Epistemic Throughput: Fundamental Limits of Attention-Constrained Inference

Lei You[1,2]

[1]Technical University of Denmark, [2]CSPaper @ Scholar7

Recent generative and tool-using AI systems can surface a large volume of candidates at low marginal cost, yet only a small fraction can be checked carefully. This creates a decoder-side bottleneck: downstream decision-makers must form reliable posteriors from many public records under scarce attention. We formalize this regime via Attention-Constrained Inference (ACI), in which a cheap screening stage processes K records and an expensive verification stage can follow up on at most B of them. Under Bayes log-loss, we study the maximum achievable reduction in posterior uncertainty per window, which we call *epistemic throughput*. Our main result is a "JaKoB" scaling law showing that epistemic throughput has a baseline term that grows linearly with verification and prevalence, and an additional *information-leverage* term that scales as $\sqrt{JKB}$, where J summarizes screening quality. Thus, expanding cheap screening can nonlinearly amplify scarce verification, even when informative records are rare. We further show that this scaling is tight in a weak-screening limit, and that in the sparse-verification regime ($B \ll K$), substantial leverage requires heavy-tailed score distributions; for light-tailed scores the amplification is only logarithmic.

Correspondence: leiyo@dtu.dk, lei@scholar7.com
Date: February 2026

1 Introduction

1.1 Background and key problem

Generative AI has fundamentally inverted the economics of information: the cost of generating plausible content has dropped to zero [1], while the cost of verifying its truth remains distinctively high. This creates a crisis in *high-stakes* domains, such as scientific discovery, software supply chains, and financial intelligence, where the system's objective is not engagement but *inference*. Unlike casual content consumption, these domains tolerate near-zero error; their goal is to reconstruct a binary latent state (truth, safety, or validity) from a noisy observable artifact. Historically, trust in these systems relied on *costly signaling* [2], where generation cost served as a proxy for quality. LLMs have shattered this separating equilibrium, flooding the channel with cheap artifacts that mimic authority but lack the underlying guarantee of truth.

This disruption is catastrophic because verification does not scale. There is no single "editor-in-chief" (or audit pipeline) with enough bandwidth to vet the global stream. In practice, verification is carried out by downstream decision-makers. Sometimes there is only one, but often there are many, and each operates under scarce attention budgets. We therefore face an *attention-constrained inference* (ACI) problem. Here, "attention" means a budget of costly inspection and verification steps, not the self-attention mechanism in Transformers. Our focus is the maximum achievable information gain under log-loss per window, which we call *epistemic throughput*. To cope with informational overload, agents naturally adopt a **two-stage inference strategy**, consistent with rational inattention [3] and visual search [4], to maximize information gain about the latent ground truth under attention constraints:

1. **Screening (K)** (broad attention): inspect K records using a low-cost, high-throughput heuristic (e.g., scanning an abstract). Screening is scalable but inherently noisy.
2. **Verification (B)** (deep attention): select a subset of B candidates from the screened pool for high-fidelity auditing (e.g., reproducing an experiment). Verification is accurate but scarce.

This abstraction matches modern retrieval augmented generation (RAG) workflows. Screening corresponds to retrieval and lightweight scoring over a large pool, while verification corresponds to consuming and integrating a small set of sources well enough to justify a low-uncertainty output.

The central operational challenge therefore shifts from channel capacity to *attention constraints*: how to use a noisy screen to aggressively filter the haystack so that the scarce verification budget is allocated only to the most promising signals.

1.2 Main contributions

JaKoB scaling law. We characterize the best achievable *information gain* under log-loss in the ACI haystack regime. For a signal with prevalence p and screening quality J, we prove a fundamental scaling law (formalized as a converse in Theorem 6 and shown achievable in a weak-screening

limit in Theorem 10):

$$\text{Gain} \;\approx\; I_{\text{ver}}\, B\left(p + c\sqrt{\tfrac{JK}{B}}\right), \qquad c \le \sqrt{\tfrac{\ln 2}{2}}. \tag{1}$$

Here, I_{ver} is the information contributed per verified informative record (defined precisely in Section 4). This factorization can be read as the verification capacity $I_{\text{ver}}B$ scaled by an *enhanced hit rate*. The square-root term reveals a tight and surprisingly universal mechanism: even weak screening (J small) can nonlinearly amplify a scarce verification budget, provided that screening can oversample at rate K/B.

- *Base precision (p):* the prevalence (sparsity) of informative records; this is the hit rate under random verification.
- *Screening boost* ($c\sqrt{JK/B}$): the precision gain enabled by oversampling. By screening a large volume (K) to fill a small verification budget (B), the system can "cherry-pick" and dramatically enrich the verified subset.

Fig. 1 offers a geometric view of the JaKoB tradeoff. The gray *Random* line is the gain from verifying blindly (hit rate p), and the gray *Oracle* line is the ceiling under perfect screening (hit rate 1). The colored curves show the JaKoB prediction for different values of JK, interpolating between these extremes. The background shading visualizes the yield per verified item.

Constructive achievability. Crucially, we show that the JaKoB limit is not merely a bound: it is efficiently achievable. A simple *score-based verification* policy (inspect K records, rank them by screening scores, and verify the top-B) achieves the scaling in a practical *weak-screening regime* (modeled via a local logistic regression).

Escaping the Gaussian trap. In the sparse verification regime ($B \ll K$), the benefit of massive screening is controlled by the *upper-tail behavior* of the screening scores. For light-tailed scores (e.g., Gaussian), leverage grows only logarithmically; for heavy-tailed scores (e.g., Pareto), leverage can be polynomial (Propositions 12–13). To sharpen intuition, Appendix A also gives a fully solved benchmark with an exact finite-block characterization of the risk-resource boundary.

1.3 Paper organization

Section 2 discusses related work. Section 3 presents the ACI model and the haystack specialization. Section 4 states the main converse, achievability, and scaling results. Section 5 concludes the paper. Appendix B contains proofs, and Appendix A provides the fully solved benchmark.

2 Related Work

Our model sits at the intersection of information theory, decision theory, and resource-constrained inference. Classical information theory focuses on *encoder-side* constraints,

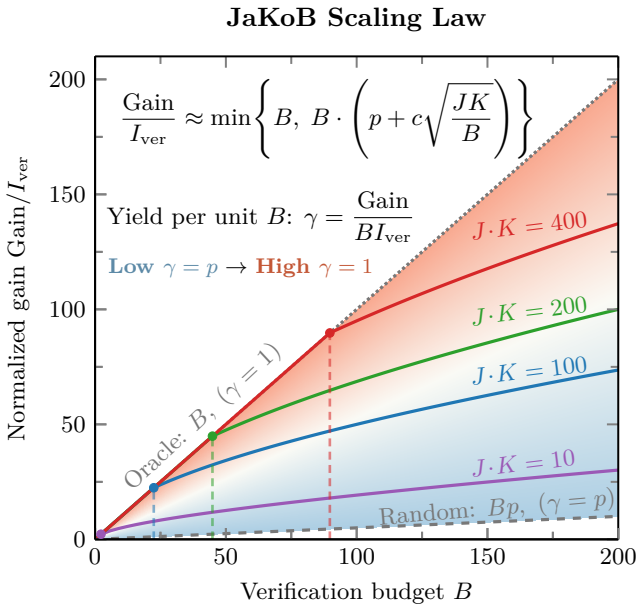

Figure 1: **JaKoB scaling and information leverage.** Blind verification yields the baseline Bp, while perfect screening caps at the oracle ceiling B. With screening quality J applied to K candidates, screening contributes an additional $\Theta(\sqrt{JKB})$ gain (clipped at B); larger JK shifts the curve upward, so the same information gain can be reached with a smaller scarce budget B and a higher per-verification yield $\gamma = \text{Gain}/(BI_{\text{ver}})$.

including rate-limited descriptions and noisy channels, as in [5] and standard texts [6]. ACI instead isolates a complementary regime in which *records are abundant and public*, but *downstream decision-makers are attention-limited*: only a small subset can be inspected and an even smaller subset can be verified. Below we organize the most relevant threads and clarify where ACI differs.

2.1 Decoder-side constraints and information acquisition

A large literature studies how resource constraints at the receiver affect communication and inference. In information theory, one line constrains *decoding resources* such as reliability, memory, or energy, and asks how these constraints modify fundamental limits; see, e.g., the thesis of [7] and related work on faulty or resource-limited iterative decoding. A second, closely related line treats *information acquisition* as an *action* under a cost constraint: the system chooses whether/how to obtain side information. This is formalized in "source coding with actions" and "vending machine" models, where actions control the quality or availability of side information [8]. ACI shares the acquisition viewpoint but differs operationally: our two-stage structure separates *cheap screening* (inspection) from *expensive high-fidelity observation* (verification), and the key coupling parameter is the screening quality (quantified by the mutual information between the screening signal and the latent truth), which governs how much screening can enrich the verified

set.

Economics provides a complementary and influential formalization of limited attention via *rational inattention*, where information-processing costs are measured through Shannon mutual information [3]. This perspective has been developed into tractable choice models [9, 10] and surveyed in [11]. Our model is aligned with the mutual-information cost philosophy, but with an engineering emphasis: we ask how a community-level inspection budget K can amplify a verification budget B in terms of Bayes log-loss.

Finally, our results relate to classic *sequential design* and *active learning* themes (e.g., choosing what to measure next) dating back to [12]. ACI is deliberately *one-shot* at the window level: screening produces a batch of scores and verification selects a subset (often top-B), making the analysis closer to selection under budget than to fully adaptive querying.

2.2 Decentralized inference and social learning

Distributed detection and estimation traditionally study networks of sensors that send messages to a fusion center or cooperate under communication constraints. Foundational results include decentralized detection with many sensors [13] and the detection/data-fusion synthesis in [14]. Modern surveys emphasize resource constraints and network structure in distributed inference [15]. These models typically assume a *designed sensing architecture* and a *communication protocol*. ACI takes a different stance: records already exist as public artifacts, and the bottleneck lies in what decoders can afford to *inspect* and *verify*.

Although ACI does not impose a multi-agent coordination constraint and applies even in the single-decoder case, its motivation of many downstream agents reusing public artifacts connects to social learning and belief aggregation. Early models of consensus and belief pooling include [16]. In economics, herd behavior and informational cascades [17, 18] show how public actions can dominate private information. More recent networked-learning work characterizes when agents learn from neighbors without a central coordinator [19]. ACI is compatible with these perspectives but focuses on a different bottleneck: we model the *production of verifiable public artifacts* under shared budgets, and evaluate the best achievable population-level log-loss.

2.3 Log-loss as an operational metric

Log-loss is central in information theory because it turns Bayes risk into conditional entropy. Beyond its role in coding under logarithmic loss [20], log-loss is the canonical scoring rule for probabilistic prediction and is deeply tied to universal coding and Bayesian mixture methods [21–23]. Recent work continues to study prediction under log-loss with side information and regret characterizations [24]. In ACI, log-loss provides a clean risk metric that naturally aligns with the objective of maximizing information gain, making the inspection and verification interaction analytically transparent.

2.4 Haystacks, rare signals, and extreme-value selection

The haystack regime, where there are many candidates but only a small fraction of items are informative, is shared with statistics on sparse or rare signal detection and multiple testing. Work on detecting sparse mixtures and rare effects highlights sharp phase transitions and the need to aggregate weak evidence [25, 26]. Verification as "testing a few" among many hypotheses also relates to multiple-comparisons ideas such as false discovery rate control [27]. In settings where tests can be pooled, group testing [28] provides another classic abstraction of scarce testing resources. ACI differs in that verification reveals high-fidelity information *only when the record is informative*, so the central object becomes *enrichment* of the verified set from cheap screening.

A complementary, information-theoretic view of selection appears in adaptive data analysis. Russo and Zou [29] bound the bias induced by data-dependent exploration by the mutual information between the selected analysis and the data, including explicit treatments of filtering and rank selection. Although their objective is to control post-selection bias rather than maximize information gain, the underlying information-usage principle is closely related to our enrichment converse: limited screening information constrains how far a top-B rule can tilt the verified set away from the baseline prevalence.

Technically, the top-B selection step makes order statistics and extreme-value theory unavoidable [30, 31]. Our tail-leverage results formalize exactly when expanding inspection K is valuable: it is valuable only to the extent that the score distribution has exploitable upper-tail mass.

2.5 Verification at scale in modern information ecosystems

Finally, ACI is motivated by empirical systems where verification is scarce relative to content generation. In NLP and fact checking, large-scale claim verification datasets and benchmarks (e.g., FEVER) formalize the pipeline of retrieving candidates and verifying a small set [32]. In the foundation-model era, community reports and analyses emphasize the gap between cheap content generation and expensive auditing [33–35]. ACI is not a proposal for a new verifier; rather, it provides *limits* that constrain *any* such pipeline under inspection/verification budgets.

A common mitigation is to ground generation in external evidence through RAG [36] and related retrieval-based pipelines. Retrieval can cheaply surface many plausible sources, but only a few can be read, cross-checked, and trusted within a limited context window or compute budget. The bottleneck remains verification attention.

The same bottleneck appears in tool-using agents that can run many low-cost queries or heuristics but can only follow up deeply on a small set of results. Representative instances include web-browsing QA agents such as WebGPT [37],

ReAct-style agents that interleave reasoning with tool calls [38], Toolformer-style self-supervised tool use [39] and modular tool-routing architectures such as MRKL systems [40]. A related prompting approach is self-ask, which can be paired with a search engine to answer follow-up questions [41]. Another relevant application is CSPaper Review that selects the most justified critiques from a pool of concurrent agent-generated reviews [42].

3 System Model

This section introduces our one-window model for attention-constrained inference (ACI) and the performance metric under log-loss. A population of producers releases a large collection of records, while a population of decoders has limited capacity to inspect them. Later in the section we specialize to a canonical *haystack* regime that captures a two-stage pipeline, cheap screening followed by expensive verification, which is the setting analyzed in Section 4. All random variables are defined on a common probability space.

Fig. 2 previews the haystack specialization. In one window, the system can screen many candidates at low cost and then follow up deeply on only a small subset. We formalize both the general model and this specialization below.

3.1 Producers, records, and decoders

Let Θ be a latent state taking values in a finite set $\mathcal{T}$, with prior P_Θ. There are m producers indexed by $i \in [m]$. Producer i observes evidence $E_i \in \mathcal{Y}$ generated according to a conditional distribution $P_{E_i|\Theta}$, and publishes a public record

$$X_i = f_i(E_i) \in \mathcal{X}.$$

We write $X^m = (X_1, \ldots, X_m)$ for the entire collection of records released in a window.

There are N decoders indexed by $j \in [N]$. We allow multiple decoders to model community settings where verified artifacts are public and can be reused. At the same time, our fundamental limits depend only on aggregate budgets, so the results apply equally to the single-decoder case $N = 1$. Decoder j starts with a prior belief π_j on Θ (capturing side information that may differ across decoders) and forms an estimate after inspecting only a small subset of the records.

3.2 Attention constraints and routing

Decoder j has an attention budget $k_j \in \mathbb{Z}_{\geq 0}$. In one window it inspects a subset

$$S_j \subseteq [m], \qquad |S_j| \leq k_j,$$

and observes the corresponding inspected information set $I_j = \{X_i : i \in S_j\}$. We define the total attention capacity

$$K := \sum_{j=1}^{N} k_j.$$

Many systems expose cheap public metadata for every record (timestamps, user features, embeddings, coarse heuristic scores, and so on). We model this as a public signal $U_i \in \mathcal{U}$ that is visible to all decoders before they choose what to inspect. A routing rule maps the collection of public signals into inspected sets. Formally, decoder j selects

$$S_j = r_j(U^m, \pi_j, \omega_j),$$

through routing r_j, with ω_j being private randomness. We assume that all private randomness variables are independent of the record-generation process, including Θ and all producer-side observations.

3.3 Log-loss risk

After observing I_j and starting from prior π_j, decoder j outputs a distribution q_j on $\mathcal{T}$. Under log-loss, the incurred loss is $-\log q_j(\Theta)$. We evaluate population-level performance by the average Bayes risk

$$\text{Risk} := \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}\left[-\log q_j(\Theta)\right]. \tag{2}$$

When q_j matches the true Bayesian posterior, the risk attains its theoretical lower bound, $\mathrm{H}(\Theta \mid I_j, \pi_j)$. This identity establishes that *the fundamental limit of inference is governed by the residual uncertainty* in the data, thereby justifying the use of Information Gain (entropy reduction) as the canonical metric for epistemic throughput.

3.4 Haystack specialization: screening and verification

The main results in Section 4 focus on a canonical haystack regime in which screening is cheap and verification is expensive. The key modeling ingredient is that only a small fraction of inspected records can become informative about Θ upon verification. We capture this sparsity using a latent type indicator.

Assumption 1 (Informative records and screening)**.** *Each inspected record has a latent type $T \in \{0, 1\}$ with $\mathbb{P}(T = 1) = p$, where $p \in (0, 1)$. An inspection reveals a screening statistic $Z \in \mathcal{Z}$. The pair (T, Z) is independent of Θ. We measure screening quality by the mutual information $J := \mathrm{I}(T; Z)$.*

Assumption 1 says that screening does not directly reveal information about the target Θ. Instead, it provides a noisy proxy for whether a record is worth verifying. Policies may map the screening statistic to a score $\eta(Z)$ and use it to decide which records to verify (for example, the Bayes score $\eta(z) = \mathbb{P}(T = 1 \mid Z = z)$ used in our achievability analysis later).

Verification is the expensive channel that can reduce uncertainty about Θ. In the canonical model we analyze, verification reveals the type together with an additional observation. Revealing T is a modeling simplification that isolates the role of verification as a distinct side channel.

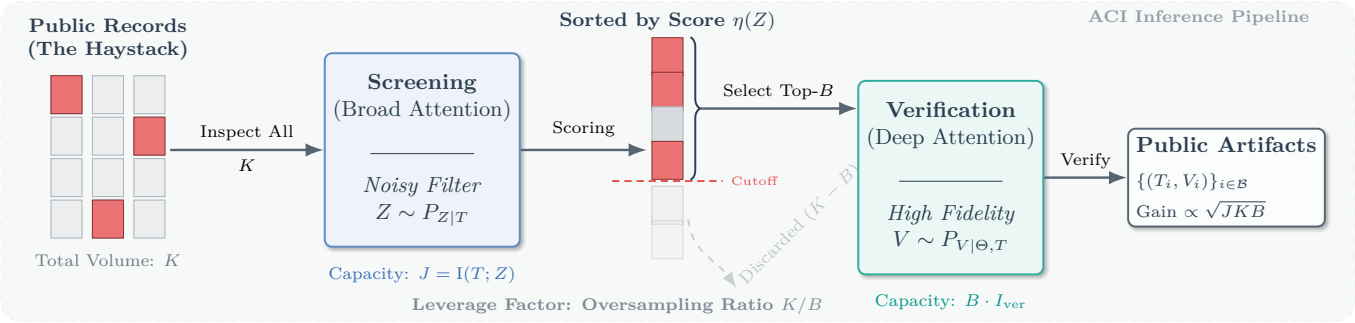

Figure 2: **The ACI Inference Pipeline.** The system operates as a two-stage information refinery in the haystack regime ($B \ll K$). A massive volume of public records (K) is first filtered by a cheap, noisy screening channel (Z) to prioritize attention. Based on the screening scores $\eta(Z)$, only the most promising top-B candidates receive expensive, high-fidelity verification (V). The *oversampling ratio* K/B acts as a leverage factor, amplifying the yield of the scarce verification budget to achieve the $\Theta(\sqrt{JKB})$ scaling.

Assumption 2 (Verification channel)**.** *If a record is verified, the decoder observes a pair (T, V). Conditioned on (Θ, T), the variable V is independent of Z. If $T = 0$, then V is independent of Θ. If $T = 1$, then V carries information about Θ. We define*

$$I_{\mathrm{ver}} := \mathrm{I}(\Theta; V \mid T = 1). \tag{3}$$

A policy in the haystack regime screens K candidate records in a window and verifies at most B of them. We index the screened candidates by $i \in [K]$ and write $\mathcal{B} \subseteq [K]$ for the (random) verified index set, with $|\mathcal{B}| \le B$. The verified outputs $\{(T_i, V_i)\}_{i\in\mathcal{B}}$ are published as public artifacts and can be reused by all decoders.

We include two specializations of the latent state Θ to cover different coupling structures across records. In the global-Θ specialization, all K records share a single target Θ; conditioned on Θ, the tuples (T_i, Z_i, V_i) are i.i.d. across i, but they are generally correlated after marginalizing over Θ. In the per-record-claim specialization, $\Theta = (\Theta_1, \ldots, \Theta_K)$ collects independent claim variables; conditioned on $(\Theta_1, \ldots, \Theta_K)$ the tuples are independent across i (not necessarily identically distributed given Θ), while marginally the tuples are i.i.d. across i. In both specializations, conditioned on (Θ, T_i) in the global model and on (Θ_i, T_i) in the decoupled model, the verification output V_i is independent of all screening statistics and all other verification outputs.

Scope of results. Our converse results (notably Lemma 4 and Theorem 6) apply to both specializations, as they depend only on the screening/verification channels, budgets (K, B), and information parameters (J, I_{ver}). Our achievability and tightness results are proved under the decoupled-claim specialization (Assumption 9), where the verification stage reduces to an optimal selection problem without global coupling.

We interpret (K, B) as community-level budgets. This is a natural abstraction when verified artifacts can be published as public objects that all decoders can use.

Assumption 3 (Public artifacts and shared access)**.** *Within one window, every verification output is published as a public artifact. All decoders can access the set of verified artifacts. Therefore, for fundamental-limit analysis under budgets (K, B), we can evaluate performance using a hypothetical agent that observes all verification outputs generated in the window. This does not introduce a fusion center into the system. It is only an evaluation device that matches what any decoder can reconstruct from public artifacts.*

4 Main Theoretical Results

This section studies the haystack specialization introduced in Section 3. We focus on the non-adversarial setting. Our goal is to quantify how inspection and verification interact under Bayes log-loss.

We use the following budgets. The community inspects K independent records and verifies at most B among those inspected records. We write $\alpha := B/K$. All logarithms are base 2. We use ln only for natural logarithms in log-odds expressions. Entropy and mutual information are measured in bits.

4.1 A fundamental tradeoff between inspection and verification

Our first question is: given budgets (K, B) and screening quality $J = \mathrm{I}(T; Z)$, how much can inspection amplify a scarce verification budget under log-loss? Our main answer is Theorem 6, which gives a universal converse bound on the information gain. The key technical ingredient is a selection enrichment bound (Lemma 4), which limits how much *any* selection rule based on Z can increase the hit rate of informative records.

Lemma 4 (Selection enrichment bound)**.** *Assume Assumption 1. Let $S \in \{0, 1\}$ be any (possibly randomized) selection rule that depends on Z and on auxiliary randomness*

independent of (T, Z)*, and let* $\alpha := \mathbb{P}(S = 1)$*. Then*

$$\mathbb{P}(T = 1 \mid S = 1) \;\le\; p + \sqrt{\frac{\ln 2}{2\alpha}\, J}. \tag{4}$$

Lemma 4 makes the role of J explicit. It states that a cheap screening score cannot create an arbitrarily clean verified set. The upper bound depends on α because enrichment is harder when verification is extremely sparse.

We now bound the information gain under budgets (K, B).

Definition 5 (Budgets (K, B))**.** *A policy uses budgets* (K, B) *if it inspects* K *i.i.d. records, and it verifies at most* B *among the inspected records. The policy may use the screening statistics* $Z_1, \ldots, Z_K$ *to select which records to verify.*

Assumption 3 lets us evaluate the community through the public artifacts created in one window. For $i \in [K]$, define the verification transcript

$$Y_i := \begin{cases} \perp, & \text{if record } i \text{ is not verified,} \\ (T_i, V_i), & \text{if record } i \text{ is verified,} \end{cases}$$

where $\perp$ is a fixed symbol. We write $\boldsymbol{Z} := (Z_1, \ldots, Z_K)$ and $\boldsymbol{Y} := (Y_1, \ldots, Y_K)$. We define $D(K, B)$ as the Bayes-optimal expected log-loss after observing $(\boldsymbol{Z}, \boldsymbol{Y})$ under budgets (K, B).

Theorem 6 (A tradeoff between verification and attention under log-loss)**.** *Assume Assumptions 1 and 2. Consider any policy that uses budgets* (K, B) *in the sense of Definition 5. Then*

$$\mathrm{H}(\Theta) - D(K, B) \;\le\; B \cdot I_{\mathrm{ver}}\left(p + \sqrt{\frac{\ln 2}{2}}\sqrt{\frac{JK}{B}}\right). \tag{5}$$

Theorem 6 is the main converse bound. The linear term Bp is the baseline gain of random verification. The square-root term quantifies the best possible amplification from inspection.

The next corollary rewrites the converse as a lower bound on the verification budget required for a target gain.

Corollary 7 (Verification required for a target gain)**.** *Assume the setting of Theorem 6. Fix any target information gain* $\Delta \in (0, \mathrm{H}(\Theta))$*. If a policy satisfies* $\mathrm{H}(\Theta) - D(K, B) \ge \Delta$*, then* B *must satisfy*

$$B \;\ge\; \frac{1}{4p^2}\left(\sqrt{\frac{\ln 2}{2}\, JK + \frac{4p\Delta}{I_{\mathrm{ver}}}} - \sqrt{\frac{\ln 2}{2}\, JK}\right)^2. \tag{6}$$

Corollary 7 separates two regimes. If $J = 0$, screening is useless and verification must scale as $\Omega(1/p)$ in the sparse limit. If JK is large, inspection can reduce the required verification budget.

4.2 Achievability with score-based verification

Theorem 6 exhibits a square-root *amplification* term of order $\sqrt{JK/B}$. A natural question is whether this term is achievable, or whether it is an artifact of the converse proof. We answer in the affirmative: in a local weak-screening model (Assumption 8), the simple top-B score-based verification policy achieves the same scaling up to constants (Theorem 10).

Let

$$\eta(z) := \mathbb{P}(T = 1 \mid Z = z)$$

denote the Bayes score of the screening statistic. We study a local regime in which $\eta(Z)$ is close to p.

Assumption 8 (Weak screening through a log-odds score)**.** *Let* $\eta(Z) = \mathbb{P}(T = 1 \mid Z)$*. There exist a scalar* $\varepsilon \ge 0$ *and a real-valued random variable* G *such that*

$$\ln \frac{\eta(Z)}{1 - \eta(Z)} = \ln \frac{p}{1 - p} + \varepsilon G \tag{7}$$

almost surely. We assume $\mathbb{E}[G] = 0$*,* $\mathbb{E}[G^2] = 1$*, and* $\mathbb{E}[|G|^3] < \infty$*. We also assume that* G *has a continuous distribution.*

Assumption 8 follows a standard Bayesian view of evidence accumulation rather than an ad hoc modeling choice. On the log-odds scale, Bayes' rule is additive: the posterior log-odds equal the prior log-odds plus a weight-of-evidence term (a log-likelihood ratio), a perspective emphasized by Good [43] in cryptanalysis (with evidence measured in "bans"). Assumption 8 adopts a local parametrization by writing this log-odds increment as εG. The qualifier *weak* refers to the regime $\varepsilon \to 0$, where $\eta(Z) = p + O(\varepsilon)$ and the screening information $J = \mathrm{I}(T; Z)$ vanishes accordingly (typically $J = O(\varepsilon^2)$). We normalize G to have mean zero and unit variance so that ε captures the overall screening strength; the remaining moment and continuity conditions are regularity assumptions used to control the asymptotics. From a modern machine-learning standpoint, equation 7 is the canonical logit-link form, closely related to logistic regression with a linear predictor [44], where G is a standardized score and ε sets its signal-to-noise level.

Fix $\alpha \in (0, 1)$. Let q_α be the $(1 - \alpha)$-quantile of G, namely

$$q_\alpha := \inf\{q \in \mathbb{R} : \mathbb{P}(G \le q) \ge 1 - \alpha\}.$$

We define the upper-tail mean

$$m_G(\alpha) := \frac{1}{\alpha}\,\mathbb{E}[G\,\mathbf{1}\{G \ge q_\alpha\}]\,. \tag{8}$$

Assumption 9 (Decoupled claims for achievability)**.** *We use the following decoupled claim model in Theorem 10. The latent state is* $\Theta = (\Theta_1, \ldots, \Theta_K)$*, where* $\Theta_1, \ldots, \Theta_K$ *are i.i.d. with prior* P_Θ*. Moreover,* Θ *is independent of* (T^K, Z^K)*. Conditioned on* (Θ_i, T_i)*, the verification output* V_i *is generated by the verification channel in Assumption 2 with* Θ *replaced by* Θ_i*. The collection* $(V_i)_{i \in [K]}$ *is conditionally independent across* i *given* (Θ, T^K)*.*

Theorem 10 (Score-based verification achieves a square–root gain)**.** *Assume Assumptions 1, 2, 8, and 9. Fix $\alpha \in (0,1)$ and set $B = \lfloor \alpha K \rfloor$. We use the following policy. We inspect K i.i.d. records and observe $Z_1, \ldots, Z_K$. We compute the scores $\eta(Z_i)$. We then verify the B records with the largest scores. Then, in the joint limit $\varepsilon \to 0$ and $K \to \infty$ with fixed α,*

$$\mathrm{H}(\Theta) - D(K,B) \;\geq\; \min\Big\{\mathrm{H}(\Theta),\; I_{\mathrm{ver}}\big(Bp + c_G(p,\alpha)\sqrt{J\,K\,B}\big)\Big\} \;+\; o\big(\sqrt{J\,K\,B}\big) \quad (9)$$

where $J = \mathrm{I}(T;Z)$ and

$$c_G(p,\alpha) := \sqrt{2\ln 2\, p(1-p)\,\alpha}\; m_G(\alpha). \quad (10)$$

Theorem 10 addresses the achievability gap. It shows that a simple score-based policy attains the same square-root scaling as the converse. The constant $m_G(\alpha)$ captures how much the policy can leverage the upper tail.

Corollary 11 (Tight square-root scaling in a weak-screening regime)**.** *Under the assumptions of Theorem 10, let $D^\star(K,B)$ be the minimum Bayes log-loss over all policies that use budgets (K,B). Then, as $\varepsilon \to 0$ and $K \to \infty$ with fixed $\alpha \in (0,1)$ and $B = \lfloor \alpha K \rfloor$,*

$$\mathrm{H}(\Theta) - D^\star(K,B) \geq I_{\mathrm{ver}}\big(Bp + c_G(p,\alpha)\sqrt{JKB}\big) + o\big(\sqrt{JKB}\big), \quad (11)$$

$$\mathrm{H}(\Theta) - D^\star(K,B) \leq I_{\mathrm{ver}}\Big(Bp + \sqrt{\tfrac{\ln 2}{2}\,JKB}\Big). \quad (12)$$

Corollary 11 gives a nontrivial inner and outer bound pair. Both bounds have the same $\sqrt{JKB}$ scaling.

4.3 The haystack regime and tail leverage

In many discovery systems, verification is much more expensive than inspection. This corresponds to $B \ll K$, namely $\alpha = B/K \to 0$. We call this the haystack regime. In this regime, the effect of screening enters through the tail mean $m_G(\alpha)$.

Using equation 10 and the relation $\alpha = B/K$, we rewrite the leading gain term in equation 9 as

$$c_G(p,\alpha)\sqrt{JKB} \;=\; \sqrt{2\ln 2\, p(1-p)}\; m_G(\alpha)\sqrt{J\,B}. \quad (13)$$

Equation equation 13 shows that expanding attention K is valuable only if $m_G(B/K)$ grows.

We now record two canonical examples. They show that the tail shape of the score determines the leverage of attention.

Proposition 12 (Gaussian scores yield logarithmic tail leverage)**.** *Assume that $G \sim \mathcal{N}(0,1)$. Then, as $\alpha \to 0$,*

$$m_G(\alpha) = \sqrt{2\log\frac{1}{\alpha}}\,\big(1 + o(1)\big). \quad (14)$$

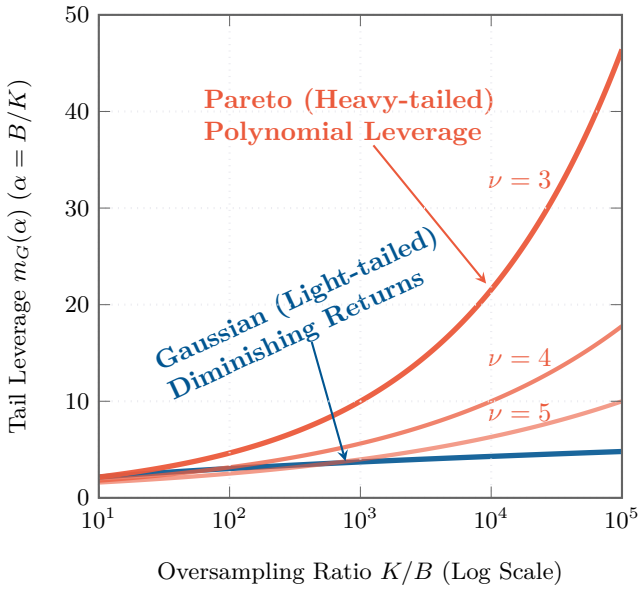

Figure 3: **Escaping the Gaussian trap: Tail leverage determines screening utility.** The figure compares the asymptotic gain from massive screening ($K \gg B$) under light-tailed (Gaussian) versus heavy-tailed (Pareto with varying ν) score distributions. While Gaussian scores yield *diminishing returns* (scaling logarithmically as $\sqrt{\ln K}$), heavy-tailed scores provide *polynomial leverage* (scaling as $K^{1/\nu}$). This illustrates the critical dichotomy in the haystack regime: expanding the screening budget K is highly effective only when the score distribution admits exploitable extremes.

Proposition 12 implies that, for light-tailed scores, expanding K yields diminishing returns. The leverage grows only as $\sqrt{\log(K/B)}$.

Proposition 13 (A Pareto right tail yields polynomial tail leverage)**.** *Let X have a Pareto distribution with exponent $\nu > 3$, namely $\mathbb{P}(X \geq x) = x^{-\nu}$ for $x \geq 1$. Define the centered and standardized score*

$$G := \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}.$$

Then, as $\alpha \to 0$,

$$m_G(\alpha) = \frac{\nu}{\nu - 1} \cdot \frac{1}{\sqrt{\mathrm{Var}(X)}}\,\alpha^{-1/\nu}\,\big(1 + o(1)\big). \quad (15)$$

Proposition 13 implies a very different scaling. For a heavy-tailed score in the Fréchet domain, attention yields polynomial leverage. This gives a precise criterion for when massive screening is most effective.

Fig. 3 visualizes these two regimes by plotting the asymptotic tail leverage $m_G(\alpha)$ (with $\alpha = B/K$) as a function of the oversampling ratio K/B.

The key message is that pushing K/B ever higher only pays off when the score distribution has exploitable extremes: for

Gaussian-like tails the benefit grows only logarithmically, whereas for Pareto-like tails it grows as a power law.

Remark 14 (A fully solved benchmark)**.** *Appendix A gives a canonical model in which the risk-resource region can be characterized exactly. The boundary is expressed through an upper-tail functional of the score distribution. This benchmark complements the global-*Θ *analysis in this section.*

5 Discussion and Conclusion

The post-generative information ecosystem has a new asymmetry. Producing and reshaping claims is cheap, while verifying them remains expensive. Modern pipelines can surface thousands of candidates through retrieval, heuristics, or lightweight model judgments, yet only a small subset can be investigated deeply with tools, experts, or reading. The binding constraint is often not access to text or the ability to phrase an answer, but the ability to allocate attention to verification that turns plausible statements into warranted belief.

This paper formalizes that constraint through ACI. In each window, a decision-maker can screen many records at low cost and verify at most B of them at high cost after observing a screening statistic. Under Bayes log-loss, we measure performance by the maximum achievable reduction in posterior uncertainty per window, which we call *epistemic throughput*. The main results characterize this throughput information-theoretically. The JaKoB scaling law isolates a simple interaction between screening quality J, screening volume K, and verification capacity B: besides the linear baseline from verification, there is a leverage term of order $\sqrt{JKB}$ (up to universal constants and the trivial ceiling $\mathrm{H}(\Theta)$). In sparse-verification regimes, leverage is governed by extreme values, which is why heavy-tailed score behavior can deliver polynomial amplification while light-tailed scores yield only logarithmic gains.

5.1 Positioning: from reliability and semantics to truthfulness

It is useful to compare ACI to established communication paradigms, because the distinction is about what is being optimized.

Traditional communication theory focuses on *transmission reliability*. The core problem is to move symbols through a physical channel with limited bandwidth and noise. Bit rate and error probability are natural metrics because the message content is taken as given, and the challenge is preserving it through the medium.

Cognitive communication emphasizes *resource adaptability*. The problem is to operate efficiently in a dynamic environment by sensing and exploiting time-varying resources, such as spectrum opportunities, under uncertainty and competition.

Semantic communication shifts attention to *meaning consistency*. When many symbol strings can express the same intent, the objective becomes task success or semantic similarity rather than exact symbol recovery. The question is whether the receiver can act on what the sender meant.

The bottleneck motivating ACI is *orthogonal* to all the three above. A claim can be transmitted perfectly in a resource-optimized way, interpreted exactly as intended, and still be false about the world. After large models became widely deployed, information streams contain an abundance of fluent, coherent statements whose truthfulness varies widely. When verification is scarce, the limiting factor is the ability to concentrate deep checks on the few items that can most improve the posterior. This is the regime in which epistemic throughput, rather than bit rate or semantic similarity, becomes the natural target.

5.2 Toward a paradigm "epistemic communication"

These considerations motivate a possible next paradigm, which we call *epistemic communication*. If this is a useful discipline, it should be defined by a distinct objective, a distinct metric, and a corresponding set of design questions.

5.2.1 Objective and metric.

Epistemic communication would focus on *truthfulness* as a system-level goal: maximize the rate at which a receiver acquires verified, decision-relevant truths under attention constraints. Epistemic throughput offers a concrete metric because it measures posterior improvement under an operational loss (log-loss) and makes the verification budget explicit. Unlike a semantic metric, it distinguishes between statements that are merely consistent in meaning and statements that are supported by evidence strong enough to change a rational posterior.

5.2.2 System boundary.

Epistemic communication is inherently end-to-end. It spans how claims are produced and packaged, how weak signals are extracted cheaply, how verification actions are chosen, and how verification outputs are represented and shared as public artifacts. ACI is a minimal model that makes these components explicit. Screening corresponds to low-cost feature extraction and ranking; verification corresponds to expensive evidence acquisition; artifacts correspond to reusable proof objects, such as citations, provenance records, or standardized evaluation reports.

5.2.3 Design principles suggested by ACI.

The limit theorems in this paper suggest several principles that are likely to persist beyond the stylized assumptions used for analysis.

First, screening creates value only insofar as it increases *selectivity*. In ACI, screening does not directly reveal Θ; it shapes which records get verified. In sparse-verification settings, average improvements in screening are often less important than the behavior of the top of the score distribution, because only the top few candidates are ever checked. This is exactly what the tail results capture: leverage is driven by rare, very high-scoring candidates.

Second, oversampling is a lever, but it has failure modes. The square-root term shows that increasing cheap screening volume can amplify scarce verification, but the amplification saturates quickly when scores are light-tailed. This yields a practical diagnostic: if doubling the screened pool does not materially improve the best few candidates, additional screening compute is unlikely to buy much epistemic progress. Conversely, engineering for heavy-tailed score behavior provides a concrete target: scoring pipelines should be designed to occasionally produce candidates that are far more verifiable and informative than typical ones.

Third, public artifacts are the medium of epistemic communication. Verification effort scales only if its outputs are cheap to reuse. This makes representation a first-class problem. Artifacts should be easy to consume, hard to counterfeit, and composable. Provenance metadata, standardized citation formats, cryptographic attestations, and structured argument summaries all fit naturally into this role. In ACI terms, better artifact design increases the downstream value of each verification action by enabling reuse across many decoders.

Finally, truthfulness is a systems property shaped by incentives and adversaries. In realistic ecosystems, the record stream is not passive. Producers may be rewarded for attention rather than accuracy, and adversaries can craft records whose screening features mimic truth. Epistemic communication therefore connects information theory to learning, mechanism design, and security: it asks how to align incentives and build auditability so that producing verifiable truth is the easiest way to succeed.

5.3 Implications for retrieval-augmented and tool-using systems

Retrieval-augmented generation and tool-use pipelines naturally fit the ACI template. Retrieval, query rewriting, and lightweight heuristics can surface many candidate snippets or hypotheses cheaply, while deep follow-up is limited by expensive tool calls, constrained context windows, human review, or time. The theory suggests evaluating screening stages by their contribution to *verification yield*, not only by average ranking quality. In particular, the heavy-tail requirement points to an engineering goal: screening should sometimes produce "clean hits" that are easy to verify and highly informative once verified, rather than merely shifting average ranks.

The model also highlights the value of reusing verification work. Caching verified sources, publishing structured citations, and storing intermediate tool traces can turn a one-off verification action into a reusable artifact. This does not increase the raw verification budget B, but it increases the epistemic return per verification by reducing repeated effort across users and across time.

5.4 Limitations and open directions

Our analysis is intentionally stylized, and its limitations outline a research agenda. First, the present model is window-based and largely i.i.d.; real systems are sequential and reflexive. Screening and verification decisions affect future data collection, user behavior, and even the content stream itself. Extending epistemic throughput to sequential settings would connect directly to active learning, bandits, and adaptive experimentation, where verification actions are chosen to maximize long-run posterior improvement.

Second, we focus on log-loss because it links optimal risk to conditional entropy and yields clean converse bounds. Other objectives, such as decision regret, calibration under abstention, or constrained false discovery, may induce different throughput notions and different limits. Third, verification costs are heterogeneous in practice and evidence sources are correlated. A richer model would allow multi-level verification ladders with varying costs and dependencies, which would better match tool chains that mix cheap checks with a few expensive audits.

Finally, the heavy-tail phenomenon raises a practical question: what architectural choices create heavy-tailed score distributions that are both useful and robust? Answering this would bridge the theory to concrete designs for ranking functions, tool orchestration, and evidence aggregation in the presence of strategic manipulation.

5.5 Closing perspective

When content is abundant and attention is scarce, the binding constraint is the rate at which systems can convert weak, cheap signals into strong, verified updates. ACI provides a language for this regime and a scaling law that quantifies its best-case behavior. More broadly, it motivates epistemic communication as a principled way to study truthfulness at scale, with epistemic throughput serving as a system-level metric for progress.

A A Fully Solved Benchmark: A Tight Risk-Resource Region

Section 4 derives general converse bounds under budgets (K, B). It also gives an achievability result for a decoupled claim model. In this appendix we complement that analysis with a benchmark that admits an exact characterization. The benchmark is relevant when each informative record carries its own independent latent claim. This decoupling removes global coupling and turns the problem into an optimal selection task.

A.1 A canonical haystack model

We consider a large pool of candidate records. Each record has an unobserved binary type $T \in \{0, 1\}$, where $T = 1$ means the record is informative and $T = 0$ means the record is uninformative. We write $p := \mathbb{P}(T = 1)$.

Each informative record is associated with an independent latent claim $\Theta \in \mathcal{T}$ with prior P_Θ. Different informative records have independent latent claims, and all latent claims share the same prior.

An inspection reveals a screening statistic $Z \in \mathcal{Z}$. The

screening channel is specified by $P_{Z|T=1}$ and $P_{Z|T=0}$. After observing $Z = z$, the posterior probability that the record is informative is

$$\eta(z) := \mathbb{P}(T = 1 \mid Z = z). \tag{16}$$

We call $\eta(Z)$ the score. The score is a sufficient statistic for selecting which inspected records to verify.

A verification reveals a pair (T, V), where $V \in \mathcal{V}$ is a verification output. We assume the following verification model.

Assumption 15 (Verification as a side channel)**.** *If $T = 0$, then V is a fixed symbol v_0 and is independent of Θ. If $T = 1$, then V is generated according to a channel $P_{V|\Theta}$.*

We will return to this benchmark at the end of the appendix with a finite-length simulation that validates the resulting boundary and its weak-screening approximation (Fig. 4).

Under log-loss, the Bayes risk for an informative verified record equals $\mathrm{H}(\Theta \mid V)$. For an uninformative verified record it equals $\mathrm{H}(\Theta)$. It is convenient to define the information contribution of verification as

$$I_{\mathrm{ver}}^{\mathrm{bm}} := \mathrm{I}(\Theta; V), \tag{17}$$

where the mutual information is computed under $P_\Theta P_{V|\Theta}$. In this benchmark, $I_{\mathrm{ver}}^{\mathrm{bm}}$ coincides with I_{ver} from equation 3.

In one window, we inspect K records and then verify exactly B of the inspected records, where $1 \le B \le K$. The inspection outcomes are i.i.d. samples $(Z_1, \ldots, Z_K)$ drawn from the mixture distribution induced by T and $P_{Z|T}$.

A verification policy maps the scores $\eta_1, \ldots, \eta_K$ to a verification set $\mathcal{B} \subseteq [K]$ with $|\mathcal{B}| = B$. We allow randomized policies.

A.2 Benchmark risk-resource region

We evaluate performance by the average log-loss over the B verified records. Let $q_i(\cdot)$ denote the predicted distribution for the latent claim of verified record i. The benchmark risk of a policy is

$$\bar{D}(K, B) := \frac{1}{B} \sum_{i \in \mathcal{B}} \mathbb{E}\left[-\log q_i(\Theta_i)\right], \tag{18}$$

where Θ_i denotes the latent claim associated with record i when $T_i = 1$.

We say that a triple (K, B, D) is achievable if there exists a policy using at most K inspections and B verifications such that $\bar{D}(K, B) \le D$. The benchmark risk-resource region is the set of all achievable triples.

The next theorem characterizes the optimal expected number of informative verifications.

Theorem 16 (Optimal hit rate under score-based selection)**.** *Let $\eta_i := \eta(Z_i)$ be the score of inspected record i, and let $\eta_{(1)} \ge \eta_{(2)} \ge \cdots \ge \eta_{(K)}$ denote the order statistics. For any verification policy,*

$$\mathbb{E}\left[\sum_{i \in \mathcal{B}} T_i\right] \le \mathbb{E}\left[\sum_{\ell=1}^{B} \eta_{(\ell)}\right]. \tag{19}$$

Moreover, the upper bound is achieved by verifying the B records with the largest scores.

Theorem 16 reduces the verification stage to an optimal top-B selection problem once we condition on inspection outcomes. The boundary of the region therefore depends on a single functional of the score distribution.

Theorem 17 (Tight tradeoff between risk and resource)**.** *Under Assumption 15, the minimum achievable risk under (K, B) equals*

$$\bar{D}^\star(K, B) = \mathrm{H}(\Theta) - \frac{I_{\mathrm{ver}}^{\mathrm{bm}}}{B}\, \mathbb{E}\left[\sum_{\ell=1}^{B} \eta_{(\ell)}\right]. \tag{20}$$

Equivalently, a triple (K, B, D) is achievable if and only if $D \ge \bar{D}^\star(K, B)$.

Theorem 17 gives a complete benchmark characterization. It also makes clear why upper-tail properties matter. The boundary is expressed through the expected sum of the top-B posterior scores among the K inspected records.

A.3 An asymptotic single-letter expression

Equation equation 20 is exact for any finite (K, B). In the large-system regime the boundary admits a single-letter form.

Let F_η be the cumulative distribution function of the score $\eta(Z)$, and let $Q_\eta(u) := \inf\{x : F_\eta(x) \ge u\}$ be the quantile function. Fix a verification fraction $\alpha \in (0, 1)$ and set $B = \lfloor \alpha K \rfloor$.

Corollary 18 (Single-letter boundary for $K \to \infty$)**.** *Assume $\eta(Z)$ has a continuous distribution. As $K \to \infty$ with $B = \lfloor \alpha K \rfloor$,*

$$\frac{1}{K}\, \mathbb{E}\left[\sum_{\ell=1}^{B} \eta_{(\ell)}\right] \to \int_{1-\alpha}^{1} Q_\eta(u)\, du. \tag{21}$$

Consequently,

$$\bar{D}^\star(K, \lfloor \alpha K \rfloor) = \mathrm{H}(\Theta) - \frac{I_{\mathrm{ver}}^{\mathrm{bm}}}{\alpha} \int_{1-\alpha}^{1} Q_\eta(u)\, du + o(1). \tag{22}$$

Corollary 18 turns the benchmark region into a single-letter expression. The integral in equation 22 is an upper-tail mean of the score distribution. It is a canonical tail-leverage quantity.

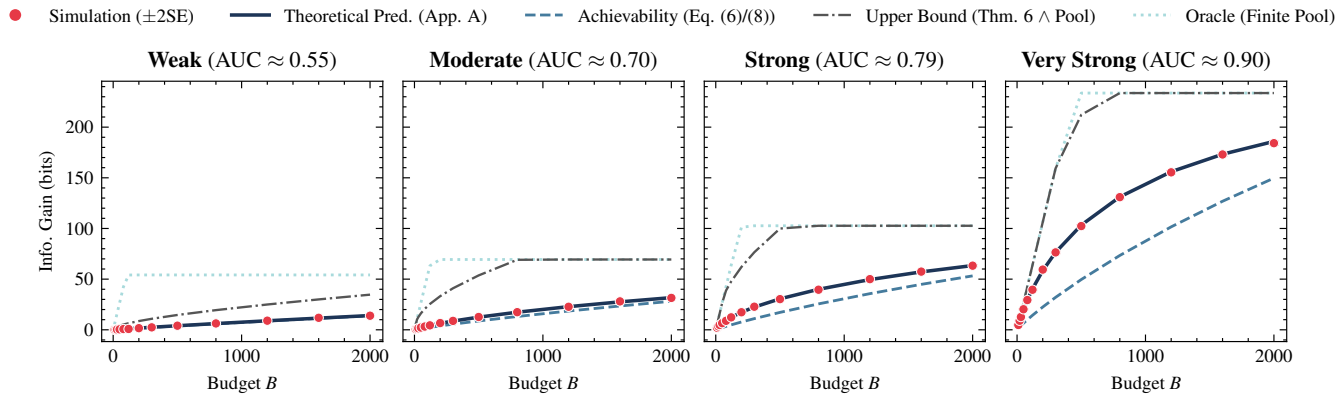

Figure 4: **Finite-length validation of the JaKoB scaling law.** We simulate screening scores from a logistic model and apply the top-B verification policy with $K = 10^4$, $\delta = 0.1$, and $p_0 = 0.01$. Markers (with ±2SE error bars) report the empirical information gain (log-loss reduction, in bits). The solid curve is the benchmark prediction from Theorem 17 (Eq. equation 20); the dashed curve is the weak-screening approximation from Theorem 10/Corollary 11. Dash-dot curves show the converse upper bound from Theorem 6 intersected with the finite-pool constraint, and the dotted curve indicates the finite-pool oracle. Across four screening strengths (AUC $\approx 0.55, 0.70, 0.79, 0.90$), the empirical gains closely track the benchmark and remain below the theoretical and oracle ceilings, while the weak-screening law becomes conservative as screening strengthens. The code to reproduce this plot is available at https://github.com/youlei202/Attention-Constrained-Inference.

A.4 Finite-length validation (simulation)

To validate that the benchmark characterization remains predictive away from the asymptotic regime, we simulate a finite pool of $K = 10^4$ inspected records under a logistic score model consistent with Assumption 8. Concretely, we generate posterior scores of the form $\eta_i = \sigma(\mathrm{logit}(p_0) + \delta G_i)$, where σ is the logistic function, $p_0 = 0.01$ is the base rate, $\delta = 0.1$ controls screening strength, and G_i is standardized. We then apply the top-B verification policy and report the empirical information gain $\mathrm{H}(\Theta) - D(K, B)$.

Fig. 4 compares these empirical gains to three theoretical references: (i) the benchmark prediction from Theorem 17 via equation 20 (solid), (ii) the weak-screening approximation implied by Theorem 10 / Corollary 11 (dashed), and (iii) the converse bound from Theorem 6 intersected with the finite-pool ceiling (dash-dot). The dotted curve is a finite-pool oracle that knows which records are informative and can therefore verify only true positives up to the pool limit.

B Proofs for Section 4

We provide proofs of the main results. All logarithms are base 2.

B.1 Proof of Lemma 4

Proof. Let S be any randomized function of Z and let $\alpha = \mathbb{P}(S = 1)$. By data processing, $\mathrm{I}(T; S) \le \mathrm{I}(T; Z) = J$. We expand $\mathrm{I}(T; S)$ using conditional relative entropy:

$$\mathrm{I}(T; S) = \alpha\, \mathrm{KL}\big(P_{T|S=1} \,\|\, P_T\big) + (1 - \alpha)\, \mathrm{KL}\big(P_{T|S=0} \,\|\, P_T\big).$$

Both KL terms are nonnegative, hence

$$\mathrm{KL}\big(P_{T|S=1} \,\|\, P_T\big) \le \frac{1}{\alpha}\, \mathrm{I}(T; S) \le \frac{J}{\alpha}.$$

We now relate KL divergence to total variation. Pinsker's inequality for base-2 KL divergence states that

$$\mathrm{TV}\big(P_{T|S=1}, P_T\big) \le \sqrt{\frac{\ln 2}{2}\, \mathrm{KL}\big(P_{T|S=1} \,\|\, P_T\big)}.$$

Since T is binary, total variation equals the absolute difference of the probability of $T = 1$:

$$\mathrm{TV}\big(P_{T|S=1}, P_T\big) = \big|\mathbb{P}(T = 1 \mid S = 1) - \mathbb{P}(T = 1)\big|.$$

Combining the last three displays yields

$$\mathbb{P}(T = 1 \mid S = 1) \le p + \sqrt{\frac{\ln 2}{2\alpha}\, J}.$$

This is equation 4. □

B.2 Proof of Theorem 6

Proof. We consider K inspected records indexed by $i \in [K]$. For record i, let (T_i, Z_i, V_i) be distributed as in Assumptions 1 and 2. Let $S_i \in \{0, 1\}$ indicate whether record i is verified. The policy may select the verified set using the entire vector $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$ and private randomness independent of (T^K, Z^K). The budget constraint $\sum_{i=1}^{K} S_i \le B$ holds almost surely.

For each record i, define the verification transcript

$$Y_i := \begin{cases} \perp, & S_i = 0, \\ (T_i, V_i), & S_i = 1, \end{cases}$$

where $\perp$ is a fixed symbol, and let $\boldsymbol{Y} := (Y_1, \ldots, Y_K)$. Under Bayes-optimal prediction, the expected log-loss equals $D(K, B) = \mathrm{H}(\Theta \mid \boldsymbol{Z}, \boldsymbol{Y})$. Therefore,

$$\mathrm{H}(\Theta) - D(K, B) = \mathrm{I}(\Theta; \boldsymbol{Z}, \boldsymbol{Y}).$$

Assumption 1 implies that (T^K, Z^K) is independent of Θ, hence $\mathrm{I}(\Theta; \boldsymbol{Z}) = 0$ and

$$\mathrm{H}(\Theta) - D(K, B) = \mathrm{I}(\Theta; \boldsymbol{Y} \mid \boldsymbol{Z}).$$

We upper bound $\mathrm{I}(\Theta; \boldsymbol{Y} \mid \boldsymbol{Z})$ by a chain-rule argument that keeps the global latent state. By the chain rule,

$$\mathrm{I}(\Theta; \boldsymbol{Y} \mid \boldsymbol{Z}) = \sum_{i=1}^{K} \mathrm{I}\big(\Theta; Y_i \mid \boldsymbol{Z}, Y^{i-1}\big).$$

If $S_i = 0$ then Y_i is constant and the term is zero. If $S_i = 1$ then $Y_i = (T_i, V_i)$. Since T_i is independent of Θ, we have

$$\mathrm{I}\big(\Theta; Y_i \mid \boldsymbol{Z}, Y^{i-1}, S_i = 1\big) = \mathrm{I}\big(\Theta; V_i \mid T_i, \boldsymbol{Z}, Y^{i-1}, S_i = 1\big).$$

Let $W_i := (\boldsymbol{Z}, Y^{i-1}, S_i = 1)$. By the record-generation model in Section 3 together with Assumption 2, we have the conditional independence $V_i \perp\!\!\!\perp W_i \mid (\Theta, T_i)$. Using the identity $\mathrm{I}(\Theta; V_i \mid T_i, W_i) = \mathrm{H}(V_i \mid T_i, W_i) - \mathrm{H}(V_i \mid \Theta, T_i)$, we obtain

$$\mathrm{I}(\Theta; V_i \mid T_i, W_i) \le \mathrm{H}(V_i \mid T_i) - \mathrm{H}(V_i \mid \Theta, T_i) = \mathrm{I}(\Theta; V_i \mid T_i).$$

Therefore,

$$\mathrm{I}(\Theta; \boldsymbol{Y} \mid \boldsymbol{Z}) \le \sum_{i=1}^{K} \mathbb{E}\left[S_i \, \mathrm{I}(\Theta; V_i \mid T_i)\right].$$

If $T_i = 0$ then V_i is independent of Θ and the mutual information term is 0. If $T_i = 1$ the term equals I_{ver} by equation 3. Thus

$$\mathrm{I}(\Theta; \boldsymbol{Y} \mid \boldsymbol{Z}) \le I_{\mathrm{ver}} \sum_{i=1}^{K} \mathbb{E}\left[S_i \, \mathbf{1}\{T_i = 1\}\right].$$

We now bound the expected number of informative verified records. Let I be a random index uniform on $[K]$ and independent of all other variables. Define $(T, Z) := (T_I, Z_I)$ and $S := S_I$. Then

$$\alpha := \mathbb{P}(S = 1) = \mathbb{E}[S] = \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}[S_i] \le \frac{B}{K}.$$

By the definition of the random index I, we have

$$\mathbb{E}\left[S \, \mathbf{1}\{T = 1\}\right] = \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}\left[S_i \, \mathbf{1}\{T_i = 1\}\right].$$

By the definition of conditional probability,

$$\begin{aligned}\mathbb{E}\left[S \, \mathbf{1}\{T = 1\}\right] &= \mathbb{P}(S = 1)\mathbb{P}(T = 1 \mid S = 1) \\ &= \alpha \mathbb{P}(T = 1 \mid S = 1).\end{aligned}$$

Combining yields

$$\sum_{i=1}^{K} \mathbb{E}\left[S_i \, \mathbf{1}\{T_i = 1\}\right] = K\alpha \, \mathbb{P}(T = 1 \mid S = 1).$$

We bound $\mathbb{P}(T = 1 \mid S = 1)$ in terms of $J = \mathrm{I}(T; Z)$. Since $\boldsymbol{Z}_{-I}$ is independent of (T, Z), the Markov chain

$$T \; - \; Z \; - \; (S, \boldsymbol{Z}_{-I})$$

holds. By data processing, $\mathrm{I}(T; S) \le \mathrm{I}(T; Z) = J$. Lemma 4 applies and yields

$$\mathbb{P}(T = 1 \mid S = 1) \le p + \sqrt{\frac{\ln 2}{2\alpha} \, J}.$$

Combining gives

$$\mathrm{I}(\Theta; \boldsymbol{Y} \mid \boldsymbol{Z}) \le I_{\mathrm{ver}} \, K\alpha \left(p + \sqrt{\frac{\ln 2}{2\alpha} \, J} \right).$$

Since $K\alpha \le B$, we have $K\alpha p \le Bp$. For the square-root term we use $K\alpha\sqrt{1/\alpha} = K\sqrt{\alpha} \le \sqrt{KB}$. This yields

$$\begin{aligned}\mathrm{H}(\Theta) - D(K, B) &\le I_{\mathrm{ver}} \left(Bp + \sqrt{\frac{\ln 2}{2} \, J \, BK} \right) \\ &= B \cdot I_{\mathrm{ver}} \left(p + \sqrt{\frac{\ln 2}{2}} \sqrt{\frac{JK}{B}} \right).\end{aligned}$$

This is equation 5. □

B.3 Proof of Corollary 7

Proof. Assume that a policy satisfies $\mathrm{H}(\Theta) - D(K, B) \ge \Delta$. If $I_{\mathrm{ver}} = 0$, then Theorem 6 implies $\mathrm{H}(\Theta) - D(K, B) = 0$ for every policy. Since $\Delta > 0$, the premise cannot hold and the corollary is vacuous. We therefore assume $I_{\mathrm{ver}} > 0$ in the remainder of the proof. By Theorem 6,

$$\Delta \le I_{\mathrm{ver}} \left(Bp + \sqrt{\frac{\ln 2}{2} \, J \, K \, B} \right).$$

We divide both sides by I_{ver} and define

$$g := \frac{\Delta}{I_{\mathrm{ver}}} \qquad \text{and} \qquad a := \sqrt{\frac{\ln 2}{2} \, J \, K}.$$

The inequality becomes $pB + a\sqrt{B} \ge g$. Let $x := \sqrt{B}$. We rewrite the constraint as a quadratic inequality

$$px^2 + ax - g \ge 0.$$

Since $p > 0$, the polynomial is convex in x and has a unique nonnegative root. Therefore

$$x \ge \frac{-a + \sqrt{a^2 + 4pg}}{2p}.$$

Squaring both sides yields

$$B \ge \frac{1}{4p^2} \left(\sqrt{a^2 + 4pg} - a \right)^2.$$

Substituting the definitions of a and g gives equation 6. □

Lemma 19 (Averages of top-score samples)**.** *Let $G_1, G_2, \dots$ be i.i.d. copies of a random variable G such that $\mathbb{E}[G^2] < \infty$. Assume that G has a continuous distribution. Fix $\alpha \in (0,1)$ and set $B = \lfloor \alpha K \rfloor$. Let $G_{(1)} \geq \cdots \geq G_{(K)}$ denote the order statistics of $(G_1, \dots, G_K)$. Then, as $K \to \infty$,*

$$\frac{1}{B}\sum_{i=1}^{B} G_{(i)} \to m_G(\alpha) \qquad \textit{in probability,}$$

where $m_G(\alpha)$ is defined in equation 8. Moreover, the convergence also holds in L^1.

B.4 Proof of Lemma 19

Proof. We first prove convergence in probability. Fix $\alpha \in (0,1)$ and set $B = \lfloor \alpha K \rfloor$. Let q_α be a $(1-\alpha)$-quantile of G as in equation 8.

Fix $\delta > 0$. Since q_α is a quantile, we have $\mathbb{P}(G \geq q_\alpha + \delta) < \alpha$ and $\mathbb{P}(G > q_\alpha - \delta) > \alpha$. Define the counting variables

$$N_+ := \sum_{i=1}^{K} \mathbf{1}\{G_i \geq q_\alpha + \delta\}, \qquad N_- := \sum_{i=1}^{K} \mathbf{1}\{G_i > q_\alpha - \delta\}.$$

By the strong law of large numbers, $N_+/K \to \mathbb{P}(G \geq q_\alpha + \delta)$ and $N_-/K \to \mathbb{P}(G > q_\alpha - \delta)$ almost surely. Therefore, the event $\{N_+ < B < N_-\}$ holds with probability tending to 1.

On this event, every sample with $G_i \geq q_\alpha + \delta$ must appear among the top B values, and every top-B sample must satisfy $G_i > q_\alpha - \delta$. This yields the sandwich bound

$$\begin{aligned}\frac{1}{B}\sum_{i=1}^{K} G_i \,\mathbf{1}\{G_i \geq q_\alpha + \delta\} &\leq \frac{1}{B}\sum_{i=1}^{B} G_{(i)} \\ &\leq \frac{1}{B}\sum_{i=1}^{K} G_i \,\mathbf{1}\{G_i > q_\alpha - \delta\},\end{aligned}$$

with probability tending to 1.

Since $\mathbb{E}[G^2] < \infty$, the random variables $G\,\mathbf{1}\{G \geq q_\alpha + \delta\}$ and $G\,\mathbf{1}\{G > q_\alpha - \delta\}$ are integrable. The strong law gives

$$\frac{1}{K}\sum_{i=1}^{K} G_i \,\mathbf{1}\{G_i \geq q_\alpha + \delta\} \to \mathbb{E}\left[G\,\mathbf{1}\{G \geq q_\alpha + \delta\}\right] \qquad \text{a.s.}$$

and the same for the upper bracket. Since $K/B \to 1/\alpha$, we obtain almost sure limits for the two normalized sums by multiplying by K/B.

We now let $\delta \downarrow 0$. As $\delta \downarrow 0$, we have the pointwise convergences

$$\begin{aligned}\mathbf{1}\{G \geq q_\alpha + \delta\} &\to \mathbf{1}\{G > q_\alpha\}, \\ \mathbf{1}\{G > q_\alpha - \delta\} &\to \mathbf{1}\{G \geq q_\alpha\}.\end{aligned}$$

Since $\mathbb{E}[|G|] < \infty$ and the integrands are dominated by $|G|$, dominated convergence yields

$$\begin{aligned}\mathbb{E}\left[G\,\mathbf{1}\{G \geq q_\alpha + \delta\}\right] &\to \mathbb{E}\left[G\,\mathbf{1}\{G > q_\alpha\}\right], \\ \mathbb{E}\left[G\,\mathbf{1}\{G > q_\alpha - \delta\}\right] &\to \mathbb{E}\left[G\,\mathbf{1}\{G \geq q_\alpha\}\right].\end{aligned}$$

Because G has a continuous distribution, $\mathbb{P}(G = q_\alpha) = 0$, hence $\mathbb{E}\left[G\,\mathbf{1}\{G > q_\alpha\}\right] = \mathbb{E}\left[G\,\mathbf{1}\{G \geq q_\alpha\}\right]$. Therefore, both limits coincide and equal $\mathbb{E}\left[G\,\mathbf{1}\{G \geq q_\alpha\}\right]$. This implies that the sandwich bounds converge to $m_G(\alpha)$. Hence $\frac{1}{B}\sum_{i=1}^{B} G_{(i)} \to m_G(\alpha)$ in probability.

We next show convergence in L^1. By Cauchy–Schwarz,

$$\left|\frac{1}{B}\sum_{i=1}^{B} G_{(i)}\right| \leq \sqrt{\frac{1}{B}\sum_{i=1}^{B} G_{(i)}^2}.$$

Since the top-B samples are a subset of all K samples, we have $\sum_{i=1}^{B} G_{(i)}^2 \leq \sum_{i=1}^{K} G_i^2$. Taking expectations yields

$$\mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B} G_{(i)}^2\right] \leq \frac{K}{B}\,\mathbb{E}[G^2] = \frac{1}{\alpha}\,\mathbb{E}[G^2].$$

Hence $\{\frac{1}{B}\sum_{i=1}^{B} G_{(i)}\}_{K \geq 1}$ is uniformly integrable. Convergence in probability together with uniform integrability implies L^1 convergence. □

B.5 Proof of Theorem 10

Proof. We consider K i.i.d. records indexed by $i \in [K]$. For record i, let (T_i, Z_i) satisfy Assumption 1. If record i is verified, the decoder observes (T_i, V_i) as in Assumption 2. We write $\eta_i := \mathbb{P}(T_i = 1 \mid Z_i)$. By Assumption 8,

$$\ln \frac{\eta_i}{1-\eta_i} = \ln \frac{p}{1-p} + \varepsilon G_i,$$

where $G_1, \dots, G_K$ are i.i.d. copies of G. The map $g \mapsto \ln \frac{\eta}{1-\eta}$ is strictly increasing in $\eta \in (0,1)$. Therefore η_i is a strictly increasing function of G_i. The policy verifies the $B = \lfloor \alpha K \rfloor$ largest scores η_i, which is equivalent to verifying the B largest values among $G_1, \dots, G_K$. We denote the verified index set by $\mathcal{B}$.

Under Bayes-optimal prediction, observing the full transcript is equivalent to observing $(Z^K, \mathcal{B}, T_\mathcal{B}, V_\mathcal{B})$, where $Z^K := (Z_1, \dots, Z_K)$ and $\mathcal{B} \subseteq [K]$ is the verified index set. Therefore,

$$D(K,B) = \mathrm{H}\big(\Theta \mid Z^K, \mathcal{B}, T_\mathcal{B}, V_\mathcal{B}\big)$$

and

$$\mathrm{H}(\Theta) - D(K,B) = \mathrm{I}\big(\Theta; Z^K, \mathcal{B}, T_\mathcal{B}, V_\mathcal{B}\big).$$

Assumption 1 implies that (T^K, Z^K) is independent of Θ. For any fixed $\varepsilon > 0$, since G has a continuous distribution, the scores are distinct with probability one. Therefore the top-B verification set $\mathcal{B}$ is a deterministic function of Z^K. Hence $(Z^K, \mathcal{B}, T_\mathcal{B})$ is independent of Θ, and

$$\mathrm{H}(\Theta) - D(K,B) = \mathrm{I}\big(\Theta; V_\mathcal{B} \mid Z^K, \mathcal{B}, T_\mathcal{B}\big).$$

Under Assumptions 2 and 9, conditioned on (Θ, T^K), the verification outputs are generated without using Z^K. Since Θ is also independent of Z^K, this implies that, conditioned

on $(\mathcal{B}, T_{\mathcal{B}})$, the pair $(\Theta, V_{\mathcal{B}})$ is independent of Z^K. Therefore

$$\mathrm{I}\big(\Theta; V_{\mathcal{B}} \mid Z^K, \mathcal{B}, T_{\mathcal{B}}\big) = \mathrm{I}\big(\Theta; V_{\mathcal{B}} \mid \mathcal{B}, T_{\mathcal{B}}\big).$$

We now expand $\mathrm{I}(\Theta; V_{\mathcal{B}} \mid \mathcal{B}, T_{\mathcal{B}})$. By Assumption 9, conditioned on $(\mathcal{B}, T_{\mathcal{B}})$, the variables V_i are independent across $i \in \mathcal{B}$. Moreover, V_i depends on Θ only through Θ_i. Therefore

$$\begin{aligned}\mathrm{I}\big(\Theta; V_{\mathcal{B}} \mid \mathcal{B}, T_{\mathcal{B}}\big) &= \mathrm{H}(V_{\mathcal{B}} \mid \mathcal{B}, T_{\mathcal{B}}) - \mathrm{H}(V_{\mathcal{B}} \mid \Theta, \mathcal{B}, T_{\mathcal{B}}) \\ &= \sum_{i \in \mathcal{B}} \mathrm{H}(V_i \mid T_i) - \sum_{i \in \mathcal{B}} \mathrm{H}(V_i \mid \Theta_i, T_i) \\ &= \sum_{i \in \mathcal{B}} \mathrm{I}\big(\Theta_i; V_i \mid T_i\big). \end{aligned} \tag{23}$$

If $T_i = 0$ the term is 0. If $T_i = 1$ the term equals I_{ver} by equation 3. Thus

$$\mathrm{H}(\Theta) - D(K, B) = I_{\mathrm{ver}} \cdot \mathbb{E}\left[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\right]. \tag{24}$$

We also have the trivial upper bound $\mathrm{H}(\Theta) - D(K, B) \le \mathrm{H}(\Theta)$. Combining yields

$$\mathrm{H}(\Theta) - D(K, B) \ge \min\left\{\mathrm{H}(\Theta),\ I_{\mathrm{ver}} \cdot \mathbb{E}\left[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\right]\right\}. \tag{25}$$

We now lower bound $\mathbb{E}\big[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\big]$. By the tower property,

$$\mathbb{E}[\mathbf{1}\{i \in \mathcal{B}\} T_i] = \mathbb{E}\big[\mathbb{E}[T_i \mid Z_i]\, \mathbf{1}\{i \in \mathcal{B}\}\big] = \mathbb{E}[\eta_i\, \mathbf{1}\{i \in \mathcal{B}\}].$$

Summing over i gives

$$\mathbb{E}\Big[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\Big] = \sum_{i=1}^{K} \mathbb{E}[\eta_i\, \mathbf{1}\{i \in \mathcal{B}\}].$$

We next relate η_i to G_i. Let $\sigma(x) := (1 + e^{-x})^{-1}$ be the logistic function and let $s := \ln \frac{p}{1-p}$. By Assumption 8, we have $\eta_i = \sigma(s + \varepsilon G_i)$. Since $\sigma(s) = p$ and $\sigma'(s) = p(1-p)$, Taylor's theorem gives

$$\eta_i = p + \varepsilon p(1-p) G_i + R_i,$$

where the remainder satisfies $|R_i| \le c_0 \varepsilon^2 G_i^2$ for a constant c_0 that depends only on p.

Using the tower property and $\mathbb{E}[T_i \mid Z_i] = \eta_i$, we have

$$\mathbb{E}\Big[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\Big] = \sum_{i=1}^{K} \mathbb{E}[\eta_i\, \mathbf{1}\{i \in \mathcal{B}\}].$$

Substituting the expansion of η_i yields

$$\begin{aligned}\mathbb{E}\Big[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\Big] &= Bp \\ &+ \varepsilon p(1-p) \sum_{i=1}^{K} \mathbb{E}[G_i\, \mathbf{1}\{i \in \mathcal{B}\}] + \sum_{i=1}^{K} \mathbb{E}[R_i\, \mathbf{1}\{i \in \mathcal{B}\}].\end{aligned}$$

Let $G_{(1)} \ge \cdots \ge G_{(K)}$ denote the order statistics. Since $\mathcal{B}$ selects the top B scores, $\sum_{i=1}^{K} G_i\, \mathbf{1}\{i \in \mathcal{B}\} = \sum_{i=1}^{B} G_{(i)}$. Lemma 19 implies

$$\frac{1}{B} \sum_{i=1}^{B} G_{(i)} \to m_G(\alpha) \qquad \text{in probability as } K \to \infty.$$

Thus

$$\sum_{i=1}^{K} \mathbb{E}[G_i\, \mathbf{1}\{i \in \mathcal{B}\}] = B\, m_G(\alpha) + o(B), \qquad \text{as } K \to \infty.$$

We now bound the remainder term. Since $\sum_{i \in \mathcal{B}} G_i^2 \le \sum_{i=1}^{K} G_i^2$, we have

$$\sum_{i=1}^{K} \mathbb{E}\left[G_i^2\, \mathbf{1}\{i \in \mathcal{B}\}\right] \le K \mathbb{E}[G^2] = K.$$

Using $|R_i| \le c_0 \varepsilon^2 G_i^2$ yields

$$\left|\sum_{i=1}^{K} \mathbb{E}[R_i\, \mathbf{1}\{i \in \mathcal{B}\}]\right| \le c_0 \varepsilon^2 \sum_{i=1}^{K} \mathbb{E}\left[G_i^2\, \mathbf{1}\{i \in \mathcal{B}\}\right] \le c_0 \varepsilon^2 K.$$

Since $B = \lfloor \alpha K \rfloor$ with fixed $\alpha \in (0, 1)$ and $\varepsilon \to 0$, the bound $c_0 \varepsilon^2 K$ is $o(\varepsilon B)$. We obtain

$$\mathbb{E}\Big[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\Big] = Bp + \varepsilon p(1-p)\, B\, m_G(\alpha) + o(\varepsilon B). \tag{26}$$

We now relate ε to $J = \mathrm{I}(T; Z)$. For binary T,

$$J = \mathrm{I}(T; Z) = \mathbb{E}\left[\mathrm{KL}\big(\mathrm{Bern}(\eta(Z)) \,\|\, \mathrm{Bern}(p)\big)\right].$$

Let $\delta := \eta(Z) - p$. Under Assumption 8, Taylor's theorem for the logistic function gives $\delta = \varepsilon p(1-p) G + O(\varepsilon^2 G^2)$. The binary relative entropy admits the expansion

$$\mathrm{KL}\big(\mathrm{Bern}(p + x) \,\|\, \mathrm{Bern}(p)\big) = \frac{x^2}{2 \ln 2\, p(1-p)} + o(x^2). \tag{27}$$

Since $\mathbb{E}[|G|^3] < \infty$, we can take expectations in equation 27 with $x = \delta$. This yields

$$\begin{aligned} J &= \frac{\mathbb{E}[\delta^2]}{2 \ln 2\, p(1-p)} + o(\varepsilon^2) \\ &= \frac{p(1-p)\, \varepsilon^2}{2 \ln 2} + o(\varepsilon^2), \qquad \varepsilon \to 0. \end{aligned}$$

Therefore

$$\varepsilon p(1-p) = \sqrt{2 \ln 2\, p(1-p)\, J} + o(\sqrt{J}).$$

We substitute into equation 26. Since $B = \lfloor \alpha K \rfloor$, we have $B\sqrt{J} = \sqrt{\alpha}\, \sqrt{JKB}\, \big(1 + o(1)\big)$. We obtain

$$\begin{aligned}\mathbb{E}\Big[\sum_{i \in \mathcal{B}} \mathbf{1}\{T_i = 1\}\Big] &= Bp \\ &+ \sqrt{2 \ln 2\, p(1-p)\, \alpha}\, m_G(\alpha)\, \sqrt{JKB} \\ &+ o\big(\sqrt{JKB}\big).\end{aligned}$$

Combining with equation 25 yields equation 9. □

B.6 Proof of Corollary 11

Proof. The upper bound is Theorem 6. For the lower bound, $D^\star(K,B)$ is the minimum Bayes log-loss over all feasible policies. It is therefore no smaller than the log-loss achieved by the score-based policy in Theorem 10. □

B.7 Proof of Proposition 12

Proof. For a continuous distribution, equation 8 reduces to a conditional mean. Let q_α be the $(1-\alpha)$-quantile of G. For $G \sim \mathcal{N}(0,1)$,

$$m_G(\alpha) = \mathbb{E}[G \mid G \geq q_\alpha].$$

A standard identity for the truncated normal gives

$$\mathbb{E}[G \mid G \geq q] = \frac{\varphi(q)}{1-\Phi(q)}, \tag{28}$$

where $\varphi(q) = \frac{1}{\sqrt{2\pi}} e^{-q^2/2}$ and Φ is the standard normal cdf.

We bound $1-\Phi(q)$ in terms of $\varphi(q)$ for $q > 0$. Using integration by parts and the identity $\varphi'(u) = -u\,\varphi(u)$ yields the upper bound

$$1-\Phi(q) \leq \frac{\varphi(q)}{q}. \tag{29}$$

A second integration by parts gives the lower bound

$$1-\Phi(q) \geq \frac{\varphi(q)}{q+1/q}. \tag{30}$$

Substituting equation 29 and equation 30 into equation 28 yields

$$q_\alpha \leq m_G(\alpha) \leq q_\alpha + \frac{1}{q_\alpha}.$$

Therefore, it is enough to identify the leading growth of q_α.

Since $\alpha = 1-\Phi(q_\alpha)$, the bound equation 29 gives

$$\alpha \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-q_\alpha^2/2}}{q_\alpha}.$$

Taking logarithms yields

$$\frac{q_\alpha^2}{2} \leq \log\frac{1}{\alpha} - \log q_\alpha + O(1).$$

This implies $q_\alpha^2 \leq 2\log(1/\alpha) + O(\log\log(1/\alpha))$. The bound equation 30 gives a matching lower bound and yields

$$q_\alpha = \sqrt{2\log\frac{1}{\alpha}}\,(1+o(1)).$$

Since $m_G(\alpha)$ differs from q_α by at most $1/q_\alpha$, we obtain equation 14. □

B.8 Proof of Proposition 13

Proof. Let X have the Pareto tail $\mathbb{P}(X \geq x) = x^{-\nu}$ for $x \geq 1$, with $\nu > 3$. The mean and variance are

$$\mu := \mathbb{E}[X] = \frac{\nu}{\nu-1}, \qquad \sigma^2 := \mathrm{Var}(X) = \frac{\nu}{(\nu-1)^2(\nu-2)}.$$

Define $G = (X-\mu)/\sigma$. For $\alpha \in (0,1)$, let q_α be the $(1-\alpha)$-quantile of G. Since G is an increasing function of X, we have

$$\mathbb{P}(G \geq q) = \mathbb{P}(X \geq \sigma q + \mu).$$

The $(1-\alpha)$-quantile of X is $x_\alpha = \alpha^{-1/\nu}$. Therefore $q_\alpha = (x_\alpha - \mu)/\sigma$.

For a continuous distribution, $m_G(\alpha) = \mathbb{E}[G \mid G \geq q_\alpha]$. We compute

$$\begin{aligned}\mathbb{E}\big[G\,\mathbf{1}\{G \geq q_\alpha\}\big] &= \frac{1}{\sigma}\,\mathbb{E}\big[(X-\mu)\,\mathbf{1}\{X \geq x_\alpha\}\big] \\ &= \frac{1}{\sigma}\Big(\mathbb{E}[X\,\mathbf{1}\{X \geq x_\alpha\}] - \mu\,\mathbb{P}(X \geq x_\alpha)\Big).\end{aligned}$$

For the Pareto density $f(x) = \nu x^{-\nu-1}$ on $x \geq 1$, we have

$$\begin{aligned}\mathbb{E}[X\,\mathbf{1}\{X \geq x_\alpha\}] &= \int_{x_\alpha}^\infty x f(x)\,dx \\ &= \nu\int_{x_\alpha}^\infty x^{-\nu}\,dx = \frac{\nu}{\nu-1}\,x_\alpha^{1-\nu}.\end{aligned}$$

We also have $\mathbb{P}(X \geq x_\alpha) = x_\alpha^{-\nu}$. Dividing by $\alpha = \mathbb{P}(X \geq x_\alpha)$ yields

$$\begin{aligned}m_G(\alpha) &= \mathbb{E}[G \mid G \geq q_\alpha] \\ &= \frac{1}{\sigma}\Big(\frac{\nu}{\nu-1}\,x_\alpha - \mu\Big) = \frac{1}{\sigma}\Big(\frac{\nu}{\nu-1}\,\alpha^{-1/\nu} - \mu\Big).\end{aligned}$$

As $\alpha \to 0$, the additive constant μ/σ is negligible compared to $\alpha^{-1/\nu}$. This gives equation 15. □

C Proofs for Appendix A

C.1 Proof of Theorem 16

Proof. Fix any verification policy and its verification set $\mathcal{B} \subseteq [K]$. We condition on the inspection outcomes $Z^K := (Z_1,\dots,Z_K)$. Under the benchmark model, types are conditionally independent given Z^K, and $\mathbb{E}[T_i \mid Z_i] = \eta(Z_i) =: \eta_i$. Therefore,

$$\mathbb{E}\Big[\sum_{i\in\mathcal{B}} T_i \mid Z^K\Big] = \sum_{i\in\mathcal{B}} \mathbb{E}[T_i \mid Z^K] = \sum_{i\in\mathcal{B}} \eta_i.$$

For fixed scores $\eta_1,\dots,\eta_K$ and a fixed budget $|\mathcal{B}| = B$, the sum is maximized by choosing the indices of the B largest scores. Let $\eta_{(1)} \geq \cdots \geq \eta_{(K)}$ denote the order statistics. Then

$$\sum_{i\in\mathcal{B}} \eta_i \leq \sum_{\ell=1}^{B} \eta_{(\ell)}.$$

Taking expectations over Z^K yields equation 19. The upper bound is achieved by verifying the B largest scores. □

C.2 Proof of Theorem 17

Proof. Consider a verified record $i \in \mathcal{B}$. If $T_i = 0$, then by Assumption 15 the verification output is the constant symbol $V_i = v_0$. The Bayes optimal prediction is the prior P_Θ, and the expected log-loss equals $\mathrm{H}(\Theta)$. If $T_i = 1$, then V_i is generated by $P_{V|\Theta}$. The Bayes optimal prediction is the posterior $P(\Theta_i \mid V_i)$, and the expected log-loss equals $\mathrm{H}(\Theta \mid V)$. By the definition of mutual information, $\mathrm{H}(\Theta \mid V) = \mathrm{H}(\Theta) - I_{\mathrm{ver}}^{\mathrm{bm}}$. Therefore,

$$\mathbb{E}\left[-\log q_i(\Theta_i)\right] = \mathrm{H}(\Theta) - I_{\mathrm{ver}}^{\mathrm{bm}}\,\mathbb{E}[T_i].$$

Averaging over the B verified records yields

$$\bar{D}(K,B) = \mathrm{H}(\Theta) - \frac{I_{\mathrm{ver}}^{\mathrm{bm}}}{B}\,\mathbb{E}\Big[\sum_{i\in\mathcal{B}} T_i\Big].$$

Minimizing $\bar{D}(K,B)$ is equivalent to maximizing $\mathbb{E}[\sum_{i\in\mathcal{B}} T_i]$. Theorem 16 gives

$$\mathbb{E}\Big[\sum_{i\in\mathcal{B}} T_i\Big] \le \mathbb{E}\Big[\sum_{\ell=1}^{B} \eta_{(\ell)}\Big].$$

Substituting into the previous display yields the converse $\bar{D}(K,B) \ge \bar{D}^\star(K,B)$ with $\bar{D}^\star$ defined in equation 20. For achievability, verify the B largest scores. Then Theorem 16 holds with equality and the Bayes predictors achieve the stated risk. □

C.3 Proof of Corollary 18

Proof. Let $\eta_1, \ldots, \eta_K$ be i.i.d. samples of $\eta(Z)$ with a continuous distribution. Fix $\alpha \in (0,1)$ and set $B = \lfloor \alpha K \rfloor$. Define the population threshold $t_\alpha := Q_\eta(1-\alpha)$. Continuity implies $\mathbb{P}(\eta = t_\alpha) = 0$. Standard order-statistics theory implies the empirical quantile converges almost surely:

$$\eta_{(B)} \to t_\alpha, \qquad \text{as } K \to \infty.$$

Since there are no ties almost surely,

$$\frac{1}{K}\sum_{\ell=1}^{B} \eta_{(\ell)} = \frac{1}{K}\sum_{i=1}^{K} \eta_i\,\mathbf{1}\{\eta_i \ge \eta_{(B)}\}.$$

The difference between this term and $\frac{1}{K}\sum_{i=1}^{K} \eta_i\,\mathbf{1}\{\eta_i \ge t_\alpha\}$ is controlled by the empirical measure of the interval between $\eta_{(B)}$ and t_α. This measure converges to 0 almost surely. Therefore,

$$\frac{1}{K}\sum_{\ell=1}^{B} \eta_{(\ell)} \to \mathbb{E}\left[\eta\,\mathbf{1}\{\eta \ge t_\alpha\}\right] = \int_{1-\alpha}^{1} Q_\eta(u)\,du.$$

This proves equation 21. Substituting into equation 20 yields equation 22. □

References

[1] Anais Galdin and Jesse Silbert. Making talk cheap: Generative ai and labor market signaling. *arXiv preprint arXiv:2511.08785*, 2025.

[2] Michael Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973.

[3] Christopher A Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

[4] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[5] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[6] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*, volume 1. John Wiley & Sons, 2006.

[7] Lav R. Varshney. *Unreliable and Resource-Constrained Decoding*. Ph.d. thesis, Massachusetts Institute of Technology, 2010.

[8] Haim Permuter and Tsachy Weissman. Source coding with a side information "vending machine". *IEEE Transactions on Information Theory*, 57(7):4530–4544, 2011.

[9] Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015. doi: 10.1257/aer.20130047.

[10] Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015.

[11] Bartosz Maćkowiak, Filip Matějka, and Mirko Wiederholt. Rational inattention: A review. *Journal of Economic Literature*, 61(1):226–273, 2023.

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

[13] John N Tsitsiklis. Decentralized detection by a large number of sensors. *Mathematics of Control, Signals and Systems*, 1(2):167–182, 1988.

[14] Pramod K Varshney. *Distributed detection and data fusion*. Springer Science & Business Media, 2012.

[15] Venugopal V Veeravalli and Pramod K Varshney. Distributed inference in wireless sensor networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):100–117, 2012.

[16] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

[17] Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.

[18] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100 (5):992–1026, 1992.

[19] Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012.

[20] Thomas A Courtade and Tsachy Weissman. Multiterminal source coding under logarithmic loss. *IEEE Transactions on Information Theory*, 60(1):740–761, 2013.

[21] Bertrand S Clarke and Andrew R Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on information theory*, 36(3):453–471, 2002.

[22] Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[23] Jorma Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information theory*, 30(4):629–636, 2003.

[24] Alankrita Bhatt and Young-Han Kim. Sequential prediction under log-loss with side information. In *Algorithmic Learning Theory*, pages 340–344. PMLR, 2021.

[25] DAVID DONOHO and JIASHUN JIN. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[26] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.

[27] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[28] Robert Dorfman. The detection of defective members of large populations. *The Annals of mathematical statistics*, 14(4):436–440, 1943.

[29] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

[30] Haikady N Nagaraja and Herbert A David. *Order statistics*. Wiley, New Jersey, 2003.

[31] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

[32] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

[33] Rishi Bommasani. On the opportunities and risks of foundation models, 2021.

[34] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[35] R OpenAI. Gpt-4 technical report. arxiv 2303.08774, 2023.

[36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, pages 9459–9474, 2020.

[37] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. URL https://arxiv.org/abs/2112.09332.

[38] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

[39] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.04761.

[40] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tennenholtz. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*, 2022. URL https://arxiv.org/abs/2205.00445.

[41] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, dec 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

[42] Lele Cao, Lei You, and R&d Team. Cspaper review: Fast, rubric-faithful conference feedback. In *Proceedings of the 18th International Natural Language Generation Conference: System Demonstrations*, pages 3–7, 2025.

[43] Isidore Jacob Good. Probability and the weighing of evidence. 1950.

[44] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.