φ_α and establishes its tightness.

THEOREM 1 **(Asymptotic optimality and tightness of the convergence rate)**. *There exists an LLM pair $(j_A^\star, j_B^\star) \in [M] \times [M]$ such that the policy $\pi := \pi^{(2)}_{j_A^\star, j_B^\star}(\alpha)$ satisfies, as $\alpha \to 0$,*

$$\mathbb{E}[C_\pi] + \mathbb{E}[g(W_\pi)] = \Phi_\alpha \, + \, O\Big((\log(1/\alpha))^{\rho-1}\Big). \tag{10}$$

Moreover, this remainder order is tight in general: letting $\pi^\star(\alpha) \in \arg\min_{\pi \in \Pi(\alpha)} \mathcal{R}(\pi; \alpha)$ denote an optimal admissible policy, as $\alpha \to 0$, we have

$$\mathcal{R}(\pi^\star(\alpha); \alpha) \, = \, \Phi_\alpha + \Omega\Big((\log(1/\alpha))^{\rho-1}\Big). \tag{11}$$

Theorem 1 establishes that our sign-based two-LLM policy is asymptotically optimal as the target error level $\alpha \to 0$. In particular, the remainder order of the proposed policy is also tight: even an optimal admissible policy cannot typically reduce the gap to Φ_α faster than $\Omega((\log(1/\alpha))^{\rho-1})$.

These conclusions have both theoretical and operational implications. Theoretically, they reveal a simple structure in the high confidence regime. Although the decision-maker may have access to many heterogeneous LLMs, optimal performance can be achieved by using at most two specialists, one that is efficient for confirming A and one that is efficient for confirming B. The complexity of sequential hypothesis experimentation, therefore, does not grow with the number of available LLMs. Operationally, the theorem gives 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 but does not require more complex orchestration. As a result, the policy remains simple and scalable under stringent reliability requirements.

4. Asymptotic Optimality of Two-LLM Policy

Despite the simplicity of the policy structure, establishing its asymptotic optimality is nontrivial. The technical difficulty arises from two key features that distinguish our setting from classical sequential testing models. First, in standard formulations of dynamic hypothesis testing or optimal stopping, 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. At each round, 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 not in learning unknown parameters but in optimally orchestrating multiple information channels 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 delay is penalized through a convex function of the total waiting time, which itself is a *random* sum of random waiting times. This nonlinearity prevents a direct application of standard optimal stopping or dynamic programming arguments. In particular, to obtain sharp asymptotic bounds as $\alpha \to 0$, we must carefully control the distribution of the stopping time and the accumulated waiting time, rather than only their expectations.

Our proof follows a matching-bounds approach centered around the deterministic universal lower bound Φ_α discussed in Section 3.1. We will show that as $\alpha \to 0$, the expected query cost and the expected waiting time penalty induced by $\pi^{(2)}_{j_A^\star, j_B^\star}$ match the *query cost* and *waiting time* components within $F(e^\star_{j_A}, e^\star_{j_B})$ (which equals Φ_α) up to lower-order terms, respectively. This alignment further implies that the expected risk under this policy matches Φ_α up to a lower-order term, yielding the asymptotic optimality of the policy.

There are three essential steps in establishing the results. We first translate the stopping requirement on the posterior LLR in Lemma 1 into deterministic linear constraints on (n^A, n^B). This enables us to construct two information budget constraints for the convex program underlying Φ_α. Moreover, these two constraints create a two-dimensional allocation structure that, through the KKT conditions, imply a sparse optimal solution and provide the key insight for why at most two LLMs are needed in the benchmark allocation. The detailed analysis is described in Section 4.1.

In the second step, we match the expected query cost and waiting time penalty induced by $\pi^{(2)}_{j_A^\star, j_B^\star}$ to their counterparts in Φ_α. The major challenge lies in matching the waiting time penalty $\mathbb{E}[g(W_\pi)]$ to its benchmark in $F(e^\star_{j_A}, e^\star_{j_B})$: Since g is convex and increasing, it may amplify the

randomness in the waiting time. We use the drift and martingale decomposition of the cumulative log likelihood ratio. Under $\theta \in \{A, B\}$,

$$L_t = \sum_{s=1}^{t} I_{j_s,\theta} + M_t^{(\theta)}, \tag{12}$$

where $I_{j_s,\theta} = \mathbb{E}_\theta[\ell_{j_s}(Y_s)]$ is the per query information rate and $M_t^{(\theta)}$ is a martingale 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. We then apply Freedman type concentration bounds to (M_t) and combine them with the structural assumptions on g (Assumption 3) to show that $\mathbb{E}[g(W_\pi)]$ only deviates from its deterministic benchmark by lower-order terms in the high confidence regime $\alpha \to 0$. The detailed analysis is described in Section 4.2

Finally, we show that the performance cannot be improved further by any admissible policy by analyzing an oracle policy that is still required to satisfy the same LLR stopping requirement but is informed of the true hypothesis $\theta \in \{A, B\}$ in advance. Because this oracle has strictly more information than any admissible policy, the minimal cost achievable by such an oracle provides a lower bound on the cost of the best admissible policy. We then show that the cost attained by the optimal oracle policy matches Φ_α up to a remainder term of the same order as that achieved by $\pi^{(2)}_{j_A^\star, j_B^\star}$. This establishes that our remainder order is tight, and hence no admissible policy can asymptotically improve upon the performance of the two-LLM policy in the high confidence regime.

4.1. A Deterministic Lower Bound

In this subsection, we detail the key results used to establish the universal lower bound Φ_α and the sparsity property of the optimal solution. In Section 4.1.1, we derive two information budget constraints that any policy π must satisfy in the high confidence regime. These constraints are implied by the posterior or LLR stopping requirement and are expressed in terms of the hypothesis dependent expected query allocations induced by π. Then, in Section 4.1.2, we construct a deterministic convex function, also expressed in terms of these hypothesis dependent expected query allocations, that lower bounds the expected cost incurred by π. This leads to the convex program whose optimal value is Φ_α, and whose KKT characterization further reveals the sparsity structure of the optimal allocation.

4.1.1. LLR requirements under asymmetry To connect the LLR requirements to the query allocation, for each policy, we let $N_j(\tau)$ denote the total number of queries sent to LLM j before stopping:

$$N_j(\tau) := \sum_{t=1}^{\tau} \mathbf{1}_{\{j_t = j\}}, \tag{13}$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function and τ is the stopping time. Correspondingly, we let

$$n_{j,A} := \mathbb{E}_A[N_j(\tau)], \qquad n_{j,B} := \mathbb{E}_B[N_j(\tau)] \tag{14}$$

be the expected query times of LLM j under $\theta = A$ and $\theta = B$ respectively. And the Bayes-averaged usage count for LLM j is defined as

$$\bar{N}_j := \xi_A \cdot n_{j,A} + \xi_B \cdot n_{j,B}. \tag{15}$$

Using these notations, we can express the expected LLR upon stopping time τ as

$$\mathbb{E}_A[L_\tau] = \sum_{j=1}^{M} I_{j,A}\mathbb{E}_A[N_j(\tau)] = \sum_{j=1}^{M} I_{j,A} n_{j,A}. \tag{16}$$

with the same logic applied to hypothesis B. Thus, we connect the LLR requirements to the query allocations. The next lemma concretizes this intuition and show that the expected information gathered from the LLMs must be at least over a certain threshold.

LEMMA 2. *Suppose an admissible policy has a stopping time τ satisfying $\tau < \infty$ almost surely under both $\mathbb{P}_A$ and $\mathbb{P}_B$, and suppose $\mathbb{E}_A[\tau] < \infty$ and $\mathbb{E}_B[\tau] < \infty$. Then there exist constants c_{err} (independent of α and of the policy) such that*

$$\sum_{j=1}^{M} I_{j,A} n_{j,A} \geq A_\alpha - K_\alpha, \tag{17}$$

$$\sum_{j=1}^{M} I_{j,B} n_{j,B} \geq B_\alpha - K_\alpha, \tag{18}$$

where $K_\alpha := c_{\text{err}}\,\alpha\left(A_\alpha + B_\alpha + C_\ell\right)$. Moreover, $K_\alpha = O(1)$.

Observe that in Lemma 1, the effective thresholds are

$$S_A(\alpha) := A_\alpha - K_\alpha, \qquad S_B(\alpha) := B_\alpha - K_\alpha, \tag{19}$$

rather than A_α and B_α. The slack term K_α accounts for adverse stopping events in which the LLR crosses the wrong threshold, for example crossing the A boundary when $\theta = B$. The choice of K_α ensures that the resulting information budget constraints hold uniformly for all admissible policies.

According to Lemma 2, any admissible policy must satisfy

$$\sum_{j=1}^{M} I_{j,A} n_{j,A} \geq S_A(\alpha), \qquad \sum_{j=1}^{M} I_{j,B} n_{j,B} \geq S_B(\alpha). \tag{20}$$

which are the information requirements under each hypothesis. These two requirements serve as the linear constraints in the convex program used for constructing the universal lower bound Φ_α.

4.1.2. Lower bound for objective function To establish the convex program, we also need to write the objective function in terms of $n_{j,A}, n_{j,B}$ that serves as a universal lower bound to the cost function. We state a lemma.

LEMMA 3. *For any policy π with $\mathbb{E}_\theta[\tau] < \infty$ for $\theta \in \{A, B\}$,*

$$\mathbb{E}_\theta[C_\pi] = \sum_{j=1}^{M} c_j \cdot n_{j,\theta}, \quad \mathbb{E}_\theta[W_\pi] = \sum_{j=1}^{M} \mu_j \, n_{j,\theta} \tag{21}$$

Under the Bayes measure,

$$\mathbb{E}[C_\pi] = \sum_{j=1}^{M} c_j \, \bar{N}_j, \quad \mathbb{E}[W_\pi] = \sum_{j=1}^{M} \mu_j \, \bar{N}_j. \tag{22}$$

Then using Jensen's inequality, we can further develop a lower bound for the waiting time penalty.

PROPOSITION 1. *Under any admissible policy π, we have*

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

where $n_{j,\theta}$ is the expected query times of LLM j under hypothesis θ.

Using Lemma 3 and Proposition 1, we are now able to construct a lower bound for the expected cost under any feasible policy π:

$$F_\alpha(n^A, n^B) := \xi_A \sum_{j=1}^{M} c_j n_{j,A} + \xi_B \sum_{j=1}^{M} c_j n_{j,B} + \xi_A g\left(\sum_{j=1}^{M} \mu_j n_{j,A}\right) + \xi_B g\left(\sum_{j=1}^{M} \mu_j n_{j,B}\right). \tag{23}$$

One key observation is that $F_\alpha(n^A, n^B)$ is convex in $\{n_j^A, n_B^j\}_{j\in[M]}$.

4.1.3. Convex program and lower bound Using Lemma 2 and Proposition 1, we are able to construct the convex optimization (LB) define below. Note that under any admissible policy π, the expected query times satisfy both linear constraints in (LB) and the objective function is a lower bound of the cost function. Thus, the optimal objective value of (LB) serves as a universal lower bound of the penalty incurred by any admissible policy.

$$\begin{aligned} \min \quad & F_\alpha(n^A, n^B) \\ \text{subject to} \quad & \sum_{j=1}^{M} I_{j,A} n_{j,A} \geq S_A(\alpha), \\ & \sum_{j=1}^{M} I_{j,B} n_{j,B} \geq S_B(\alpha), \\ & n_{j,A}, n_{j,B} \geq 0 \quad \forall j \in [M] \end{aligned} \tag{LB}$$

Furthermore, according to the definition of $F_\alpha(n^A, n^B)$, it is not difficult to show that, under optimality, the inequality constraints must be tight. We state a lemma.

LEMMA 4 **(Tightness of information constraints)**. *The optimal objective value of* (LB) *stays unchanged if the inequalities are replaced by equalities:* $\sum_{j=1}^{M} I_{j,A} n_{j,A} = S_A(\alpha)$ *and* $\sum_{j=1}^{M} I_{j,B} n_{j,B} = S_B(\alpha)$, *for any* $j \in [M]$.

Using this formulation, we can instead focus on the the convex problem with equality constraints, denoted as (LB')

$$\begin{aligned} \Phi_\alpha := \min \quad & F_\alpha(n^A, n^B) \\ \text{subject to} \quad & \sum_{j=1}^{M} I_{j,A} n_{j,A} = S_A(\alpha), \\ & \sum_{j=1}^{M} I_{j,B} n_{j,B} = S_B(\alpha), \\ & n_{j,A}, n_{j,B} \geq 0 \quad \forall j \in [M] \end{aligned} \tag{LB'}$$

Our goal is to show that the penalty incurred by the best two-LLM policy matches the universal lower bound Φ_α asymptotically. To conveniently compare the information gain with cost or waiting time, we define *information fractions*, *cost-efficiency* and *time-efficiency* for each $\theta \in \{A, B\}$ as

$$w_{j,\theta} := \frac{I_{j,\theta} \cdot n_{j,\theta}}{S_A(\theta)}, \qquad \kappa_{j,\theta} := \frac{c_j}{I_{j,A}}, \qquad \eta_{j,\theta} := \frac{\mu_j}{I_{j,B}}$$

where c_j is the per-query cost and μ_j is the mean of waiting time. Correspondingly, we define the average per-unit-information cost and waiting time under hypothesis $\theta \in \{A, B\}$:

$$\kappa^\theta(w) := \sum_{j=1}^{M} \kappa_{j,\theta} w_j, \qquad \tau^\theta(w) := \sum_{j=1}^{M} \eta_{j,\theta} w_j.$$

Using these definitions, we can rewrite the objective function of (LB') in terms of $w^A := (w_1^A, w_2^A, ..., w_M^A), w^B := (w_1^B, w_2^B, ..., w_M^B)$, which can be interpreted as the per

$$\begin{aligned} F_\alpha(n^A, n^B) = {} & \xi_A S_A(\alpha) \kappa^A(w^A) + \xi_B S_B(\alpha) \kappa^B(w^B) \\ & + \xi_A g\big(S_A(\alpha) \tau^A(w^A)\big) + \xi_B g\big(S_B(\alpha) \tau^B(w^B)\big) := G_\alpha(w^A, w^B). \end{aligned} \tag{24}$$

Thus (LB') can be further transformed to the following convex problem

$$\Phi_\alpha := \min G_\alpha(w^A, w^B) \quad \text{subject to} \quad (w^A, w^B) \in \Delta_M \times \Delta_M \tag{ALO}$$

where $\Delta_M := \left\{ w \in \mathbb{R}_+^M : \sum_{j=1}^{M} w_j = 1 \right\}$. Now, leveraging the Karush-Kuhn-Tucker (KKT) conditions underpinning the optimal solution of a convex programming, we are able to establish the universal lower bound corresponding to a two-LLM information allocation scheme.

THEOREM 2. *For any sufficiently small α we have $\mathbb{E}[C_\pi]+\mathbb{E}[g(W_\pi)] \geq \Phi_\alpha$ where for any admissible policy π. Moreover, there exists a finite constant $\bar{\alpha}>0$, such that when $\alpha<\bar{\alpha}$, we have*

$$\min_{(w^A,w^B)\in\Delta_M\times\Delta_M} G_\alpha(w^A,w^B) = \min_{(i,j)\in[M]\times[M]} G_\alpha(e_i,e_j),$$

where e_i is a vector is an all-zero vector except for the i-th entry being one. Furthermore, the optimal entries $i^\star$ and $j^\star$ satisfy

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

An important implication of Theorem 2 is that, for sufficiently small α, the convex problem admits an optimal solution that uses at most two LLMs, with one LLM active under $\theta=A$ and one LLM active under $\theta=B$.

To illustrate the intuition, let $(w^{A\star},w^{B\star})$ be an optimal solution to (ALO). Suppose there exist two distinct indices $j\neq k$ such that $w_j^{A\star}>0$ and $w_k^{A\star}>0$ (the discussion on w^{B*} will follow the same logic). The KKT stationarity conditions imply that these active coordinates must have the same marginal contribution to the objective, which yields

$$(\kappa_{j,A}-\kappa_{k,A})+g'(y^\star)\,(\eta_{j,A}-\eta_{k,A})=0. \tag{25}$$

Equation (25) provides a necessary condition that must be satisfied by any pair j,k that are both active in $w^{A\star}$. In particular, to satisfy this odentity, one of the following two cases must occur.

- **Case 1.** $\eta_{j,A}=\eta_{k,A}$ and $\kappa_{j,A}=\kappa_{k,A}$. In this case, the two LLMs are indistinguishable in both time efficiency and cost efficiency. Therefore, we may construct an alternative optimal solution to the convex problem by shifting the entire mass $w_j^{A\star}+w_k^{A\star}$ to either coordinate j or k while keeping all other coordinates unchanged.
- **Case 2.** $\eta_{j,A}\neq\eta_{k,A}$ and $g'(y^\star)=K_{jk}$, where

  $$K_{jk}:=\frac{\kappa_{k,A}-\kappa_{j,A}}{\eta_{j,A}-\eta_{k,A}}.$$

  In this case, the two LLMs are genuinely different, but the identity $g'(y^\star)=K_{jk}$ implies that they have the same marginal objective tradeoff between time efficiency and cost efficiency at the optimum. Consequently, any convex combination supported on $\{j,k\}$ is optimal. Similar to case 1, we can construct an alternative optimal solution by shifting the entire mass $w_j^{A\star}+w_k^{A\star}$ to either j or k while keeping all other coordinates unchanged. Moreover, we can show that as $\alpha\to 0$, this case can only occur when $\rho=1$ because there are only finitely many critical values $\{K_{jk}\}$ and when $\rho>1$, under Assumption 3, as $\alpha\to 0$, the derivative $g'(y^\star)\to\infty$ and exceeds every finite K_{jk}.

Applying the same argument symmetrically to $w^{B\star}$, we conclude that for all sufficiently small α there exists an optimal solution to (ALO) that is supported on at most one coordinate in w^A and at most one coordinate in w^B. Equivalently, the benchmark allocation can be chosen to use only two active LLMs at most, one for each hypothesis.

4.2. Matching the lower bound

Motivated by the optimal solution of the convex program, we consider the policy $\pi^{(2)}_{j_A^\star, j_B^\star}$ where

$$(j_A^\star, j_B^\star) \in \arg\min_{(i,j)\in[M]\times[M]} G_\alpha(e_i, e_j)$$

is a pair minimizing the convex program. According to the definition in (24), we can further write

$$\Phi_\alpha = G_\alpha(e_{j_A^\star}, e_{j_B^\star}) = \xi_A S_A(\alpha)\kappa_{j_A^\star,A} + \xi_B S_B(\alpha)\kappa_{j_B^\star,B} + \xi_A g\Big(S_A(\alpha)\eta_{j_A^\star,A}\Big) + \xi_B g\Big(S_B(\alpha)\eta_{j_B^\star,B}\Big),$$

Ideally, we can show that under the two-LLM policy $\tilde{\pi} = \pi_{j_A^\star, j_b^\star}$, we have

$$\mathbb{E}[C_{\tilde{\pi}}] + \mathbb{E}[g(W_{\tilde{\pi}})] = \Phi_\alpha(1 + o(1)) \text{ as } \alpha \to 0$$

which implies that $\tilde{\pi}$ is asymptotically optimal, and this is exactly the first part of Theorem 1. To establish the results, we need to match both the query cost component and the waiting time penalty component.

4.2.1. Characterizing the expected query cost and expected waiting time We state a proposition.

PROPOSITION 2 **(Expected cost and waiting time)**. *Under the sign-based policy $\pi^{(2)}_{j_A^\star, j_B^\star}$, the expected query cost and expected waiting time satisfy, for all sufficiently small α, we have*

$$\mathbb{E}[C_\pi] = \xi_A S_A(\alpha)\,\kappa_{j_A^\star,A} + \xi_B S_B(\alpha)\,\kappa_{j_B^\star,B} + O(1), \quad \mathbb{E}[W_\pi] = \xi_A S_A(\alpha)\,\eta_{j_A^\star,A} + \xi_B S_B(\alpha)\,\eta_{j_B^\star,B} + O(1).$$

The proof for Proposition 2 relies on bounding the expected *wrong-side* query counts (i.e., $n_{j_A^\star,B}, n_{j_B^\star,A}$). Note that according to Lemma 3 and the definition of $\bar{N}_j$ in (15), we have

$$\begin{aligned}
\mathbb{E}[C_\pi] &= c_{j_A^\star} \cdot (n_{j_A^\star,A} + n_{j_B^\star,A}) + c_{j_B^\star} \cdot (n_{j_B^\star,A}, n_{j_B^\star,B}) \\
&= c_{j_A^\star} \cdot \xi_A \cdot n_{j_A^\star,A} + c_{j_B^\star} \cdot \xi_B \cdot n_{j_B^\star,B} + \Big(c_{j_A^\star} \cdot \xi_B \cdot n_{j_A^\star,B} + c_{j_B^\star} \cdot \xi_A \cdot n_{j_B^\star,A}\Big) \\
&= \kappa_{j_A^\star,A} \cdot \xi_A \cdot S_A(\alpha) \cdot \frac{n_{j_A^\star,A} \cdot I_{j_A^\star,A}}{S_A(\alpha)} + \kappa_{j_B^\star,B} \cdot \xi_B \cdot S_B(\alpha) \cdot \frac{n_{j_B^\star,B} \cdot I_{j_B^\star,B}}{S_B(\alpha)} \\
&\quad + \Big(c_{j_A^\star} \cdot \xi_B \cdot n_{j_A^\star,B} + c_{j_B^\star} \cdot \xi_A \cdot n_{j_B^\star,A}\Big)
\end{aligned} \tag{26}$$

Similarly, we have

$$\mathbb{E}[W_\pi] = \eta_{j_A^\star,A}\cdot\xi_A\cdot S_A(\alpha)\cdot\frac{n_{j_A^\star,A}\cdot I_{j_A^\star,A}}{S_A(\alpha)} + \eta_{j_B^\star,B}\cdot\xi_B\cdot S_B(\alpha)\cdot\frac{n_{j_B^\star,B}\cdot I_{j_B^\star,B}}{S_B(\alpha)} + \left(c_{j_A^\star}\cdot\xi_B\cdot n_{j_A^\star,B} + c_{j_B^\star}\cdot\xi_A\cdot n_{j_B^\star,A}\right). \tag{27}$$

By establishing (26) and (27), we can then delegate the proof of Proposition 2 to the analysis on the relations between $S_\theta(\alpha)$ and $n_{j_A^\star,\theta}, n_{j_B^\star,\theta}$ for $\theta = A, B$. We state a lemma.

LEMMA 5. *Under the policy $\pi := \pi^{(2)}_{j_A^\star,j_B^\star}$, for all sufficiently small α, we have*

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

Moreover, we have $n_{j_B^\star,A} = O(1),\ n_{j_A^\star,B} = O(1)$.

According to the Lemma 2, Lemma 5 is sufficient to prove Proposition 2. However, establishing the result for Lemma 5 is complicated since both $n_{j_A^\star,B}$ and $n_{j_B^\star,A}$ are governed by the LLR evolution and the stopping rule defined in Lemma 1. The key idea is to construct a simple random walk based on the LLR evolution that starts from the same initial level and upper bounds the upward drifts of LLR. This construction delegates the task of bounding $n_{j_A^\star,B}$ and $n_{j_B^\star,A}$ to bounding the downward drifts of the constructed random walk, which is achieved by leveraging Hoeffding's inequality.

4.2.2. Asymptotics of the waiting cost $\mathbb{E}[g(W_\pi)]$. We now address the challenge of matching the expected waiting time penalty incurred by $\pi = \pi^{(2)}_{j_A^\star,j_B^\star}$ to its deterministic benchmark $\xi_A g\left(\sum_{j=1}^M \mu_j n_{j,A}\right) + \xi_B g\left(\sum_{j=1}^M \mu_j n_{j,B}\right)$ underpinning the lower bound Φ_α. Let $W_\pi^{(\theta)}$ denote the total waiting time, conditioning on the hypothesis θ. To establish the results, the key step is to anchor

$$W_\pi = \sum_{t=1}^{\tau} T_{j_t,t} = \sum_{n=1}^{N^A} T_{j_A^\star,n} + \sum_{n'=1}^{N^B} T_{j_B^\star,n'} \text{ under a hypothesis } \theta,$$

to its conditional mean

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

According to Lemma 5, under each hypothesis $\theta = A, B$, when $\alpha \to 0$, the query will gradually concentrate on $j_\theta^\star$ 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 $N^\theta \to \mathbb{E}_\theta[\tau]$ for each θ and further control the deviation of the cumulative waiting time to its mean conditioned on the realized τ. We state a lemma.

LEMMA 6. *For each $\theta \in \{A, B\}$, we have*

1. *For $S_\pi := \sum_{t=1}^{\tau} \mu_{j_t}$, we have $\mathbb{E}_\theta\left[(W_\pi - S_\pi)^2\right] \le \psi^2\, \mathbb{E}_\theta[\tau]$.*

2. *We have* $\mathbb{P}_\theta(|\tau - \mathbb{E}_\theta[\tau]| > k \cdot \sqrt{\mathbb{E}_\theta[\tau] \cdot \log \mathbb{E}_\theta[\tau]}) = O\left(\frac{1}{(m_\alpha^A)^2}\right)$ *for some constant* $k > 0$.

The first part of the lemma is driven by the sub-Gaussian assumption imposed on waiting time. The second part of the lemma is driven by the drift decomposition in (12) together with an application of Freedman's inequality, which provides sharp concentration for the associated martingale terms and thereby controls both the LLR drift and W_π. Specifically, for each $\theta \in \{A, B\}$, $t \geq 1$ and $x > 0$,

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

where C_ℓ and v_ℓ^2 are defined in (5) and (6). Using this sharp concentration bound, we can further show that for any $p > 1$,

$$\mathbb{E}[\tau^p] \;\leq\; C_p\,(\mathbb{E}[\tau])^p \qquad \text{uniformly over } \alpha \in (0, \alpha_0],$$

for some constant $C_p < \infty$ independent of α. Moreover, since $\mathbb{E}[W_\pi] \geq \underline{\mu}\,\mathbb{E}[\tau]$, Minkowski's inequality yields

$$\sup_{\alpha\in(0,\alpha_0]} \mathbb{E}\left[\left(\frac{W_\pi}{\mathbb{E}[W_\pi]}\right)^p\right] \;\leq\; \frac{m_p}{\underline{\mu}^{\,p}} \sup_{\alpha\in(0,\alpha_0]} \frac{\mathbb{E}[\tau^p]}{(\mathbb{E}[\tau])^p} \;<\; \infty,$$

which implies uniform integrability and establishes the second part of the proposition.

Finally, 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 $M_t^{(\theta)}$. This eventually enables us to match the lower bound. We formalize the argument in the following proposition.

PROPOSITION 3. *Under* $\tilde{\pi} := \pi^{(2)}_{j_A^\star, j_B^\star}$, *we have* $\frac{\mathbb{E}_\theta[g(W_{\tilde{\pi}})]}{g(m_{\tilde{\pi}}^\theta)} = O(1/\log(1/\alpha))$ *for* $\theta = A, B$.

5. 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, latency, 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. This suggests that effective AI decision systems can remain structurally simple even as the number of available models grows.

References

Abdallah, Aisha, Mohd Aizaini Maarof, Anazida Zainal. 2016. Fraud detection system: A survey. *Journal of Network and Computer Applications* **68** 90–113.

Aggarwal, Pranjal, Aman Madaan, Yiming Yang, Mausam. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. *arXiv preprint arXiv:2305.11860* .

Ao, Ruicheng, Gan Luo, David Simchi-Levi, Xinshang Wang. 2025. Optimizing LLM inference: Fluid-guided online scheduling with memory constraints. *arXiv preprint arXiv:2504.11320* .

Bari, Agrim, Parikshit Hegde, Gustavo de Veciana. 2025. Optimal scheduling algorithms for llm inference: Theory and practice. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* **9**(3). doi:10.1145/3771574.

Chen, Lingjiao, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, James Zou. 2024a. Are more LLM calls all you need? towards the scaling properties of compound AI systems. *Advances in Neural Information Processing Systems*, vol. 37.

Chen, Lingjiao, Matei Zaharia, James Zou. 2024b. Frugalgpt: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research* .

Chen, Zixi, Yinyu Ye, Zijie Zhou. 2025. Adaptively robust llm inference optimization under prediction uncertainty. *arXiv preprint arXiv:2508.14544* doi:10.48550/arXiv.2508.14544.

Chernoff, Herman. 1959. Sequential design of experiments. *The Annals of Mathematical Statistics* **30**(3) 755–770.

Dayanik, Savas, Semih O Sezer. 2012. Multisource bayesian sequential binary hypothesis testing problem. *Annals of Operations Research* **201**(1) 99–130.

Ding, Dujian, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, Ahmed Hassan Awadallah. 2024. Hybrid llm: Cost-efficient and quality-aware query routing. *International Conference on Learning Representations*.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325* .

Durrett, Rick. 2019. *Probability: Theory and Examples*, vol. 49. Cambridge University Press.

Gorwa, Robert, Reuben Binns, Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* **7**(1).

Gou, Zhibin, Zhihong Shao, Yadong Gong, Yelong Shen, Yujiu Yang, Nan Duan, Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738* .

Huang, Baihe, Shanda Li, Tianhao Wu, Yiming Yang, Ameet Talwalkar, Kannan Ramchandran, Michael I. Jordan, Jiantao Jiao. 2025. Sample complexity and representation ability of test-time scaling paradigms. *arXiv preprint arXiv:2506.05295* .

Huang, Jingkai, Will Ma, Zhengyuan Zhou. 2026. Optimal Bayesian stopping for efficient inference of consistent LLM answers. *arXiv preprint arXiv:2602.05395* .

Jaillet, Patrick, Jiashuo Jiang, Konstantina Mellou, Marco Molinaro, Chara Podimata, Zijie Zhou. 2025. Online scheduling for LLM inference with KV cache constraints. *arXiv preprint arXiv:2502.07115* .

Jung, Jaehyeok, Dongwoo Lee, Jinwoo Kim. 2025. Trust or escalate: LLM judges with provable guarantees. *International Conference on Learning Representations*.

Khraisat, Ansam, Iqbal Gondal, Peter Vamplew, Joarder Kamruzzaman. 2019. Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity* **2**(1) 1–22.

Kim, Alex, Keonwoo Kim, Sangwon Yoon. 2024. Debate: Devil's advocate-based assessment and text evaluation. *Findings of the Association for Computational Linguistics: ACL 2024*. 1885–1897.

Lambez, Tidhar, Kobi Cohen. 2022. Anomaly search with multiple plays under delay and switching costs. *IEEE Transactions on Signal Processing* **70** 174–189.

Lattimore, Tor, Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press.

Li, Yiwei, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, Kan Li. 2024. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *arXiv preprint arXiv:2401.10480* .

Li, Yueying, Jim Dai, Tianyi Peng. 2025. Throughput-optimal scheduling algorithms for llm inference and ai agents. *arXiv preprint arXiv:2504.07347* doi:10.48550/arXiv.2504.07347.

Naghshvar, Mohammad, Tara Javidi. 2013. Active sequential hypothesis testing. *The Annals of Statistics* **41**(6) 2703–2738.

Ong, Isaac, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, Ion Stoica. 2025. Routellm: Learning to route llms from preference data. *International Conference on Learning Representations*.

OpenAI. 2025. Introducing GPT-5. URL `https://openai.com/index/introducing-gpt-5/`.

Setlur, Amrith, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146* .

Singhi, Nishad, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, Anna Rohrbach. 2025. When to solve, when to verify: Compute-optimal problem solving and generative verification for llm reasoning. *arXiv preprint arXiv:2504.01005* .

Snell, Charlie, Jaehoon Lee, Kelvin Xu, Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314* .

Song, Yun, Georgios Fellouris. 2017. Asymptotically optimal sequential design for rank aggregation. *Operations Research* **65**(1) 246–263.

Tsopelakos, Aristomenis, Georgios Fellouris. 2022. Sequential anomaly detection under sampling constraints. *IEEE Transactions on Information Theory* **69**(12) 8126–8146.

Vaidhiyan, Nidhin Koshy, Rajesh Sundaresan. 2015. Active search with a cost for switching actions. *arXiv preprint arXiv:1505.02358* .

Verga, Pat, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796* .

Viola, Paul A., Michael J. Jones. 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*, vol. 1. IEEE, I–511–I–518. doi:10.1109/CVPR.2001.990517.

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

Wang, Junlin, Jue Wang, Ben Athiwaratkun, Ce Zhang, James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692* .

Wang, Meixuan, Yinyu Ye, Zijie Zhou. 2025. Llm serving optimization with variable prefill and decode lengths. *arXiv preprint arXiv:2508.06133* doi:10.48550/arXiv.2508.06133.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* .

Wu, Yangzhen, Zhiqing Sun, Shanda Li, Sean Schieber, Jiantao Jiao, Lingpeng Xie, Dale Schuurmans, Bo Dai. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724* .

Xu, Zhixiang, Matt Kusner, Kilian Weinberger, Minmin Chen. 2013. Cost-sensitive tree of classifiers. Sanjoy Dasgupta, David McAllester, eds., *Proceedings of the 30th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 28. PMLR, 133–141.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601* .

Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* .

Zhang, Lunjun, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240* .

Appendix

The appendix is organized as follows. In Appendix B, we provide proofs for all theoretical results except for Theorems 1 and 2. In Appendix A, we prove some useful results. In Appendix C, we provide the proof for the upper bound result (Theorem 1). In Appendix D, we provide structural properties of the optimal solution to the lower bound problem.

Appendix A: Additional Preliminary Results

In this section, we prove some preliminary technical results that will be repeatedly used in the proofs of main results.

A.1. Likelihood Ratio Martingale and Change of Measure

We define the likelihood ratio process as

$$\Lambda_t = e^{L_t}, \qquad t \geq 0.$$

In this subsection, we provide some preliminary properties of Λ_t.

LEMMA A.1 **(Likelihood ratio martingale)**. *Under $\mathbb{P}_B$, $(\Lambda_t)_{t\geq 0}$ is a nonnegative $(\mathcal{G}_t)$-martingale with*

$$\mathbb{E}_B[\Lambda_t \mid \mathcal{G}_{t-1}] = \Lambda_{t-1}, \qquad \mathbb{E}_B[\Lambda_t] = 1, \qquad t \geq 0.$$

Similarly, under $\mathbb{P}_A$, the process $(\Lambda_t^{-1})_{t\geq 0}$ is a nonnegative $(\mathcal{G}_t)$-martingale with $\mathbb{E}_A[\Lambda_t^{-1}] = 1$.

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

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

because $e^{\ell_{j_t}(y)} = \mathbb{P}(Y = y \mid \theta = A, j_t)/\mathbb{P}(Y = y \mid \theta = B, j_t)$. Thus $(\Lambda_t)_{t\geq 0}$ is a martingale under $\mathbb{P}_B$ with $\mathbb{E}_B[\Lambda_0] = 1$. Taking expectations gives $\mathbb{E}_B[\Lambda_t] = 1$ for all t. The argument under $\mathbb{P}_A$ is analogous: for $Z_t := \Lambda_t^{-1} = e^{-L_t}$ we have

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

LEMMA A.2 **(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,*

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

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

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

Proof. We prove the first identity; the second is analogous. Since τ is almost surely finite and (Λ_t) has bounded increments, the optional stopping theorem gives

$$\mathbb{E}_B[\Lambda_\tau] = \mathbb{E}_B[\Lambda_0] = 1.$$

Given any bounded $\mathcal{G}_\tau$-measurable random variable X, we can derive the following equality:

$$\begin{aligned}\mathbb{E}_B[\Lambda_t X \mid \mathcal{G}_{t-1}] &= \Lambda_{t-1} \sum_{y\in\{A,B\}} P_B(Y_t = y \mid \mathcal{G}_{t-1}) e^{\ell_{j_t}(y)} \mathbb{E}[X \mid \mathcal{G}_{t-1}, Y_t = y] \\ &= \Lambda_{t-1} \sum_{y\in\{A,B\}} P_A(Y_t = y \mid \mathcal{G}_{t-1}) \mathbb{E}[X \mid \mathcal{G}_{t-1}, Y_t = y] \\ &= \Lambda_{t-1} \mathbb{E}_A[X \mid \mathcal{G}_{t-1}],\end{aligned}$$

where the first equality is because X is $\mathcal{G}_t$-measurable and j_t is determined by $\mathcal{G}_{t-1}$.

Then, we prove the statement $\mathbb{E}_B[\Lambda_t X] = \mathbb{E}_A[X]$ by induction. First, when $t = 0$, we have $\mathbb{E}_B[\Lambda_0 X] = \mathbb{E}_A[X]$ because $\Lambda_0 = 1$ and $\mathcal{G}_0$ is empty. Then, given the statement holds for $t = k-1$, we prove the statement for $t = k$:

$$\mathbb{E}_B[\Lambda_k X] = \mathbb{E}_B[\mathbb{E}_B[\Lambda_k X \mid \mathcal{G}_{k-1}]] = \mathbb{E}_B[\Lambda_{k-1}\mathbb{E}_A[X \mid \mathcal{G}_{k-1}]].$$

Since the statement holds for $t = k-1$ and $\mathbb{E}_A[X \mid \mathcal{G}_{k-1}]$ is $\mathcal{G}_{t-1}$-measurable, we have

$$\mathbb{E}_B[\Lambda_{k-1}\mathbb{E}_A[X \mid \mathcal{G}_{k-1}]] = \mathbb{E}_A[\mathbb{E}_A[X \mid \mathcal{G}_{k-1}]] = \mathbb{E}_A[X],$$

which implies that $\mathbb{E}_B[\Lambda_k X] = \mathbb{E}_A[X]$ and the statement holds for $t = k$. Thus, we have $\mathbb{E}_B[\Lambda_t X] = \mathbb{E}_A[X]$ for any t. Similarly, the remaining results can also be proved. □

A.2. Error Probability Bounds

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

LEMMA A.3 **(Exponential error bounds)**. *For the posterior-threshold stopping rule with threshold $\alpha \in (0, 1/2)$, there exists a constant satisfying*

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

Proof. We prove the first inequality; the second is analogous. Under $\theta = A$, the error event $\{D = B\}$ occurs if and only if $L_\tau \le -B_\alpha$. By the change-of-measure identity (Lemma A.2),

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

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

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

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 $\mathbb{P}_B(D = A) \le e^{-A_\alpha} \le c_{\text{err}} \cdot \alpha$. □

A.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. We let $N_j(t)$ denote the total number of queries sent to LLM j until period t:

$$N_j(t) := \sum_{\ell=1}^{t} \mathbf{1}_{\{j_t=j\}}.$$

LEMMA A.4 **(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. We prove the claim under $\mathbb{P}_A$; the argument under $\mathbb{P}_B$ is identical with signs adjusted. Define

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

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

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

As noted in the remark above, under $\mathbb{P}_A$ we have

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

so

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

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:

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

□

A.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 A.5 **(Freedman's inequality; Theorem 5.2.9 in Durrett 2019)**. *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 $V_t := \sum_{s=1}^{t} \mathbb{E}[X_s^2 \mid \mathcal{F}_{s-1}]$ 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 $V_t \leq v$ almost surely, we have*

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

Then, we prove the concentration inequalities.

LEMMA A.6 **(Freedman-type concentration for the LLR martingales)**. *Under Assumptions 1-3, for each $\theta \in \{A, B\}$, each integer $t \geq 1$, and each $x > 0$,*

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

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

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

and under $\mathbb{P}_B$,

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

Proof. Fix $\theta \in \{A, B\}$ and $t \geq 1$. From Lemma A.4, 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 A.5 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 A.4. From Lemma A.4, we have $|D_s^{(\theta)}| \leq 2C_\ell$ almost surely for all $s \geq 1$. Thus the increment bound in Lemma A.5 holds with

$$b := 2C_\ell.$$

The predictable quadratic variation is, by (6),

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

By Lemma A.5 with the choices $b = 2C_\ell$ and $v = v_\ell^2 t$, for each $x \geq 0$,

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

Thus, we have

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

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

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

Therefore,

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

A.5. Overshoot and Expected Log-Likelihood Ratio 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 log-likelihood ratio L_τ typically overshoots the boundary by a small amount. The following lemma provides a uniform bound on this overshoot.

LEMMA A.7 **(Uniform overshoot bound)**. *For the posterior-threshold stopping rule, the overshoot is uniformly bounded:*

$$|L_\tau - A_\alpha| \le C_\ell \quad \text{on } \{L_\tau \ge A_\alpha\},$$

and

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

where C_ℓ is the bound from (5).

Proof. 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 $\{L_\tau \ge A_\alpha\}$, 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 (5). Since $L_\tau \ge A_\alpha$ by assumption, we have

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

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

The next lemma gives upper and lower bounds on the expected cumulative log-likelihood ratio at stopping under each hypothesis.

LEMMA A.8 **(Expected log-likelihood ratio at stopping)**. *Suppose the policy π satisfies $\mathbb{E}_A[\tau] < \infty$ and $\mathbb{E}_B[\tau] < \infty$. Then for all sufficiently small $\alpha > 0$,*

$$A_\alpha - c_{\mathrm{err}}\alpha \cdot (A_\alpha + B_\alpha + C_\ell) \le \mathbb{E}_A[L_\tau] \le A_\alpha + C_\ell, \tag{A.1}$$

$$B_\alpha - c_{\mathrm{err}}\alpha \cdot (A_\alpha + B_\alpha + C_\ell) \le \mathbb{E}_B[-L_\tau] \le B_\alpha + C_\ell. \tag{A.2}$$

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

Proof. We prove (A.1); the proof of (A.2) is analogous. When $\theta = A$, we can decompose the expectation by the decision outcome:

$$\begin{aligned}\mathbb{E}_A[L_\tau] &= \mathbb{E}_A[L_\tau \mathbb{1}_{\{D=A\}}] + \mathbb{E}_A[L_\tau \mathbb{1}_{\{D=B\}}] \\ &= \mathbb{E}_A[L_\tau \mathbb{1}_{\{L_\tau \ge A_\alpha\}}] + \mathbb{E}_A[L_\tau \mathbb{1}_{\{L_\tau \le -B_\alpha\}}].\end{aligned}$$

On the event $\{L_\tau \ge A_\alpha\}$, Lemma A.7 gives $A_\alpha \le L_\tau \le A_\alpha + C_\ell$. On the event $\{L_\tau \le -B_\alpha\}$, we have $-B_\alpha - C_\ell \le L_\tau \le -B_\alpha < 0$. Therefore,

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

By Lemma A.3, we have $\mathbb{P}_A(D=B) \le c_{\mathrm{err}}\alpha$ and hence

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

which establishes the lower bound in (A.1). For the upper bound, note that on $\{L_\tau \ge A_\alpha\}$ we have $L_\tau \le A_\alpha + C_\ell$ and on $\{L_\tau \le -B_\alpha\}$ we have $L_\tau \le 0$. Therefore,

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

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

The proof of (A.2) is identical with $-L_\tau$ and B_α in place of L_τ and A_α.

□

A.6. Concentration of Total Waiting Time

In this subsection, we present concentration results for the total waiting time W_π. Recall that given the stopping time under a policy π, the total waiting time is

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

where $j_t \in \{1,\dots,M\}$ is the LLM selected at time t, and $T_{j_t,t}$ is the corresponding random waiting time. Let

$$S_\pi := \sum_{t=1}^{\tau} \mu_{j_t} = \sum_{j=1}^{M} \mu_j N_j(\tau)$$

denote the random total mean waiting time under policy π. We also define a coarser filtration

$$\mathcal{H}_\tau := \sigma\big(\{j_s, Y_s\}_{s=1}^{\tau}\big).$$

In the following lemma, we prove that $W_\pi - S_\pi$ is a sub-Gaussian variable.

LEMMA A.9 **(Conditional sub-Gaussianity of W_π).** *Under Assumption 2, for any $\theta \in \{A,B\}$ and any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}_\theta\big[e^{\lambda(W_\pi - S_\pi)} \,\big|\, \mathcal{H}_\tau, \theta\big] \;\le\; \exp\Big(\frac{\psi^2\lambda^2}{2}\,\tau\Big) \quad \textit{almost surely.}$$

In particular, conditional on $(\theta, \mathcal{H}_\tau)$, the centered variable $W_\pi - S_\pi$ is sub-Gaussian with variance proxy at most $\psi^2\tau$. Furthermore, for any $\theta \in \{A,B\}$ and any $x > 0$,

$$\mathbb{P}_\theta\big(|W_\pi - S_\pi| \ge x \,\big|\, \mathcal{H}_\tau, \theta\big) \;\le\; 2\exp\Big(-\frac{x^2}{2\psi^2\tau}\Big) \quad \textit{almost surely,}$$

and hence

$$\mathbb{P}_\theta\big(|W_\pi - S_\pi| \ge x\big) \;\le\; 2\,\mathbb{E}_\theta\Big[\exp\Big(-\frac{x^2}{2\psi^2\tau}\Big)\Big].$$

Proof. Fix $\theta \in \{A,B\}$. We will establish the moment generating function (MGF) bound conditional on $(\theta, \mathcal{H}_\tau)$. By definition, we have $W_\pi - S_\pi = \sum_{t=1}^{\tau}(T_{j_t,t} - \mu_{j_t})$. By Assumption 1, the array $(T_{j,t})_{j,t}$ is independent of $(\theta, (Y_t)_{t\ge1})$ and of the policy's internal randomization. Consequently, conditional on $(\theta, \mathcal{H}_\tau)$:

- The LLM indices $(j_t)_{t\le\tau}$ 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 $(T_{j_t,t})_{t\le\tau}$ are independent of $(\theta, \mathcal{H}_\tau)$ and mutually independent across t (since $(T_{j,t})_{t\ge1}$ is i.i.d. in t for each fixed j by Assumption 1).

Then, we can compute the conditional MGF as follows:

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

By Chernoff bound, we have

$$\mathbb{P}_\theta\big(W_\pi - S_\pi \ge x \,\big|\, \mathcal{H}_\tau, \theta\big) \le e^{-\lambda x}\mathbb{E}_\theta\big[e^{\lambda(W_\pi - S_\pi)} \,\big|\, \mathcal{H}_\tau, \theta\big] \le \exp\Big(-\lambda x + \frac{\psi^2\lambda^2\tau}{2}\Big).$$

Then, we set $\lambda = \frac{x}{\psi^2\tau}$, and obtain that $\mathbb{P}_\theta\big(W_\pi - S_\pi \ge x \,\big|\, \mathcal{H}_\tau, \theta\big) \le \exp\Big(-\frac{x^2}{2\psi^2\tau}\Big)$. Similarly, we have $\mathbb{P}_\theta\big(W_\pi - S_\pi \le -x \,\big|\, \mathcal{H}_\tau, \theta\big) \le \exp\Big(-\frac{x^2}{2\psi^2\tau}\Big)$. Thus, we have

$$\mathbb{P}_\theta\big(|W_\pi - S_\pi| \ge x \,\big|\, \mathcal{H}_\tau, \theta\big) \le 2\exp\Big(-\frac{x^2}{2\psi^2\tau}\Big) \quad \text{almost surely,}$$

and

$$\mathbb{P}_\theta\big(|W_\pi - S_\pi| \ge x\big) \le \mathbb{E}_\theta\Big[2\exp\Big(-\frac{x^2}{2\psi^2\tau}\Big)\Big] = 2\,\mathbb{E}_\theta\Big[\exp\Big(-\frac{x^2}{2\psi^2\tau}\Big)\Big]. \qquad \square$$

Then, in the following lemma, we show that the second moment of $W_\pi - S_\pi$ is bounded.

LEMMA A.10 **(Second-moment bound for the waiting-time fluctuation)**. *Under Assumption 2, for each $\theta \in \{A, B\}$,*

$$\mathbb{E}_\theta\big[(W_\pi - S_\pi)^2\big] \le \psi^2\,\mathbb{E}_\theta[\tau].$$

Consequently, under the Bayes measure,

$$\mathbb{E}\big[(W_\pi - S_\pi)^2\big] \le \psi^2\,\mathbb{E}[\tau].$$

Proof. We prove the bound under $\mathbb{P}_\theta$ for a fixed $\theta \in \{A, B\}$; the Bayes bound then follows by averaging over θ. For each $t \ge 1$, define

$$D_t := T_{j_t,t} - \mu_{j_t},$$

where j_t is the LLM selected at time t (which is $\mathcal{H}_t$-measurable and hence $\mathcal{H}_\tau$-measurable for $t \le \tau$). We have $\mathbb{E}_\theta[D_t \mid \mathcal{H}_\tau, \theta] = 0$ and D_t's are conditionally independent given $(\theta, \mathcal{H}_\tau)$. Since $W_\pi - S_\pi = \sum_{t=1}^{\tau} D_t$, for each fixed $t \le \tau$, we have

$$\mathrm{Var}_\theta(D_t \mid \mathcal{H}_\tau, \theta) = \mathbb{E}\big[(T_{j_t,t} - \mu_{j_t})^2 \mid \mathcal{H}_\tau, \theta\big] = \mathbb{E}\big[(T_{j_t,t} - \mu_{j_t})^2 \mid j_t\big] = \mathrm{Var}(T_{j_t,t}) \le \psi^2,$$

where the inequality is due to Lemma 5.4 in Lattimore and Szepesvári (2020). Since $(D_t)_{t\le\tau}$ are conditionally independent given $(\theta, \mathcal{H}_\tau)$ and each has conditional mean zero, we have

$$\mathbb{E}_\theta\big[(W_\pi - S_\pi)^2 \mid \mathcal{H}_\tau, \theta\big] = \mathrm{Var}_\theta\Big(\sum_{t=1}^{\tau} D_t \,\Big|\, \mathcal{H}_\tau, \theta\Big) = \sum_{t=1}^{\tau} \mathrm{Var}_\theta(D_t \mid \mathcal{H}_\tau, \theta) \le \psi^2\,\tau.$$

Then, we have

$$\mathbb{E}_\theta\big[(W_\pi - S_\pi)^2\big] \le \psi^2\, \mathbb{E}_\theta[\tau],$$

and

$$\mathbb{E}\big[(W_\pi - S_\pi)^2\big] \le \psi^2\, \mathbb{E}[\tau]. \qquad \square$$

A.7. Expected Number of Wong Queries

In this subsection, we provide upper bounds for the expected number of "wrong-side" queries under the policy $\pi^{(2)}_{j_A^\star, j_B^\star}$.

LEMMA A.11 **(Finite expected number of "wrong-side" queries)**. *Consider the sign-based policy $\pi^{(2)}_{j_A^\star, j_B^\star}$ for any fixed α. Then there exists a constant $C^A_{\text{neg}} > 0$, independent of α, such that*

$$n_{j_B^\star, A} = \mathbb{E}_A[N_{j_B^\star}] \le C^A_{\text{neg}} \qquad \textit{for all sufficiently small } \alpha. \tag{A.3}$$

Similarly, there exists a constant $C^B_{\text{pos}} > 0$, independent of α, such that

$$n_{j_A^\star, B} = \mathbb{E}_B[N_{j_A^\star}] \le C^B_{\text{pos}} \qquad \textit{for all sufficiently small } \alpha. \tag{A.4}$$

Proof. We prove (A.3); the argument for (A.4) is symmetric.

Under $\mathbb{P}_A$, the conditional expectation of the LLR increment at time t given j_t is $I_{j_t, A} > 0$. In particular, whether we query $j_A^\star$ or $j_B^\star$, the LLR has strictly positive drift under A. Therefore, $(L_t)_{t\ge 0}$ is a random walk with uniformly bounded increments and strictly positive drift under $\mathbb{P}_A$.

Each time $j_B^\star$ is queried under $\theta = A$, the LLR at the beginning of that round satisfies $L_{t-1} + \delta < 0$ by the definition of the policy $\pi^{(2)}_{j_A^\star, j_B^\star}$. Thus $N_{j_B^\star}$ is exactly the number of time indices $t \le \tau$ for which the process is below $-\delta$ just before querying:

$$N_{j_B^\star} = \sum_{t=1}^{\tau} \mathbb{1}_{\{j_t = j_B^\star\}} = \sum_{t=1}^{\tau} \mathbb{1}_{\{L_{t-1} < -\delta\}}. \tag{A.5}$$

Therefore, $N_{j_B^\star}$ is bounded above by the total number of visits of the process (L_t) to the value $-\delta$ before it crosses the upper boundary A_α.

In our setting, the increments of L_t are not i.i.d. because the policy may switch between $j_A^\star$ and $j_B^\star$. However, each increment belongs to the finite set

$$\{\ell_{j_A^\star}(A), \ell_{j_A^\star}(B), \ell_{j_B^\star}(A), \ell_{j_B^\star}(B)\},$$

and under $\mathbb{P}_A$ all these increments have strictly positive expectation (since $I_{j,A} = \mathbb{E}_A[\ell_j(Y)] > 0$ for both $j \in \{j_A^\star, j_B^\star\}$). Therefore, we can bound the behavior of (L_t) using a stochastic dominance argument.

In the following, we construct a auxiliary random walk to prove the result. Let

$$m_A := \min\{I_{j_A^\star, A}, I_{j_B^\star, A}\} > 0$$

be the minimal drift under $\theta = A$ across the two LLMs. Define an auxiliary random walk $(\tilde{L}_t)_{t\ge 0}$ with $\tilde{L}_0 = 0$ and i.i.d. increments $\tilde{X}_t$ such that:

- $\mathbb{E}[\tilde{X}_t] = m_A > 0$,
- $|\tilde{X}_t| \le C_\ell$ almost surely.

To verify that such a random walk can be constructed: we can take $\tilde{X}_t$ to be a random variable on $[-C_\ell, C_\ell]$ with mean m_A. For example, $\tilde{X}_t$ can be a shifted and scaled Rademacher random variable. The key property we need is $\tilde{X}_t$ has the smallest drift among all possible LLR increments under $\mathbb{P}_A$, while maintaining bounded support.

We claim that whenever $L_t < -\delta$, the process $(L_s - L_t)_{s \ge t}$ stochastically dominates $(\tilde{L}_s - \tilde{L}_t)_{s \ge t}$ in the sense that starting from the same initial position, L_t drifts upward at least as fast as $\tilde{L}_t$.

To see this formally, note that for each $s \ge t$:

$$\mathbb{E}_A[L_s - L_t \mid \mathcal{G}_t] = \sum_{u=t+1}^{s} I_{j_u,A} \ge \sum_{u=t+1}^{s} m_A = (s-t)m_A = \mathbb{E}[\tilde{L}_s - \tilde{L}_t].$$

This shows that conditional on any realization up to time t, the expected increment of (L_s) is at least as large as that of $(\tilde{L}_s)$. By a standard coupling argument, we can construct both processes on the same probability space such that

$$L_s \ge L_t + (\tilde{L}_s - \tilde{L}_t) \quad \text{almost surely for all } s \ge t,$$

whenever $L_t = \tilde{L}_t$. Therefore, if $L_t < -\delta$, then starting from this point, L_s reaches $-\delta$ before $\tilde{L}_s$ does.

Then, we are ready to bound the total number of wrong-side queries. Let N_{neg} denote the total number of visits to the value $-\delta$ for $\tilde{L}_t$:

$$N_{\text{neg}} := \sum_{t=1}^{\infty} \mathbb{1}_{\{\tilde{L}_t < -\delta\}}.$$

By the strong law of large numbers, since $\tilde{X}_t$ are i.i.d. with $\mathbb{E}[\tilde{X}_t] = m_A > 0$, we have

$$\frac{\tilde{L}_t}{t} = \frac{1}{t}\sum_{s=1}^{t} \tilde{X}_s \xrightarrow[t\to\infty]{a.s.} m_A > 0.$$

Therefore, $\tilde{L}_t \to \infty$ almost surely as $t \to \infty$. This implies that $\tilde{L}_t$ can visit the the finite value $-\delta$ only finitely many times almost surely. In particular, we have

$$\mathbb{E}[N_{\text{neg}}] = \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}_{\{\tilde{L}_t < -\delta\}}\right] = \sum_{t=1}^{\infty} \mathbb{E}[\mathbb{1}_{\{\tilde{L}_t < -\delta\}}] = \sum_{t=1}^{\infty} \mathbb{P}(\tilde{L}_t < -\delta) < \infty.$$

The finiteness follows from standard random walk theory: for a random walk with positive drift and bounded increments, the probability $\mathbb{P}(\tilde{L}_t < 0)$ decays exponentially in t. Specifically, by Hoeffding's inequality, there exist constants $c, \gamma > 0$ such that

$$\mathbb{P}(\tilde{L}_t < -\delta) \le c\exp(-\gamma t) \quad \text{for all } t \ge 1.$$

Therefore,

$$\sum_{t=1}^{\infty} \mathbb{P}(\tilde{L}_t < -\delta) \le \sum_{t=1}^{\infty} c\exp(-\gamma t) = \frac{c\exp(-\gamma)}{1-\exp(-\gamma)} < \infty.$$

By the stochastic dominance established in Step 4, whenever $L_t < -\delta$, the process (L_s) reaches $-\delta$ before $(\tilde{L}_s)$ does (when both start from the same negative value). Therefore, L_t spends less time in the negative half-line than $\tilde{L}_t$:

$$N_B = \sum_{t=1}^{\tau} \mathbb{1}_{\{L_{t-1} < -\delta\}} \le \sum_{t=1}^{\infty} \mathbb{1}_{\{L_{t-1} < -\delta\}} \le \sum_{t=1}^{\infty} \mathbb{1}_{\{\tilde{L}_{t-1} < -\delta\}} \le N_{\text{neg}} + 1.$$

The last inequality accounts for the shift in indices. Taking expectations under $\mathbb{P}_A$:

$$n_{B,A} = \mathbb{E}_A[N_B] \le \mathbb{E}_A[N_{\text{neg}}] + 1 =: C^A_{\text{neg}} < \infty,$$

where C^A_{neg} does not depend on the thresholds (A_α, B_α) and hence is independent of α.

The proof of (A.4) under $\theta = B$ is identical, swapping the roles of (A, B) and $(j^\star_A, j^\star_B)$. Under $\mathbb{P}_B$, the LLR has negative drift (i.e., $-L_t$ has positive drift), so $L_t \to -\infty$ almost surely. Therefore, L_t can visit the positive half-line only finitely many times, and the expected number of queries to $j^\star_A$ under $\mathbb{P}_B$ is bounded by a constant C^B_{pos} independent of α. □

Appendix B: Proofs of Lemmas and Propositions

In this section, we provide proofs for theoretical results, excluding Theorems 1 and 2.

B.1. Proof of Lemma 1

Given the historical LLM choices $j_{1:t}$ and outputs $Y_{1:t}$, the posterior belief of $\theta = A$ can be computed as follows:

$$\begin{aligned}
\mathbb{P}(\theta = A \mid Y_{1:t}, j_{1:t}) &= \frac{\xi_A \cdot \mathbb{P}(Y_{1:t} \mid \theta = A, j_{1:t})}{\xi_A \cdot \mathbb{P}(Y_{1:t} \mid \theta = A, j_{1:t}) + \xi_B \cdot \mathbb{P}(Y_{1:t} \mid \theta = B, j_{1:t})} \\
&= \frac{\xi_A \cdot \Pi_{\ell=1}^t \mathbb{P}(Y_\ell \mid \theta = A, j_\ell)}{\xi_A \cdot \Pi_{\ell=1}^t \mathbb{P}(Y_\ell \mid \theta = A, j_\ell) + \xi_B \cdot \Pi_{\ell=1}^t \mathbb{P}(Y_\ell \mid \theta = B, j_\ell)} \\
&= \frac{e^{\delta + L_t}}{1 + e^{\delta + L_t}},
\end{aligned}$$

where the first equation is due to Bayes' theorem, the second equation is due to Assumption 1, and the last equation is due to (4). Then the posterior belief of $\theta = B$ is

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

The stopping criterion in (1) can rewritten as follow:

$$\begin{aligned}
\mathbb{P}(\theta = A \mid Y_{1:t}, j_{1:t}) \ge 1 - \alpha \quad &\Leftrightarrow \quad \frac{e^{\delta + L_t}}{1 + e^{\delta + L_t}} \ge 1 - \alpha \quad \Leftrightarrow \quad L_t \ge \log \frac{1-\alpha}{\alpha} - \delta = A_\alpha, \\
\mathbb{P}(\theta = B \mid Y_{1:t}, j_{1:t}) \ge 1 - \alpha \quad &\Leftrightarrow \quad \frac{1}{1 + e^{\delta + L_t}} \ge 1 - \alpha \quad \Leftrightarrow \quad -L_t \ge \log \frac{1-\alpha}{\alpha} + \delta = B_\alpha.
\end{aligned}$$ □

B.2. Proof of Lemma 2

We proceed in three steps.

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

By Lemma A.4, under $\mathbb{P}_A$ we can write

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

where $(M_t^{(A)})_{t \ge 0}$ is a martingale under $\mathbb{P}_A$ with respect to the natural filtration $(\mathcal{G}_t)_{t \ge 0}$, 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) $\tau < \infty$ almost surely under $\mathbb{P}_A$,

(ii) $\mathbb{E}_A[\tau] < \infty$,

(iii) The increments of $M_t^{(A)}$ are bounded: $|M_t^{(A)} - M_{t-1}^{(A)}| \le |L_t - L_{t-1}| + |I_{j_t,A}| \le 2C_\ell$ almost surely for all $t \ge 1$.

Given the above conditions, by the Optional Stopping Theorem (see Theorem 5.7.3 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 $N_j(\tau)$ is defined in (13).

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

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

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 A.4, 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 A.3, there exists a constant $c_{\text{err}} > 0$ (independent of α and of the policy) such that for all sufficiently small α,

$$\mathbb{P}_A(D = B) \le c_{\text{err}}\alpha, \qquad \mathbb{P}_B(D = A) \le c_{\text{err}}\alpha. \tag{B.1}$$

According to Lemma A.8, we obtain

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

where $K_\alpha := c_{\text{err}}\alpha(A_\alpha + B_\alpha + C_\ell)$. By Step 1, $\mathbb{E}_A[L_\tau] = \sum_j I_{j,A} n_{j,A}$, so we have established (17). A similar argument can induce (18).

Step 3: Upper bound of K_α/S_α. Since $x \log \frac{1}{x} \le 1/e$, we can also deduce that

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

Thus, we can set $C_K = \frac{2c_{\text{err}}}{e} + c_{\text{err}} C_\ell$. □

B.3. Proof of Lemma 3

By definition, the total cost is

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

Taking expectations under $\mathbb{P}_\theta$ and using $\mathbb{E}_\theta[\tau] < \infty$ gives (21) by dominated convergence (since costs are uniformly bounded). The proof of W_π is identical with μ_j in place of c_j. The Bayes measure identities (22) follow immediately from

$$\mathbb{E}[C_\pi] = \xi_A \mathbb{E}_A[C_\pi] + \xi_B \mathbb{E}_B[C_\pi] = \sum_{j=1}^{M} c_j \big(\xi_A \mathbb{E}_A[N_j(\tau)] + \xi_B \mathbb{E}_B[N_j(\tau)]\big),$$

and similarly for $\mathbb{E}[W_\pi]$. □

B.4. Proof of Proposition 1

We prove the inequality in three steps: First we apply Jensen's inequality, then we verify the necessary integrability condition, and finally we substitute the expression for $\mathbb{E}[W_\pi]$ from Lemma 3.

First, we need to verify that $\mathbb{E}[W_\pi] < \infty$ so that Jensen's inequality can be applied. By definition,

$$W_\pi = \sum_{t=1}^{\tau} T_{j_t,t},$$

where τ is the stopping time and $j_t \in \{1, \dots, M\}$ is the LLM selected at time t.

Using the law of iterated expectations and Assumption 1:

$$\mathbb{E}[W_\pi] = \mathbb{E}\Big[\sum_{t=1}^{\tau} T_{j_t,t}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\sum_{t=1}^{\tau} T_{j_t,t} \,\Big|\, \mathcal{H}_\tau\Big]\Big] = \mathbb{E}[S_\pi] \le \mu_{\max}\mathbb{E}[\tau] < \infty,$$

where $S_\pi := \sum_{t=1}^{\tau} \mu_{j_t}$ as defined in Lemma A.9.

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

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

Substituting the expression of $\mathbb{E}[W_\pi]$ in (22) into the above inequality, we have:

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

B.5. Proof of Lemma 4

We prove this by showing that any minimizer of (LB) must satisfy both constraints with equality. Suppose $(n^{A,*}, n^{B,*})$ is a minimizer of F_α subject to (20), but

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

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

Define the scaling factor

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

Since $\sum_{\ell=1}^{M} I_{\ell,A} n^\star_{\ell,A} > S_A(\alpha)$ by assumption, we have $\lambda < 1$. Consider the scaled vector

$$\tilde{n}_{j,A} := \lambda\, n^\star_{j,A}, \qquad j = 1, \ldots, M,$$

and keep $\tilde{n}^B := n^{B,*}$ unchanged. Since the objective is non-decreasing in $n_{j,A}$'s, it can be verified that $(\tilde{n}^A, n^{B,*})$ is feasible and $F_\alpha(\tilde{n}^A, n^{B,*}) < F_\alpha(n^{A,*}, n^{B,*})$. Therefore, any minimizer must satisfy the first constraint with equality. The same argument applies to the second constraint. □

B.6. Proof of Lemma 5

We prove the statements under hypothesis A; the case B is analogous.

First, according to (16), under the policy $\pi^{(2)}_{j^\star_A, j^\star_B}$, we obtain $\mathbb{E}_A[L_\tau] = I_{j^\star_A,A}\, n_{j^\star_A,A} + I_{j^\star_B,A}\, n_{j^\star_B,A}$. Second, according to Lemma A.8, we have $\mathbb{E}_A[L_\tau] \le A_\alpha + C_\ell$. Combining the above two results, we have:

$$I_{j^\star_A,A}\, n_{j^\star_A,A} + I_{j^\star_B,A}\, n_{j^\star_B,A} \le A_\alpha + C_\ell. \tag{B.2}$$

First, since $I_{j^\star_B,A} > 0$, we have

$$I_{j^\star_A,A}\, n_{j^\star_A,A} = A_\alpha + C_\ell - I_{j^\star_B,A}\, n_{j^\star_B,A} \le A_\alpha + C_\ell = S_A(\alpha) + K_\alpha + C_\ell.$$

Second, since $I_{j^\star_B,A} > 0$ and $n_{j^\star_B,A} \le C^A_{\text{neg}}$, we have

$$I_{j^\star_A,A}\, n_{j^\star_A,A} = A_\alpha + C_\ell - I_{j^\star_B,A}\, n_{j^\star_B,A} \ge A_\alpha + C_\ell - I_{j^\star_B,A} C^A_{\text{neg}}$$

Moreover, according to Lemma A.11, we have $n_{j^\star_B,A} \le C^A_{\text{neg}}$ and $n_{j^\star_A,B} \le C^B_{\text{pos}}$. □

B.7. Proof of Proposition 2

We prove the result for $\mathbb{E}[C_\pi]$; the proof for $\mathbb{E}[W_\pi]$ is analogous. We can write the total query cost as

$$C_\pi = \sum_{t=1}^{\tau} c_{j_t} = c_{j^\star_A} N_{j^\star_A}(\tau) + c_{j^\star_B} N_{j^\star_B}(\tau)$$

Taking expectations under the prior (ξ_A, ξ_B) using the law of total expectation, we have:

$$\begin{aligned}
\mathbb{E}[C_\pi] &= \xi_A \mathbb{E}_A[C_\pi] + \xi_B \mathbb{E}_B[C_\pi] \\
&= \xi_A (c_{j^\star_A} n_{j^\star_A,A} + c_{j^\star_B} n_{j^\star_A,A}) + \xi_B (c_{j^\star_A} n_{j^\star_A,B} + c_{j^\star_B} n_{j^\star_B,B}) \\
&= c_{j^\star_A} \cdot \xi_A \cdot n_{j^\star_A,A} + c_{j^\star_B} \cdot \xi_B \cdot n_{j^\star_B,B} + \left(c_{j^\star_A} \cdot \xi_B \cdot n_{j^\star_A,B} + c_{j^\star_B} \cdot \xi_A \cdot n_{j^\star_B,A} \right) \\
&= \kappa_{j^\star_A,A} \cdot \xi_A \cdot S_A(\alpha) \cdot \frac{n_{j^\star_A,A} \cdot I_{j^\star_A,A}}{S_A(\alpha)} + \kappa_{j^\star_B,B} \cdot \xi_B \cdot S_B(\alpha) \cdot \frac{n_{j^\star_B,B} \cdot I_{j^\star_B,B}}{S_B(\alpha)} \\
&\quad + \left(c_{j^\star_A} \cdot \xi_B \cdot n_{j^\star_A,B} + c_{j^\star_B} \cdot \xi_A \cdot n_{j^\star_B,A} \right).
\end{aligned}$$

According to Lemma 5, we can deduce that $\mathbb{E}[C_\pi] = \xi_A S_A(\alpha) \kappa_{j^\star_A,A} + \xi_B S_B(\alpha) \kappa_{j^\star_B,B} + O(1)$. □

B.8. Proof of Lemma 6

The first statement is proved in Lemma A.10. In the following, we prove the second statement when $\theta = A$. The proof for the case where $\theta = B$ is similar.

Define $\Delta := \sqrt{\mathbb{E}_A[\tau] \cdot \log \mathbb{E}_\theta[\tau]}$. We let $N^B = \sum_{s=1} \mathbf{1}\{j_s = j_B^*\}$ and $N^A = \sum_{s=1} \mathbf{1}\{j_s = j_A^*\}$. We first consider $\mathbb{P}_A(\tau - E_A[\tau] > \Delta)$. We let $\bar{n} = c_1 \cdot \log(\mathbb{E}_A[\tau])$, then we have

$$\mathbb{P}_A(\tau > \mathbb{E}_A[\tau] + \Delta) \le \mathbb{P}_A(N^B \ge \bar{n}) + \mathbb{P}_A\left(\tau > \mathbb{E}_A[\tau] + \Delta, N^B \le \bar{n}\right). \tag{B.3}$$

Recall that for any t, we have $\gamma_1, \gamma_2 > 0$ such that

$$\mathbb{P}_A(j_t = j_B^*) = \mathbb{P}(L_{t-1} < \delta) \le \mathbb{P}(\tilde{L}_{t-1} < \delta) \le \gamma_1 \exp(-\gamma_2 t) \quad \text{for all } t \ge 1.$$

where $\tilde{L}_{t-1}$ is the auxiliary random defined in the proof of Lemma A.11 and the inequality has been proved in Lemma A.11. This inequality further implies that under proper choice of c_1 in $\bar{n} = c_1 \cdot \log(\mathbb{E}_A[\tau])$, we have

$$\mathbb{P}_A(N^B \ge \bar{n}) \le \gamma_1 \exp(-\gamma_2 \cdot \bar{n}) = O\left(\frac{1}{(\mathbb{E}_A[\tau])^2}\right) = O\left(\frac{1}{(m_\alpha^A)^2}\right), \tag{B.4}$$

where the inequality holds since if $N^B \ge \bar{n}$, then the last time when j_B^* is queried before τ must be greater than $\bar{n}$.

We now bound $\mathbb{P}_A\left(\tau > \mathbb{E}_A[\tau] + \Delta, N^B \le \bar{n}\right)$, according to Lemma A.7, we have

$$A_\alpha - C_\ell \le L_\tau \le A_\alpha + C_\ell$$

On the event $\{\tau > t, \theta = A\}$ we must have

$$L_t = \sum_{s=1}^{t} I_{j_s}^A + M_t^{(A)} = I_{j_A^*}^A \cdot N^A + I_{j_B^*}^A \cdot N^B + M_t^{(A)} < A_\alpha$$

So if we let $t := \lfloor \mathbb{E}_A[\tau] + \Delta \rfloor - \bar{n}$ for some constant $\bar{n} \in \mathbf{Z}_+$,

$$\{\tau > \mathbb{E}_A[\tau] + \Delta, N^B \le \bar{n}\} \subseteq \left\{M_t^{(A)} \le A_\alpha - t \cdot I_{j_A^*}^A\right\}.$$

Since $\tau = N^A + N^B$ and $\mathbb{E}_A[N^B] = O(1)$, when $\Delta = k \cdot \sqrt{\mathbb{E}_A[\tau] \cdot \log \mathbb{E}_A[\tau]}$ for some $k > 0$, we have

$$tI_{j_A^*}^A \ge (\mathbb{E}_A[\tau] + \Delta - \bar{n} - 1) \cdot I_{j_A^*}^A \ge \mathbb{E}_A\left[\sum_{s=1}^{\tau} I_{j_s}\right] + (\Delta - \bar{n} - 1) \cdot I_{j_A^*}^A + \mathbb{E}_A[N^B] \cdot (I_{j_A^*}^A - I_{j_B^*}^A) = A_\alpha + c_2 \cdot \Delta + K_2$$

for some constant $c_2, K_2 > 0$ that are independent of α. Therefore, under such choice of Δ, we have

$$tI_{j_A^*}^A - A_\alpha \ge c_3 \cdot \Delta \text{ for some } c_3 > 0.$$

Applying Lemma A.6, we have

$$\mathbb{P}_A\left(\tau > \mathbb{E}_A[\tau] + \Delta, N^B \le \bar{n}\right) \le \mathbb{P}_A\left(M_t^{(A)} \le A_\alpha - t \cdot I_{j_A^*}^A\right) \le 2\exp\left(-\frac{c_3^2 \Delta^2}{2\left(v_\ell^2 t + 2C_\ell \cdot c_3 \Delta\right)}\right)$$

Since $\mathbb{E}_A[\tau] + \Delta - \bar{n} - 1 \le t \le \mathbb{E}_A[\tau] + \Delta - \bar{n}$ and $\mathbb{E}_A[\tau] = O(m_\alpha^A) = O(\log \frac{1}{\alpha})$, there must exist some $k > 0$ such that when $\Delta = k \cdot \sqrt{\mathbb{E}_A[\tau] \cdot \log \mathbb{E}_A[\tau]}$, we have

$$\frac{c_3^2 \Delta^2}{2\left(v_\ell^2 t + 2C_\ell \cdot c_3 \Delta\right)} \ge 2 \log \mathbb{E}_A[\tau]$$

Hence

$$\mathbb{P}_A\left(\tau > \mathbb{E}_A[\tau] + \Delta, N^B \le \bar{n}\right) \le \frac{1}{(\mathbb{E}_A[\tau])^2} = O\left(\frac{1}{(m_\alpha^A)^2}\right) \tag{B.5}$$

Finally, by (B.3), (B.4) and (B.5), we have

$$\mathbb{P}_A(\tau - E_A[\tau] > \Delta) = O\left(\frac{1}{(m_\alpha^A)^2}\right).$$

We now bound the other side $\mathbb{P}_A\left(\tau < \mathbb{E}_A[\tau] - \Delta\right)$. On the event $\{\tau \le t, \theta = A\}$ for some $t > 0$, we have

$$I_{j_A^*}^A \cdot t + M_{\tau\wedge t}^{(A)} > L_\tau = \sum_{s=1}^{t} I_{j_s}^A + M_\tau^{(A)} > A_\alpha$$

So if we let $t := \lceil \mathbb{E}_A[\tau] - \Delta \rceil$, this implies.

$$M_{\tau\wedge t}^{(A)} > A_\alpha - t \cdot I_{j_A^*}^A$$

Since $\tau = N^A + N^B$ and $\mathbb{E}_A[N^B] = O(1)$, when $\Delta = k \cdot \sqrt{\mathbb{E}_A[\tau] \cdot \log \mathbb{E}_A[\tau]}$ for some $k > 0$, we have

$$tI_{j_A^*}^A \le (\mathbb{E}_A[\tau] - \Delta + 1) \cdot I_{j_A^*}^A \le \mathbb{E}_A\left[\sum_{s=1}^{\tau} I_{j_s}\right] - \Delta \cdot I_{j_A^*}^A + \mathbb{E}_A[N^B] \cdot |I_{j_A^*}^A - I_{j_B^*}^A| = A_\alpha - c_3 \cdot \Delta + K_3$$

for some constant $c_3, K_3 > 0$ that are independent of α. Therefore, under such choice of Δ, we have

$$A_\alpha - tI_{j_A^*}^A \ge c_4 \cdot \Delta \text{ for some } c_4 > 0.$$

Applying Lemma A.6, we have

$$\mathbb{P}_A\left(\tau < \mathbb{E}_A[\tau] - \Delta\right) \le \mathbb{P}_A\left(M_{\tau\wedge t}^{(A)} > c_4 \cdot \Delta\right) \le 2\exp\left(-\frac{c_4^2\Delta^2}{2\left(v_\ell^2 t + 2C_\ell \cdot c_4\Delta\right)}\right)$$

Since $\mathbb{E}_A[\tau] - \Delta - 1 \le t \le \mathbb{E}_A[\tau] - \Delta$ and $\mathbb{E}_A[\tau] = O(m_\alpha^A) = O(\frac{1}{\alpha})$, there must exist some $k > 0$ such that when $\Delta = k \cdot \sqrt{\mathbb{E}_A[\tau] \cdot \log \mathbb{E}_A[\tau]}$, we have

$$\frac{c_4^2\Delta^2}{2\left(v_\ell^2 t + 2C_\ell \cdot c_4\Delta\right)} \ge 2\log \mathbb{E}_A[\tau]$$

Therefore,

$$\mathbb{P}_A\left(\tau < \mathbb{E}_A[\tau] - \Delta\right) = O\left(\frac{1}{(m_\alpha^A)^2}\right) \tag{B.6}$$

Combining with (B.5), we have

$$\mathbb{P}_A\left(|\tau - \mathbb{E}_A[\tau]| \le \Delta\right) = O\left(\frac{1}{(m_\alpha^A)^2}\right). \tag{B.7}$$

□

B.9. Proof of Claim 1 in Appendix C

Write

$$\mu_A := \mu_{j_A^\star}, \qquad \mu_B := \mu_{j_B^\star}, \qquad N_A := \sum_{t=1}^{\tau} \mathbf{1}\{j_t = j_A^\star\}, \qquad N_B := \sum_{t=1}^{\tau} \mathbf{1}\{j_t = j_B^\star\}.$$

Since $N_A + N_B = \tau$, we have

$$S_\pi = \mu_A N_A + \mu_B N_B = \mu_A \tau + (\mu_B - \mu_A) N_B.$$

Hence

$$m_\alpha^A = \mu_A \mathbb{E}_A[\tau] + (\mu_B - \mu_A)\mathbb{E}_A[N_B],$$

and therefore

$$S_\pi - m_\alpha^A = \mu_A(\tau - \mathbb{E}_A[\tau]) + (\mu_B - \mu_A)(N_B - \mathbb{E}_A[N_B]).$$

Using $(x+y)^2 \le 2x^2 + 2y^2$, we obtain

$$\mathbb{E}_A[(S_\pi - m_\alpha^A)^2] \le 2\mu_A^2 \operatorname{Var}_A(\tau) + 2(\mu_B - \mu_A)^2 \mathbb{E}_A[(N_B - \mathbb{E}_A[N_B])^2]. \tag{B.8}$$

By Lemma A.11, we have $\mathbb{E}_A[N_B^2] \le C_B$, implying that the second term is $O(1)$. It remains to bound $\operatorname{Var}_A(\tau)$. Let

$$I_A := I_{j_A^\star, A}, \qquad I_B := I_{j_B^\star, A}.$$

Under $\theta = A$, the drift-martingale decomposition gives

$$L_\tau = I_A N_A + I_B N_B + M_\tau^{(A)} = I_A \tau + (I_B - I_A) N_B + M_\tau^{(A)},$$

so

$$\tau - \mathbb{E}_A[\tau] = \frac{L_\tau - \mathbb{E}_A[L_\tau]}{I_A} - \frac{I_B - I_A}{I_A}(N_B - \mathbb{E}_A[N_B]) - \frac{M_\tau^{(A)}}{I_A}.$$

Hence, by $(x+y+z)^2 \le 3x^2 + 3y^2 + 3z^2$,

$$\operatorname{Var}_A(\tau) \le \frac{3}{I_A^2}\mathbb{E}_A[(L_\tau - \mathbb{E}_A[L_\tau])^2] + \frac{3(I_B - I_A)^2}{I_A^2}\mathbb{E}_A[(N_B - \mathbb{E}_A[N_B])^2] + \frac{3}{I_A^2}\mathbb{E}_A[(M_\tau^{(A)})^2].$$

Now:

- We first show that

  $$\mathbb{E}_A[(L_\tau - \mathbb{E}_A[L_\tau])^2] = O(1).$$

  By Lemma A.8, we have

  $$\mathbb{E}_A[L_\tau] = A_\alpha + O(1).$$

  Therefore,

  $$\mathbb{E}_A[(L_\tau - \mathbb{E}_A[L_\tau])^2] \le 2\mathbb{E}_A[(L_\tau - A_\alpha)^2] + 2(\mathbb{E}_A[L_\tau] - A_\alpha)^2.$$

  Since $\mathbb{E}_A[L_\tau] - A_\alpha = O(1)$ by Lemma A.8, it remains to show that

  $$\mathbb{E}_A[(L_\tau - A_\alpha)^2] = O(1).$$

  We decompose according to the terminal decision:

  $$\mathbb{E}_A[(L_\tau - A_\alpha)^2] = \mathbb{E}_A[(L_\tau - A_\alpha)^2 \mathbf{1}_{\{D=A\}}] + \mathbb{E}_A[(L_\tau - A_\alpha)^2 \mathbf{1}_{\{D=B\}}].$$

On the event $\{D = A\}$, Lemma A.7 implies

$$A_\alpha \le L_\tau \le A_\alpha + C_\ell,$$

hence

$$(L_\tau - A_\alpha)^2 \mathbf{1}_{\{D=A\}} \le C_\ell^2.$$

Therefore,

$$\mathbb{E}_A[(L_\tau - A_\alpha)^2 \mathbf{1}_{\{D=A\}}] \le C_\ell^2.$$

On the event $\{D = B\}$, Lemma A.7 gives

$$-B_\alpha - C_\ell \le L_\tau \le -B_\alpha,$$

so

$$|L_\tau - A_\alpha| \le A_\alpha + B_\alpha + C_\ell.$$

Thus

$$\mathbb{E}_A[(L_\tau - A_\alpha)^2 \mathbf{1}_{\{D=B\}}] \le (A_\alpha + B_\alpha + C_\ell)^2 \, \mathbb{P}_A(D = B).$$

By Lemma A.3,

$$\mathbb{P}_A(D = B) \le e^{-B_\alpha} = O(\alpha).$$

Moreover, since $A_\alpha = \log(1/\alpha) - \delta + O(\alpha)$ and $B_\alpha = \log(1/\alpha) + \delta + O(\alpha)$, we have

$$A_\alpha + B_\alpha = 2\log(1/\alpha) + O(1) = O(\log(1/\alpha)).$$

Hence

$$(A_\alpha + B_\alpha + C_\ell)^2 \, \mathbb{P}_A(D = B) = O\big(\log^2(1/\alpha)\big) \cdot O(\alpha) = O\big(\alpha \log^2(1/\alpha)\big) = o(1).$$

Combining the two parts, we obtain

$$\mathbb{E}_A[(L_\tau - A_\alpha)^2] \le C_\ell^2 + o(1) = O(1).$$

Therefore,

$$\mathbb{E}_A[(L_\tau - \mathbb{E}_A[L_\tau])^2] \le 2\,O(1) + 2\,O(1) = O(1).$$

- We bound the second moment of N_B using the comparison-walk quantity

$$N_{\text{neg}} := \sum_{t \ge 1} \mathbf{1}\{\widetilde{L}_t < -\delta\}.$$

By the proof of Lemma A.11, under $\theta = A$ we have

$$N_B \le N_{\text{neg}} + 1 \qquad \text{a.s.}$$

and there exist constants $c, \gamma > 0$ such that

$$\mathbb{P}_A(\widetilde{L}_t < -\delta) \le c e^{-\gamma t}, \qquad t \ge 1.$$

Therefore,

$$\mathbb{E}_A[N_{\text{neg}}^2] = \sum_{t\geq 1} \mathbb{P}_A(\widetilde{L}_t < -\delta) + 2 \sum_{1\leq s<t} \mathbb{P}_A(\widetilde{L}_s < -\delta, \widetilde{L}_t < -\delta)$$

$$\leq \sum_{t\geq 1} \mathbb{P}_A(\widetilde{L}_t < -\delta) + 2\sum_{t\geq 2}(t-1)\mathbb{P}_A(\widetilde{L}_t < -\delta) \leq \sum_{t\geq 1} ce^{-\gamma t} + 2\sum_{t\geq 2}(t-1)ce^{-\gamma t} < \infty.$$

Hence there exists a constant $C_B < \infty$, independent of α, such that

$$\mathbb{E}_A[N_B^2] \leq \mathbb{E}_A[(N_{\text{neg}}+1)^2] \leq C_B.$$

It follows that

$$\mathbb{E}_A[(N_B - \mathbb{E}_A[N_B])^2] = \text{Var}_A(N_B) \leq \mathbb{E}_A[N_B^2] \leq C_B = O(1).$$

- Since $(M_t^{(A)})_{t\geq 0}$ is a square-integrable martingale, the process

$$(M_t^{(A)})^2 - V_t^{(A)}$$

is also a martingale. Hence for each $n \geq 1$, optional stopping at $\tau \wedge n$ gives

$$\mathbb{E}_A\left[(M_{\tau\wedge n}^{(A)})^2 - V_{\tau\wedge n}^{(A)}\right] = 0,$$

so

$$\mathbb{E}_A[(M_{\tau\wedge n}^{(A)})^2] = \mathbb{E}_A[V_{\tau\wedge n}^{(A)}].$$

By Lemma A.10,

$$V_{\tau\wedge n}^{(A)} \leq v_\ell^2(\tau \wedge n) \quad \text{a.s.},$$

therefore

$$\mathbb{E}_A[(M_{\tau\wedge n}^{(A)})^2] \leq v_\ell^2\, \mathbb{E}_A[\tau \wedge n].$$

Letting $n \to \infty$, Fatou's lemma on the left-hand side and monotone convergence on the right-hand side yield

$$\mathbb{E}_A[(M_\tau^{(A)})^2] \leq v_\ell^2\, \mathbb{E}_A[\tau].$$

Therefore

$$\text{Var}_A(\tau) \leq C_1 + C_2\mathbb{E}_A[\tau] \leq C_3\mathbb{E}_A[\tau],$$

since $\mathbb{E}_A[\tau] \geq 1$. Returning to (B.8), we conclude that

$$\mathbb{E}_A[(S_\pi - m_\alpha^A)^2] \leq C_A\mathbb{E}_A[\tau]$$

for some constant $C_A < \infty$ independent of α. Finally, write

$$G_\alpha^A = \{\theta = A\} \cap \widehat{G}_\alpha^A.$$

Because conditioning on G_α^A forces $\theta = A$,

$$\mathbb{E}[(S_\pi - m_\alpha^A)^2 \mid G_\alpha^A] = \mathbb{E}_A[(S_\pi - m_\alpha^A)^2 \mid \widehat{G}_\alpha^A] \leq \frac{\mathbb{E}_A[(S_\pi - m_\alpha^A)^2]}{\mathbb{P}_A(\widehat{G}_\alpha^A)}.$$

Using the unconditional bound above,

$$\mathbb{E}[(S_\pi - m_\alpha^A)^2 \mid G_\alpha^A] \le \frac{C_A \, \mathbb{E}_A[\tau]}{\mathbb{P}_A(\widehat{G}_\alpha^A)}.$$

Since

$$\mathbb{P}(G_\alpha^A) = \xi_A \mathbb{P}_A(\widehat{G}_\alpha^A) \qquad \text{and} \qquad \mathbb{E}[\tau] \ge \xi_A \mathbb{E}_A[\tau],$$

we obtain

$$\mathbb{E}[(S_\pi - m_\alpha^A)^2 \mid G_\alpha^A] \le \frac{C \, \mathbb{E}[\tau]}{\mathbb{P}(G_\alpha^A)}.$$

This proves the claim. □

Appendix C: Proof of Theorem 1

In this section, we first establish the upper bound in (10), and then prove the lower bound in (11).

First, by Proposition 2, we have

$$\mathbb{E}[C_\pi] = \xi_A S_A(\alpha) \, \kappa_{j_A^\star, A} + \xi_B S_B(\alpha) \, \kappa_{j_B^\star, B} + O(1), \quad \mathbb{E}[W_\pi] = \xi_A S_A(\alpha) \, \eta_{j_A^\star, A} + \xi_B S_B(\alpha) \, \eta_{j_B^\star, B} + O(1).$$

Define $\Lambda_\alpha := \xi_A S_A(\alpha) \, \eta_{j_A^\star, A} + \xi_B S_B(\alpha) \, \eta_{j_B^\star, B}$. We have $\mathbb{E}[W_\pi] = \Lambda_\alpha + O(1) = \Theta\left(\log(1/\alpha)\right)$. In the following, it suffices to prove that $\mathbb{E}[g(W_\pi)] - g(\mathbb{E}[W_\pi]) = O\big((\log(1/\alpha))^{\rho-1}\big)$.

Note that

$$\mathbb{E}[g(W_\pi)] = \xi_A \cdot \mathbb{E}_A[g(W_\pi)] + \xi_B \cdot \mathbb{E}_B[g(W_\pi)] \tag{C.1}$$

We let $m_\alpha^A = \mathbb{E}_A[W_\pi], m_\alpha^B = \mathbb{E}_B[W_\pi]$. Then

$$m_\alpha = \xi_A \cdot m_\alpha^A + \xi_B \cdot m_\alpha^B$$

Since g function is polynomial, we have

$$Z_\alpha^\rho = \frac{g(W_\pi)}{g(m_\alpha^A)}, \qquad \hat{Z}_\alpha^\rho = \frac{g(W_\pi)}{g(m_\alpha^B)} \tag{C.2}$$

where $Z_\alpha = W_\pi / m_\alpha^A, \hat{Z}_\alpha = W_\pi / m_\alpha^B$. Observe that for any $q > 0$, the map $z \mapsto z^q$ is C^2, by Taylor's theorem with Lagrange remainder around $z = 1$, there exists ξ between 1 and z such that

$$z^q = f(1) + f'(1)(z-1) + \frac{1}{2} f''(\xi)(z-1)^2 = 1 + q(z-1) + \frac{1}{2} f''(\xi)(z-1)^2 \tag{C.3}$$

where $\xi \le \max\{z, 1\}$. Therefore, we have

$$\mathbb{E}_A[Z_\alpha^q - 1] = \mathbb{E}_A\left[q(Z_\alpha - 1) + \frac{1}{2} \cdot f''(\xi_{Z_\alpha}) \cdot (Z_\alpha - 1)^2\right] = \mathbb{E}_A\left[\frac{1}{2} \cdot f''(\xi_{Z_\alpha}) \cdot (Z_\alpha - 1)^2\right] \tag{C.4}$$

$$\mathbb{E}_B[\hat{Z}_\alpha^q - 1] = \mathbb{E}_B\left[q(\hat{Z}_\alpha - 1) + \frac{1}{2} \cdot f''(\xi_{\hat{Z}_\alpha}) \cdot (\hat{Z}_\alpha - 1)^2\right] = \mathbb{E}_B\left[\frac{1}{2} \cdot f''(\xi_{\hat{Z}_\alpha}) \cdot (\hat{Z}_\alpha - 1)^2\right] \tag{C.5}$$

Here we use ξ_{Z_α} ($\xi_{\hat{Z}_\alpha}$) since its value depends on the realized $Z_\alpha(\hat{Z}_\alpha)$.

Define Events To facilitate our analysis, we define "typical" event and "tail" events.

$$G_\alpha^A := \{|Z_\alpha - 1| \le \varepsilon, \theta = A\}, \quad G_\alpha^B := \{|\hat{Z}_\alpha^A - 1| \le \hat{\varepsilon}, \theta = B\}$$

and

$$\bar{G}_\alpha^A := \{|Z_\alpha - 1| > \varepsilon, \theta = A\}, \quad \bar{G}_\alpha^B := \{|\hat{Z}_\alpha - 1| > \hat{\varepsilon}, \theta = B\}$$

We let $\varepsilon := \sqrt{1/m_\alpha^A} \cdot \log m_\alpha^A$ and $\hat{\varepsilon} := \sqrt{1/m_\alpha^B} \cdot \log m_\alpha^B$. Then G_α^A, G_α^B can be alternatively written as

$$G_\alpha^A = \{|W_\pi - m_\alpha^A| \le \sqrt{m_\alpha^A} \log m_\alpha^A, \theta = A\}, \quad G_\alpha^B = \{|W_\pi - m_\alpha^B| \le \sqrt{m_\alpha^B} \log m_\alpha^B, \theta = B\}$$

Thus, we can write

$$\mathbb{E}_A[Z_\alpha^q - 1] = \mathbb{E}_A[(Z_\alpha^q - 1) \cdot \mathbf{1}_{\{G_\alpha^A\}}] + \mathbb{E}_A[(Z_\alpha^q - 1) \cdot \mathbf{1}_{\{\bar{G}_\alpha^A\}}] \tag{C.6}$$

$$\mathbb{E}_B[\hat{Z}_\alpha^q - 1] = \mathbb{E}_B[(\hat{Z}_\alpha^q - 1) \cdot \mathbf{1}_{\{G_\alpha^B\}}] + \mathbb{E}_B[(\hat{Z}_\alpha^q - 1) \cdot \mathbf{1}_{\{\bar{G}_\alpha^B\}}] \tag{C.7}$$

And our goal is to show

$$\mathbb{E}_A[Z_\alpha^q - 1] = O\left(\frac{1}{m_\alpha}\right), \qquad \mathbb{E}_B[Z_\alpha^q - 1] = O\left(\frac{1}{m_\alpha}\right).$$

Once they hold, by (C.2), we have

$$\frac{\mathbb{E}[g(W_\pi)]}{\xi_A \cdot \mathbb{E}_A[g(m_\alpha^A)] + \xi_B \cdot \mathbb{E}_B[g(m_\alpha^B)]} = \frac{\xi_A \cdot \mathbb{E}_A[g(W_\pi)] + \xi_B \cdot \mathbb{E}_B[g(W_\pi)]}{\xi_A \cdot \mathbb{E}_A[g(m_\alpha^A)] + \xi_B \cdot \mathbb{E}_B[g(m_\alpha^B)]} = O\left(\frac{1}{m_\alpha}\right). \tag{C.8}$$

Typical Event G_α^A or G_α^B. Note that

$$W_\pi = \sum_{t=1}^{\tau} Y_t, \quad S_\pi = \sum_{t=1}^{\tau} \mu_{J_t} \quad \text{with} \quad \mu_j = \mathbb{E}[Y_t \mid J_t = j],$$

and

$$(W_\pi - m_\alpha^A)^2 \le 2 \cdot (W_\pi - S_\pi)^2 + 2 \cdot (S_\pi - m_\alpha^A)^2.$$

We now bound $\mathbb{E}[(W_\pi - S_\pi)^2 \mid G_\alpha^A]$ and $\mathbb{E}[(S_\pi - m_\alpha)^2 \mid G_\alpha^B]$ follow the same logics.

- For the first term, we have

$$\mathbb{E}[(W_\pi - S_\pi)^2 \mid G_\alpha^A] = \frac{\mathbb{E}[(W_\pi - S_\pi)^2 \mathbf{1}_{G_\alpha^A}]}{\mathbb{P}(G_\alpha^A)} \le \frac{\mathbb{E}[(W_\pi - S_\pi)^2]}{\mathbb{P}(G_\alpha^A)} \le \frac{\psi^2 \mathbb{E}[\tau]}{\mathbb{P}(G_\alpha^A)}.$$

  where the last inequality holds by Lemma A.10.

- For the second term, note that on G_α^A, We state a claim (whose proof is provided in Appendix B.9).

  Claim 1 There exists a constant $C < \infty$, independent of α, such that for all sufficiently small α,

$$\mathbb{E}\big[(S_\pi - m_\alpha^A)^2 \mid G_\alpha^A\big] \le \frac{C\,\mathbb{E}[\tau]}{\mathbb{P}(G_\alpha^A)}.$$

Therefore, when $\varepsilon \le \frac{1}{2}$, which holds when α is sufficiently small since $\varepsilon = \frac{\log m_\alpha^A}{\sqrt{m_\alpha^A}}$ and $m_\alpha = O(\log 1/\alpha)$, we have

$$\mathbb{E}[(Z_\alpha^q - 1) \cdot \mathbf{1}_{\{G_\alpha^A\}}] \le \mathbb{E}\left[\frac{f''(\xi_{Z_\alpha})}{2} \cdot (Z_\alpha - 1)^2 \cdot \mathbf{1}_{\{G_\alpha^A\}}\right] \le \frac{\bar{C} \cdot \mathbb{E}[\tau]}{(m_\alpha^A)^2}. \tag{TE-A}$$

where the last inequality holds since $f''(\xi_{Z_\alpha})$ is bounded by a constant on G_α^A for any $q > 0$. Similarly, we have

$$\mathbb{E}[(\hat{Z}_\alpha^q - 1) \cdot \mathbf{1}_{\{G_\alpha^B\}}] \le \mathbb{E}\left[\frac{f''(\xi_{Z_\alpha})}{2} \cdot (\hat{Z}_\alpha - 1)^2 \cdot \mathbf{1}_{\{G_\alpha^B\}}\right] \le \frac{\bar{C} \cdot \mathbb{E}[\tau]}{(m_\alpha^B)^2}. \tag{TE-B}$$

Tail Event $\bar{G}_\alpha^A = \{|Z_\alpha - 1| > \varepsilon, \theta = A\}$ **or** $\bar{G}_\alpha^B = \{|\hat{Z}_\alpha - 1| > \hat{\varepsilon}, \theta = B\}$. Note that on event $\bar{G}_\alpha^A$, we have

$$f''(\xi_{Z_\alpha}) \cdot (Z_\alpha - 1)^2 = \frac{q(q-1)\xi_{Z_\alpha}^{\rho-2}}{2} \cdot (Z_\alpha - 1)^2 \le \frac{q(q-1)}{2} \cdot \max\{Z_\alpha^\rho, Z_\alpha^2\} \le q(q-1) \cdot (Z_\alpha^\rho + Z_\alpha^2)$$

Note that, for any q, we have

$$\mathbb{E}[Z_\alpha^q \mathbf{1}_{\{\bar{G}_\alpha^A\}}] \le \mathbb{E}[Z_\alpha^q]^{1/2} \cdot [\mathbb{P}(\bar{G}_\alpha^A)]^{1/2} \tag{C.9}$$

We now show

$$\mathbb{E}[Z_\alpha^q] = O(1), \quad \mathbb{P}(\bar{G}_\alpha^A) = O\left(\frac{1}{(m_\alpha^A)^2}\right) \tag{C.10}$$

Bound $\mathbb{E}[Z_\alpha^q]$. Note that

$$Z_\alpha = \frac{W_\pi}{m_\alpha^A} \le \frac{S_\pi}{m_\alpha^A} + \frac{|W_\pi - S_\pi|}{m_\alpha^A}.$$

And for $q \ge 1$, since $\left(\frac{u+v}{2}\right)^q \le \frac{u^q + v^q}{2}$ always holds, we have

$$\mathbb{E}[Z_\alpha^q] \le 2^{q-1}\left(\mathbb{E}\Big[\Big(\frac{S_\pi}{m_\alpha}\Big)^q\Big] + \mathbb{E}\Big[\Big(\frac{|W_\pi - S_\pi|}{m_\alpha}\Big)^q\Big)\right). \tag{C.11}$$

(i) Bounding S_π/m_α. Since the set of LLMs is finite and $\mu_j < \infty$, let $\mu_{\max} := \max_j \mu_j$. Then $S_\pi \le \mu_{\max}\tau$, hence

$$\mathbb{E}\Big[\Big(\frac{S_\pi}{m_\alpha^A}\Big)^q\Big] \le \mu_{max}^q \cdot \mathbb{E}\Big[\Big(\frac{\tau}{m_\alpha^A}\Big)^q\Big].$$

According to Freeman Inequality, there exists some constants $K_0, c > 0$, such that for all sufficiently small α,

$$\mathbb{P}(\tau > K m_\alpha) \le e^{-cK m_\alpha} \qquad \text{for all } K \ge K_0. \tag{C.12}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\tau}{m_\alpha^A}\right)^q\right] &= q\int_0^\infty u^{q-1}\,\mathbb{P}\left(\frac{\tau}{m_\alpha} > u\right) du \\
&\le q\int_0^{K_0} u^{q-1}\,du \;+\; q\int_{K_0}^\infty u^{q-1}\,\mathbb{P}(\tau > u m_\alpha)\,du \\
&\le K_0^q \;+\; q\int_{K_0}^\infty u^{q-1}\,e^{-cum_\alpha}\,du.
\end{aligned} \tag{C.13}$$

For the second term, we let $v = cum_\alpha$, then $u = v/(cm_\alpha)$ and $du = dv/(cm_\alpha)$, which implies

$$\begin{aligned}
\int_{K_0}^\infty u^{q-1}e^{-cum_\alpha}\,du &= \frac{1}{(cm_\alpha)^q}\int_{cK_0m_\alpha}^\infty v^{q-1}e^{-v}\,dv \\
&\le \frac{1}{(cm_\alpha)^q}\int_0^\infty v^{q-1}e^{-v}\,dv = \frac{\Gamma(q)}{(cm_\alpha)^q}.
\end{aligned} \tag{C.14}$$

Combining (C.13)–(C.14) yields

$$\mathbb{E}\left[\left(\frac{\tau}{m_\alpha}\right)^q\right] \le K_0^q + \frac{q\,\Gamma(q)}{(cm_\alpha)^q} \le K_0^q + \frac{q\,\Gamma(q)}{c^q} =: C_q,$$

where C_q is a constant independent of α. Consequently, since $S_\pi \le \mu_{\max}\tau$, we have for the same q,

$$\mathbb{E}\left[\left(\frac{S_\pi}{m_\alpha}\right)^q\right] \le \mu_{\max}^q\,\mathbb{E}\left[\left(\frac{\tau}{m_\alpha}\right)^q\right] \le \mu_{\max}^q C_q \tag{C.15}$$

(ii) Bounding $(W_\pi - S_\pi)/m_\alpha$. By the conditional sub-Gaussian bound for the waiting-time noise, for each fixed q there exists $M_q < \infty$ such that

$$\mathbb{E}\big[|W_\pi - S_\pi|^q \mid \mathcal{H}_\tau\big] \le M_q \cdot (\psi^2 \tau)^{q/2}.$$

Taking expectations and dividing by m_α^q yields

$$\mathbb{E}\Big[\Big(\frac{|W_\pi - S_\pi|}{m_\alpha}\Big)^q\Big] \le \frac{M_q \psi^q}{m_\alpha^q}\,\mathbb{E}[\tau^{q/2}] = M_q \cdot \psi^q\,\mathbb{E}\Big[\Big(\frac{\tau}{m_\alpha}\Big)^{q/2}\Big] \cdot m_\alpha^{-q/2}.$$

By similar analysis as in (i), we can show that for some $\tilde{C}_q$ that is independent of α, we have

$$\mathbb{E}\Big[\Big(\frac{|W_\pi - S_\pi|}{m_\alpha}\Big)^q\Big] < \tilde{C}_q \cdot m_\alpha^{-q/2}. \tag{C.16}$$

Combining (C.11), (C.15), and (C.16), for some constant $\hat{C}_q$, we have

$$\mathbb{E}[Z_\alpha^q] < \hat{C}_q$$

Bound $\mathbb{P}(\bar{G}_\alpha^A)$. It remains to show that

$$\mathbb{P}(\bar{G}_\alpha^A) = O\left(\frac{1}{(m_\alpha^A)^2}\right) \tag{C.17}$$

Recall that $x_\alpha^A := \varepsilon m_\alpha^A$. By definition,

$$\mathbb{P}(\bar{G}_\alpha^A) = \xi_A \mathbb{P}_A\big(|W_\pi - m_\alpha^A| > x_\alpha^A\big) \tag{C.18}$$

Recall that

$$S_\pi := \sum_{t=1}^{\tau} \mu_{j_t}.$$

Then

$$W_\pi - m_\alpha^A = (W_\pi - S_\pi) + (S_\pi - m_\alpha^A),$$

so by the union bound,

$$\mathbb{P}_A\big(|W_\pi - m_\alpha^A| > x_\alpha^A\big) \le \mathbb{P}_A\left(|W_\pi - S_\pi| > \frac{x_\alpha^A}{2}\right) + \mathbb{P}_A\left(|S_\pi - m_\alpha^A| > \frac{x_\alpha^A}{2}\right). \tag{C.19}$$

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

Part 1. We first consider $\mathbb{P}_A(|W_\pi - S_\pi| > x_\alpha^A/2)$. We let $E := \{|\tau - \mathbb{E}_A[\tau]| \le \Delta\}$ where

$$\Delta = k \cdot \sqrt{\mathbb{E}_A[\tau] \cdot \log \mathbb{E}_A[\tau]} \text{ for some } k > 0$$

Note that

$$\mathbb{P}_A\left(|W_\pi - S_\pi| > \frac{x_\alpha^A}{2}, E\right) \le \mathbb{P}_A\left(|W_\pi - S_\pi| > \frac{x_\alpha^A}{2}, E\right) + P_A(\bar{E}) \tag{C.20}$$

According to Lemma A.9,

$$\mathbb{P}_\theta\big(|W_\pi - S_\pi| \ge \frac{x_\alpha^A}{2} \,\big|\, \mathcal{H}_\tau, \theta, E\big) \;\le\; 2\exp\Big(-\frac{(x_\alpha^A)^2}{4\psi^2 \cdot \tau}\Big)$$

Thus

$$\mathbb{P}_A\left(|W_\pi - S_\pi| > \frac{x_\alpha^A}{2}\Big|E\right) \le 2\mathbb{E}\left[\exp\Big(-\frac{(x_\alpha^A)^2}{4\psi^2\cdot\tau}\Big)\Big|E\right] \le 2\mathbb{E}\left[\exp\Big(-\frac{(x_\alpha^A)^2}{4\psi^2\cdot(\mathbb{E}_A[\tau]+\Delta])}\Big)\right],$$

which further implies that

$$\mathbb{P}_A\left(|W_\pi - S_\pi| > \frac{x_\alpha^A}{2}, E\right) \le 2\mathbb{E}\left[\exp\Big(-\frac{(x_\alpha^A)^2}{4\psi^2\cdot(\mathbb{E}_A[\tau]+\Delta])}\Big)\right]\cdot\mathbb{P}_A(E). \tag{C.21}$$

Then, by Lemma 6, we have $\mathbb{P}_A(\bar{E}) = \mathbb{P}_A(|\tau - \mathbb{E}_A[\tau]| > \Delta) = O\left(\frac{1}{(m_\alpha^A)^2}\right)$ when $\Delta = k\cdot\sqrt{\mathbb{E}_A[\tau]\cdot\log\mathbb{E}_A[\tau]}$ for some $k>0$.

Since

$$(x_\alpha^A)^2 = m_\alpha^A\cdot\log^2 m_\alpha^A,\quad \mathbb{E}_A[\tau] = O(m_\alpha^A),\quad m_\alpha^A = O\left(\log(1/\alpha)\right),$$

when $\alpha\to 0$, the RHS of (C.21) can be of the order $O\left(\frac{1}{(m_\alpha^A)^2}\right)$. Finally, by (C.20), we conclude that

$$\mathbb{P}_A\left(|W_\pi - S_\pi| > \frac{x_\alpha^A}{2}\right) = O\left(\frac{1}{(m_\alpha^A)^2}\right). \tag{C.22}$$

Part 2. Bound for $\mathbb{P}_A(S_\pi - m_\alpha^A > x_\alpha^A/2)$. We write

$$\mu_A := \mu_{j_A^\star},\qquad \mu_B := \mu_{j_B^\star},\qquad N_A := \sum_{t=1}^{\tau}\mathbf{1}\{j_t = j_A^\star\},\qquad N_B := \sum_{t=1}^{\tau}\mathbf{1}\{j_t = j_B^\star\}.$$

Since $N_A + N_B = \tau$,

$$S_\pi = \mu_A N_A + \mu_B N_B = \mu_A\tau + (\mu_B - \mu_A)N_B.$$

Taking $\mathbb{E}_A$ gives

$$m_\alpha^A = \mu_A\mathbb{E}_A[\tau] + (\mu_B - \mu_A)\mathbb{E}_A[N_B],$$

and therefore

$$S_\pi - m_\alpha^A = \mu_A(\tau - \mathbb{E}_A[\tau]) + (\mu_B - \mu_A)(N_B - \mathbb{E}_A[N_B]).$$

Thus,

$$\mathbb{P}_A\left(S_\pi - m_\alpha^A > \frac{x_\alpha^A}{2}\right) \le \mathbb{P}_A\left(\mu_A(\tau - \mathbb{E}_A[\tau]) > \frac{x_\alpha^A}{4}\right) + \mathbb{P}_A\left(|\mu_B - \mu_A|\,|N_B - \mathbb{E}_A[N_B]| > \frac{x_\alpha^A}{4}\right). \tag{C.23}$$

For the first term, since $\mathbb{E}_A[N_B] = O(1)$ by Lemma A.11, we have

$$m_\alpha^A = \mu_A\mathbb{E}_A[\tau] + O(1),$$

hence $\mathbb{E}_A[\tau] \le C_\tau m_\alpha^A$ for some constant $C_\tau > 0$ and all sufficiently small α. Therefore,

$$\left\{\mu_A(\tau - \mathbb{E}_A[\tau]) > \frac{x_\alpha^A}{4}\right\} \subseteq \left\{\tau > \mathbb{E}_A[\tau] + \frac{x_\alpha^A}{4\mu_A}\right\}$$

Since $x_\alpha^A > \Delta$ when α is sufficiently small, we have

$$\mathbb{P}_A\left(\mu_A(\tau - \mathbb{E}_A[\tau]) > \frac{x_\alpha^A}{4}\right) \le O\left(\frac{1}{(m_\alpha^A)^2}\right) \tag{C.24}$$

For the second term, we use the second-moment bound for the wrong-side count: there exists a constant $C_B < \infty$, independent of α, such that

$$\mathbb{E}_A[N_B^2] \le C_B.$$

Hence $\mathrm{Var}_A(N_B) \le C_B$, and Chebyshev's inequality yields

$$\mathbb{P}_A\left(|\mu_B - \mu_A|\,|N_B - \mathbb{E}_A[N_B]| > \frac{x_\alpha^A}{4}\right) \le \frac{16(\mu_B-\mu_A)^2\,\mathrm{Var}_A(N_B)}{(x_\alpha^A)^2} \le \frac{\hat{C}}{(m_\alpha^A)^2}. \tag{C.25}$$

Combining (C.24), and (C.25), we can bound the RHS of (C.23):

$$\mathbb{P}_A\left(S_\pi - m_\alpha^A > \frac{x_\alpha^A}{2}\right) \le O\left(\frac{1}{(m_\alpha^A)^2}\right). \tag{C.26}$$

Combination. Substituting (C.22) and (C.26) into (C.19), we have

$$\mathbb{P}_A\big(|W_\pi - m_\alpha^A| > x_\alpha^A\big) \le O\left(\frac{1}{(m_\alpha^A)^2}\right).$$

Finally, by (C.18), we have

$$\mathbb{P}(\bar{G}_\alpha^A) = O\left(\frac{1}{(m_\alpha^A)^2}\right).$$

According to (C.9), (C.10), we have

$$\mathbb{E}[Z_\alpha^q \mathbf{1}_{\{\bar{G}_\alpha^A\}}] = O\left(\frac{1}{m_\alpha^A}\right) \tag{C.27}$$

And by (TE-A), we have

$$\mathbb{E}[(Z_\alpha^q - 1)\mathbf{1}_{\{\bar{G}_\alpha^A\}}] + \mathbb{E}[(Z_\alpha^q - 1)\mathbf{1}_{\{G_\alpha^A\}}] = O\left(\frac{1}{m_\alpha^A}\right). \tag{C.28}$$

Similarly, we can show that

$$\mathbb{E}[(Z_\alpha^q - 1)\mathbf{1}_{\{\bar{G}_\alpha^B\}}] + \mathbb{E}[(Z_\alpha^q - 1)\mathbf{1}_{\{G_\alpha^B\}}] = O\left(\frac{1}{m_\alpha^B}\right). \tag{C.29}$$

Since m_α^A, m_α^B are both of the order $\log\frac{1}{\alpha}$, we obtain

$$\boxed{\mathbb{E}[C_\pi] + \mathbb{E}[g(W_\pi)] = \Phi_\alpha + O\big((\log(1/\alpha))^{\rho-1}\big).}$$

Lastly, we prove the lower bound in (11). The proof proceeds in four steps.

Step 1 (Hindsight relaxation). Fix $\alpha \in (0, 1/2)$ and consider the posterior-threshold stopping rule $\Pi(\alpha)$ in Definition 1. According to Remark 1, we only need to consider policies with $\mathbb{E}_A[\tau] < \infty$ and $\mathbb{E}_B[\tau] < \infty$. Define the *hindsight* (oracle) policy class $\Pi^\circ(\alpha)$ as the set of policies that are allowed to depend on the ground truth $\theta \in \{A, B\}$ in addition to the observable history, i.e.,

$$j_t = \phi_t^\circ(\theta, \mathcal{G}_{t-1}),$$

while using the same stopping time τ.

Intuitively, hindsight policies are strictly more powerful: they know in advance which hypothesis is true and can tailor their LLM choices accordingly (i.e., the posterior needs to satisfy the same stopping criterion), whereas admissible policies must learn this information sequentially.

Formally, every admissible policy is also a hindsight policy that simply ignores θ. Hence

$$\Pi(\alpha) \subseteq \Pi^\circ(\alpha),$$

and therefore

$$\inf_{\pi\in\Pi(\alpha)}\big(\mathbb{E}[C_\pi]+\mathbb{E}[g(W_\pi)]\big)\ \geq\ \inf_{\pi^\circ\in\Pi^\circ(\alpha)}\big(\mathbb{E}[C_{\pi^\circ}]+\mathbb{E}[g(W_{\pi^\circ})]\big).$$

Step 2 (Optimal hindsight policy. Let $(j_A^\star, j_B^\star)$ be the optimal two-LLM design indices in Theorem 1. Under hindsight, the ground truth θ is known in advance. Therefore, when $\theta = A$, it is optimal to always query the LLM with the best efficiency under A, namely $j_A^\star$; when $\theta = B$, it is optimal to always query $j_B^\star$. Hence, the optimal hindsight policy reduces to a single-LLM policy $\pi^\dagger$ under each hypothesis:

$$j_t \equiv \begin{cases} j_A^\star, & \theta = A, \\ j_B^\star, & \theta = B, \end{cases}$$

with the same stopping rule. To prove the lemma, it suffices to lower bound the performance of this benchmark policy and show

$$\mathbb{E}[C_{\pi^\dagger}]+\mathbb{E}[g(W_{\pi^\dagger})]\ \geq\ \Phi_\alpha\ +\ \Omega\bigg(\frac{g(\Lambda_\alpha)}{\Lambda_\alpha}\bigg). \tag{C.30}$$

Step 3 (Mean waiting time under $\pi^\dagger$ matches Λ_α up to $O(1)$). Write $\tau^\dagger$ and $W^\dagger$ for the stopping time and waiting time under $\pi^\dagger$. Under $\theta = A$, $\pi^\dagger$ always queries $j_A^\star$, hence by the drift identity at stopping (Proposition 2) and the expected boundary crossing bound (Lemma A.8 and Lemma A.7),

$$I_{j_A^\star,A}\,\mathbb{E}_A[\tau^\dagger]\ =\ \mathbb{E}_A[L_{\tau^\dagger}]\ =\ A_\alpha + O(1), \qquad \text{so} \qquad \mathbb{E}_A[\tau^\dagger]\ =\ \frac{A_\alpha}{I_{j_A^\star,A}} + O(1).$$

Similarly, under $\theta = B$,

$$I_{j_B^\star,B}\,\mathbb{E}_B[\tau^\dagger]\ =\ \mathbb{E}_B[-L_{\tau^\dagger}]\ =\ B_\alpha + O(1), \qquad \text{so} \qquad \mathbb{E}_B[\tau^\dagger]\ =\ \frac{B_\alpha}{I_{j_B^\star,B}} + O(1).$$

By the waiting-time decomposition, we have

$$\mathbb{E}_A[W^\dagger] = \mu_{j_A^\star}\mathbb{E}_A[\tau^\dagger], \qquad \mathbb{E}_B[W^\dagger] = \mu_{j_B^\star}\mathbb{E}_B[\tau^\dagger].$$

Bayes-averaging gives

$$m_\alpha^\dagger := \mathbb{E}[W^\dagger] = \xi_A\mu_{j_A^\star}\mathbb{E}_A[\tau^\dagger] + \xi_B\mu_{j_B^\star}\mathbb{E}_B[\tau^\dagger] = \xi_A A_\alpha \eta_{j_A^\star,A} + \xi_B B_\alpha \eta_{j_B^\star,B} + O(1),$$

where $\eta_{j,\theta} = \mu_j / I_{j,\theta}$. Recalling $S_A(\alpha) = A_\alpha - K_\alpha$ and $S_B(\alpha) = B_\alpha - K_\alpha$, we have

$$\Lambda_\alpha = \xi_A S_A(\alpha)\eta_{j_A^\star,A} + \xi_B S_B(\alpha)\eta_{j_B^\star,B} = \xi_A A_\alpha \eta_{j_A^\star,A} + \xi_B B_\alpha \eta_{j_B^\star,B} - K_\alpha(\xi_A\eta_{j_A^\star,A} + \xi_B\eta_{j_B^\star,B}).$$

Since $K_\alpha = O(\alpha\log(1/\alpha))$ is uniformly bounded for small α, it follows that

$$m_\alpha^\dagger = \Lambda_\alpha + O(1), \qquad \text{and} \qquad \Lambda_\alpha = \Theta(\log(1/\alpha)). \tag{C.31}$$

Step 4 (Quantitative Jensen gap via the same typical/tail decomposition as the upper bound proof). Let $m := \mathbb{E}[W^\dagger]$ and $Z := W^\dagger/m$. Then $\mathbb{E}[Z] = 1$ and $\mathbb{E}[(Z-1)^2] = \mathrm{Var}(W^\dagger)/m^2$. Using the same "typical event / tail event" structure as the proof of the upper bound (only the inequality direction changes), one obtains the lower bound

$$\mathbb{E}[g(W^\dagger)] - g(m)\ \geq\ c_2\, g(m)\,\mathbb{E}[(Z-1)^2]\ =\ c_2\, g(m)\,\frac{\mathrm{Var}(W^\dagger)}{m^2}, \tag{C.32}$$

for some constant $c_2 > 0$ and all sufficiently small α. (On the typical event $\{|Z-1| \le \varepsilon\}$, regular variation $g(x) = x^\rho L(x)$ and uniform convergence of L give a lower bound of $g(mZ)/g(m)$ by $(1-\varepsilon)Z^\rho$; then a quadratic lower bound for Z^ρ on $[1-\varepsilon, 1+\varepsilon]$ yields a contribution proportional to $\mathbb{E}[(Z-1)^2]$; the tail event is controlled using the same sub-Gaussian concentration and Cauchy–Schwarz steps as in Theorem 8.1.).

Note that Under $\theta = A$, conditional on $\tau^\dagger$ we have $\mathrm{Var}(W^\dagger \mid \tau^\dagger, \theta = A) = \mathrm{Var}(T_{j_A^\star,1})\,\tau^\dagger$. Taking expectations yields $\mathrm{Var}_A(W^\dagger) \ge \mathrm{Var}(T_{j_A^\star,1})\,\mathbb{E}_A[\tau^\dagger]$. Similarly, $\mathrm{Var}_B(W^\dagger) \ge \mathrm{Var}(T_{j_B^\star,1})\,\mathbb{E}_B[\tau^\dagger]$. Since $\mathbb{E}_A[\tau^\dagger] = \Theta(A_\alpha)$ and $\mathbb{E}_B[\tau^\dagger] = \Theta(B_\alpha)$, both are $\Theta(\log(1/\alpha))$, we obtain

$$\mathrm{Var}(W^\dagger) \ge c_0 \log(1/\alpha) \qquad \text{and hence} \qquad \frac{\mathrm{Var}(W^\dagger)}{(\mathbb{E}[W^\dagger])^2} \ge \frac{c_1}{\mathbb{E}[W^\dagger]}, \tag{C.33}$$

for some constants $c_0, c_1 > 0$ and all sufficiently small α, using $\mathbb{E}[W^\dagger] = \Theta(\log(1/\alpha))$ from (C.31).

Combining (C.32) with (C.33) gives

$$\mathbb{E}[g(W^\dagger)] - g(m) \ge c_3 \frac{g(m)}{m}.$$

Finally, since $m = \Lambda_\alpha + O(1)$ by (C.31) and $g \in RV_\rho$, the fixed-shift regular-variation bound used in Step 4 of the proof of Theorem 8.1 implies $\frac{g(m)}{m} = \Theta\big(\frac{g(\Lambda_\alpha)}{\Lambda_\alpha}\big)$. Therefore

$$\mathbb{E}[g(W^\dagger)] - g(\Lambda_\alpha) \ge \mathbb{E}[g(W^\dagger)] - g(m) \ge c_4 \frac{g(\Lambda_\alpha)}{\Lambda_\alpha} = \Omega\left((\log(1/\alpha))^{\rho-1}\right), \tag{C.34}$$

for some constant $c_4 > 0$ and all sufficiently small α. Combining with Step 1, this completes the proof.

Appendix D: Proof of Theorem 2

In this section, we provide the proof for Theorem 2. Let

$$F(x,z) = \xi_A \sum_{j=1}^M \kappa_{j,A} \cdot x_j + \xi_B \sum_{j=1}^M \kappa_{j,B} \cdot z_j + \xi_A g\left(\sum_{j=1}^M \eta_{j,A} \cdot x_j\right) + \xi_B g\left(\sum_{j=1}^M \eta_{j,B} \cdot z_j\right)$$

$$\mathcal{X}_S = \left\{(x,z) \in \mathbb{R}_+^M \times \mathbb{R}_+^M : \sum_{j=1}^M x_j = S_A(\alpha),\ \sum_{j=1}^M z_j = S_B(\alpha)\right\}.$$

Since $\mathcal{X}_S$ is compact and F is continuous and convex, a minimizer $(x^\star, z^\star)$ exists. For $\theta \in \{A, B\}$, write

$$y_\theta^\star := \xi_\theta \sum_{j=1}^M \eta_{j,\theta} \cdot x_j^\star.$$

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

$$\xi_A \kappa_{j,A} + \xi_A \eta_{j,A} g'(y_A^\star) \begin{cases} = \lambda_x, & \text{if } x_j^\star > 0, \\ \ge \lambda_x, & \text{if } x_j^\star = 0, \end{cases}, \qquad \xi_B \kappa_{j,B} + \xi_B \eta_{j,B} g'(y_B^\star) \begin{cases} = \lambda_z, & \text{if } z_j^\star > 0, \\ \ge \lambda_z, & \text{if } z_j^\star = 0. \end{cases} \tag{D.1}$$

In the following, we prove that there exists an optimal solution with $x^\star = e_{j^\star}$ and $z^\star = e_{k^\star}$. We present the proof for $x^\star = e_{j^\star}$ in the following and the proof for $z^\star = e_{k^\star}$ is similar.

Suppose two distinct indices $j \ne k$ are active ($x_j^\star, x_k^\star > 0$). Subtracting the equalities in (D.1) gives

$$(\kappa_{j,A} - \kappa_{k,A}) + g'(y_A^\star)\,(\eta_{j,A} - \eta_{k,A}) = 0. \tag{D.2}$$

Case 1: $\eta_{j,A} \ne \eta_{k,A}$**.**

Equation (D.2) forces

$$g'(y_A^\star) = K_{jk} := \frac{\kappa_{k,A} - \kappa_{j,A}}{\eta_{j,A} - \eta_{k,A}}. \tag{D.3}$$

Case 2: $\eta_{j,A} = \eta_{k,A}$**.**

Then (D.2) 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 (x_j, x_k) with $x_j, x_k > 0$, we can transfer all mass to a single LLM (say, set $\tilde{x}_j = x_j + x_k$ and $\tilde{x}_k = 0$) without changing the objective value:

$$F(\tilde{x}, z) = F(x, z).$$

Thus, even when two such equivalent LLMs are active, there exists an equally optimal solution that is concentrated.

Conclusion from KKT: Unless $g'(y_A^\star)$ equals one of the finitely many values $\{K_{jk}\}_{\eta_{j,A} \neq \eta_{k,A}}$ (in which case two distinct LLMs can be active), at most one coordinate can be active. In the degenerate case where $\kappa_{j,A} = \kappa_{k,A}$ and $\eta_{j,A} = \eta_{k,A}$, multiple LLMs may be active, but we can always find an equally optimal concentrated solution by moving all mass to one LLM.

We now show that for large $S_A(\alpha)$, $g'(y_A^\star)$ avoids all K_{jk} (except possibly in knife-edge cases), so generically there is at most one active coordinate. Note that

$$y_A^\star = \xi_A \sum_{j=1}^{M} \eta_{j,A} \cdot x_j^\star \cdot z_j^\star \geq \eta_{\min} \cdot S_A(\alpha),$$

where $\eta_{\min} := \max_{j,y} \eta_{j,y} < \infty$. Therefore, as $\alpha \to \infty$, we have $y_A^\star \to \infty$. 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 $g'(x) = c\rho \cdot x^{\rho-1}$. Then, as x tends to infinity, we have $g'(x)$ tends to infinity, and hence avoid all K_{jk}'s. Thus, as $\alpha \to 0$, we have $y_A^* \to \infty$ and hence $g'(y_A^*)$ 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 $s := x_j + x_k > 0$ and define the reduced one-dimensional objective

$$\phi_s(\theta) := \xi_A \kappa_{j,A} \theta s + \xi_A \kappa_{k,A} (1-\theta) s + c \cdot \Big(u + \xi_A \eta_{j,A} \theta s + \xi_A \eta_{k,A} (1-\theta) s \Big), \qquad \theta \in [0,1],$$

where $u := \sum_{\ell \notin \{j,k\}} x_\ell / \eta_\ell$. Since $\phi_s(\theta)$ is linear in s, the minimum of ϕ_s lies at an endpoint and the allocation can be concentrated on a single coordinate.

The trivial linear tie $\kappa_{j,A} = \kappa_{k,A}$ and $\eta_{j,A} = \eta_{k,A}$ gives $\phi_s'(\theta) \equiv 0$, so every θ is optimal; choosing an endpoint again yields a concentrated minimizer.

Conclusion: Combining Subcases (a) and (b), for all sufficiently large $S_A(\alpha)$ every minimizer has at most one positive coordinate; in the tie $\eta_{j,A} = \eta_{k,A}, \eta_{j,A} = \eta_{k,A}$ an endpoint solution is equally optimal. Since $\sum_j x_j = S_A(\alpha) > 0$, exactly one coordinate equals $S_A(\alpha)$; i.e., every minimizer is concentrated. □