The Courtroom Challenge—A Conjecture on the Unavoidable Cost of LLM Verification

John Repsys[*]

August 3, 2025

Abstract

Large-scale generative models hallucinate, and ad hoc patches hide the true cost of verification. We pose a falsifiable ultimatum: automate validation, prove it unnecessary, or pay the explicit human-in-the-loop bill. We conjecture that for any unbounded open-world data stream with Shannon entropy rate $h > 0$, a system that lacks both a persistent precedent store and a friction-priced validation loop will accumulate unbounded prequential regret. To test this "validation trilemma" we formalise a state machine, a cost ledger, and an impossibility triangle, then invite the community to break it. A 100-query toy stream illustrates how the ledger is logged; optimising the validator is left to future work. We also introduce the Courtroom Model, a wrapper that subjects every claim—textual, visual, or otherwise—to precedent search, automated semantic checks, and selective escalation to an external judge. Each output is checked against persistent vector memory, and conflicting evidence reopens prior "case law." The challenge stands: thrive without paying the validation bill—or show why the bill is inescapable.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in generating context-rich outputs by associating patterns from large corpora. However, this power comes with a critical limitation: LLMs often produce plausible-sounding but factually incorrect conclusions, a phenomenon known as hallucination (Marcus, 2020; Shinn et al., 2023). Without structured verification, these errors risk being perpetuated and even amplified over time, undermining trust in LLM-generated knowledge. Addressing this issue requires a system that not only generates associations but rigorously validates and refines them. Throughout this paper we use the term *Courtroom* metaphorically: it denotes an epistemic due-process layer that challenges and verifies LLM outputs, not a system for practicing law or adjudicating legal disputes.

To address these shortcomings we pose the *Courtroom Challenge*: a hybrid state machine in which every LLM claim is tested against a persistent precedent store and—when automatic checks falter—escalated under a tunable *validation_preference* dial (formal spec in Section 11). The design rewires known parts (generator, retrieval, human oversight) but prices each byte, cycle, and minute of validation. Retrieval-augmented generation (RAG) provides the raw materials (Lewis et al., 2020; Fan et al., 2024). Recent advances—Self-RAG's self-verification (Asai et al., 2024), Iter-RetGen's retrieval-generation synergy (Shao et al., 2023), MemoRAG's compressed long-term memory (Qian et al., 2024), JoT's judgment-of-thought prompting (Park et al., 2024), and Chain-of-Verification

---

[*]email: john.repsys@gmail.com
[1]Baseline script and toy data: https://github.com/johnrepsys/courtroom-model

RAG's answer-driven re-checks (He et al., 2024)—motivate our persistent precedent store but still validate only per answer. Human-alignment work such as Constitutional AI (Bai et al., 2022) motivates the escalation branch. Whereas MemoRAG retains past evidence without auditing it, our buffer makes each accepted verdict both permanent and costly. Unlike debate frameworks (J. Chen et al., 2024; Madaan et al., 2023) or formal provers (Wang et al., 2025), we emphasise continuous cross-session verification: knowledge is never final, only current.

Why a conjecture first? Well-posed problems often precede the data needed to solve them. We therefore publish the validation trilemma now: automate open-world verification, prove it unnecessary, or price human judgment honestly. Curves can wait until the target is unambiguous.

Scope and contribution. This is a conjecture note—*no curves, by design.* We provide a falsifiable protocol and open-source meta-configuration so that anyone can measure cost-adjusted regret. Generators are already benchmarked; the Courtroom Challenge benchmarks verification. We invite the community to break the conjecture or beat the baseline and thereby close the last mile in trustworthy AI.

2 The Courtroom Model: Legal Precedent & Debate

Humans think by generating ideas and challenging them. In scientific and philosophical inquiry, ideas are not just created—they are debated, tested, and refined. As discussed in Section 1, the Courtroom Model integrates LLMs for idea generation, FAISS/Parquet for memory, and feedback loops for validation. In this section, we'll explore how these elements interact in greater detail.

Once an LLM generates an idea, it enters the 'courtroom'—where it is tested against structured precedent and subjected to rigorous debate. Ideas are challenged and refined, ensuring that only those that withstand scrutiny are accepted as truth. The Courtroom Model keeps every claim—new or old—under continual scrutiny, revising or discarding assumptions as fresh evidence arrives and thereby maintaining an evolving knowledge base rather than a static one.

Picture cumulative human knowledge as the area under a quarter-circle frontier. LLM association pushes the curve outward; the Courtroom validator folds it inward wherever new evidence pierces old beliefs. Expansion and maintenance run in lock-step, keeping the frontier smooth instead of ballooning into noisy bubbles. The prequential-regret metric defined in Section 11.5 is the algebraic mirror of this movie: outward growth reduces error, but only if the validator's inward checks prevent unchecked drift.

The Courtroom Model introduces a hybrid framework that distinguishes itself from current AI methodologies by integrating associative reasoning (LLMs) with the persistent structure of legal precedents (FAISS/Parquet) and dynamic friction-priced validation (see Figure 1). This combination enables the model to adapt, incorporate external judgment, and apply continuous scrutiny—features largely absent in systems like ReConcile or MA-LoT.

Rather than requiring manual review of every idea, the Courtroom Model automatically flags associations that meet certain thresholds of uncertainty or inconsistency with established facts. This approach allows humans to focus on the most critical evaluations, ensuring that their intervention is directed toward areas where it adds the most value.
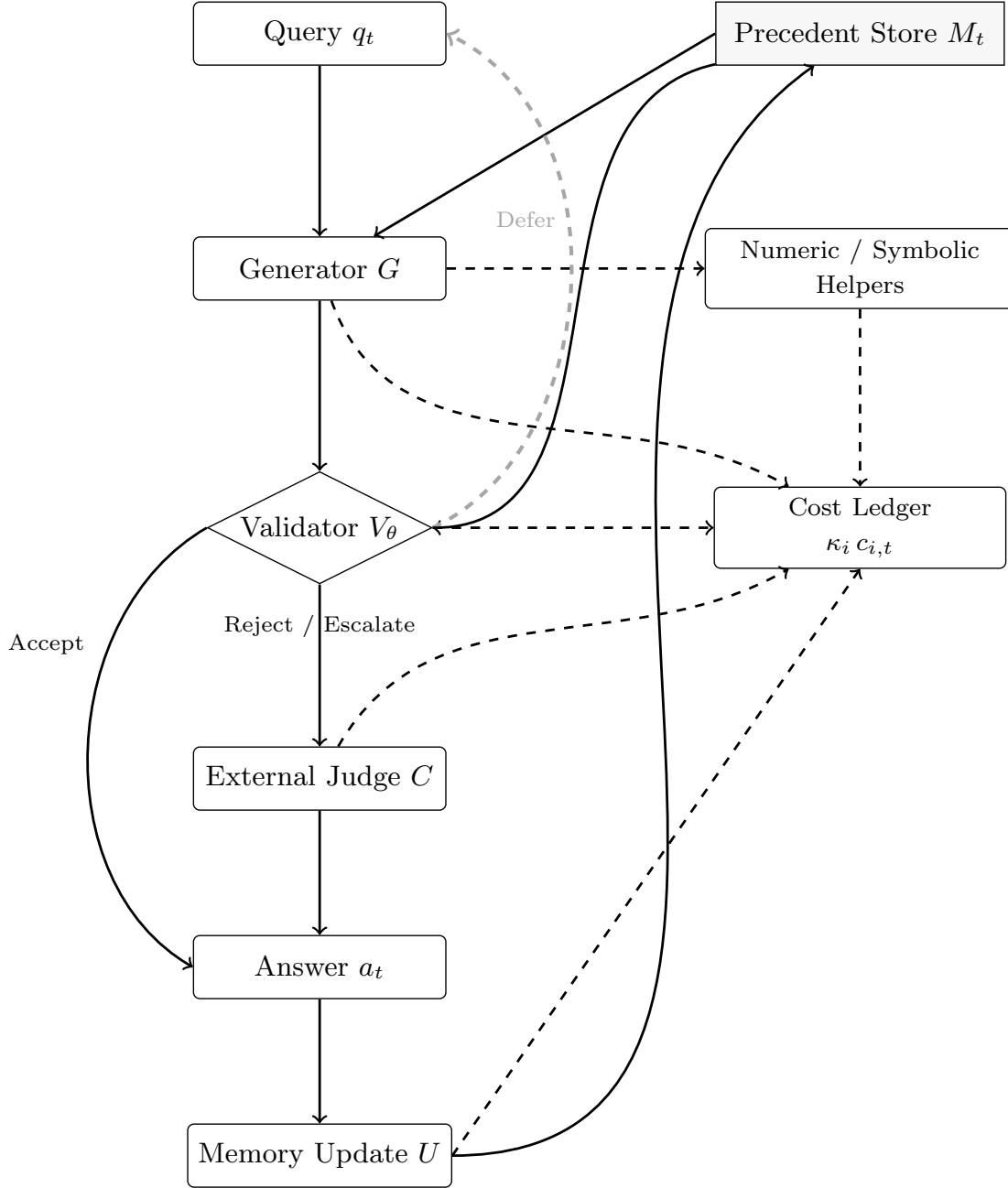
Figure 1: One step of the Courtroom loop. Solid arrows: primary query flow. Dashed arrows: helper calls and cost logging. The cost bus aggregates all resource writes into a single ledger. On DEFER the round is skipped (no cost, no memory write). If C cannot supply an authoritative value—because the fact is genuinely unknown—the query transitions to a terminal UNRESOLVED state; no memory write occurs.

Illustrative trace (informal)

Query q_t. *"Wells Fargo's 2018 10-K introduces a line `Debt securities`. (1) Map this line item to the existing FY-2017 bucket taxonomy, and (2) report the FY-2018 amount in that bucket."*

1. Generator $G(q_t, M_t) \longrightarrow h_t$ The precedent store M_t has no FY-2018 descriptor for `"Debt Securities"`. Embedding search ranks (`WFC, "Investment Securities", FY2017, 10.664 B`) as the nearest neighbour (high string similarity). The embedding vector and its nearest neighbour are cached and handed to the validator in the same step. Assuming the new label is merely a rename, the LLM—assisted by a local XBRL helper that extracts the numeric value—relabels the bucket and proposes

$$h_t = \text{"Bucket: \texttt{"Investment Securities"}; FY-2018 amount: \$10.664 B."}$$

The helper itself is numerically correct; the label choice is the error.

2. Validator $V_\theta(q_t, h_t, M_t) \longrightarrow$ ESCALATE Using the same lookup result, the validator sees that the query descriptor `WFC | "Debt Securities" | FY2018` is only $\cos = 0.906$ similar to its nearest neighbour `"Investment Securities"`. Because the illustration fixes the retention threshold at $\tau = 0.95$, this similarity falls outside the allowed neighbourhood, so the claim is escalated—no domain-specific rule required.

3. External judge C A human analyst consults the 2018 10-K and confirms that `Debt Securities` is a new line-item, not yet present in the taxonomy. The analyst extracts the FY-2018 amount

$$a_t = \text{"Bucket: \texttt{"Debt Securities"}; FY-2018 amount: \$14.406 B."}$$

4. Memory update $U(\dots) \longrightarrow M_{t+1}$ Because `"Debt Securities"` represents a genuinely new bucket (label and composition differ from `"Investment Securities"`), U appends a fresh descriptor vector `WFC|"Debt Securities"|FY2018` $\rightarrow$ 14.406 B. The existing `"Investment Securities"` tuples remain intact; an optional *alias edge* can later be added if domain experts decide the two buckets should roll up to the same parent taxonomy node.

Resource i	Quantity $c_{i,t}$	Weight κ_i	$\kappa_i c_{i,t}$
Vector write (bytes)	1 536 B	1.0×10^{-6} /byte	0.0015
FAISS rebuild (sec)	0.20 s	0.05 /s	0.0100
Human review (minutes)	1.00 min	1.00 /min	1.0000
Step cost			**1.0115**

Table 1: Friction priced for this step (arbitrary κ-weights).

Cost ledger entry $\mathbf{WriteCost}_t = \sum_i \kappa_i \, c_{i,t}$

Storage dial. The precedent store need not—indeed, should not—memorise every numeric fact. A single *retention knob* (τ) decides whether a freshly extracted descriptor vector is written back or simply reuses its nearest neighbour. Let $S(x,y)$ be any similarity score where "larger means closer" (or equivalently a distance $D(x,y)$ where "smaller means closer"). The generic rule is

$$\text{write\_new} = \begin{cases} \big[S(x,y) < \tau\big], & \text{(similarity form)} \\ \big[D(x,y) > \tau\big], & \text{(distance form).} \end{cases}$$

Cosine and L_2 are special cases, but a system designer is free to plug in any monotone metric—e.g. Mahalanobis distance, dot-product in a learned embedding space, or even a hybrid score that blends semantic and numeric keys—so long as the inequality is flipped to match the metric's notion of proximity. Section 11.8 analyses the cost/error trade-off under the distance form, and Appendix A.1 lists this dial among the falsifiability-critical degrees of freedom. Selecting which descriptor fields to embed is itself tunable; Section 7 (Q5) sketches how the Courtroom loop could learn or prune schema columns over time.

3 A Practical Implementation: Financial Data as a Reasoning Playground

While much of the LLM + formal proof conversation focuses on pure math (e.g., Lean), the Courtroom Model has broad applications, including complex domains like financial data. In this space, ambiguity and structure meet, and semantic checks become crucial:

- SEC filings often contain inconsistent naming and formats. LLMs can provide initial opinions on relationships between known metrics.

- Vector embeddings allow semantic similarity checks with prior precedent, using metrics such as L_2-distance or cosine similarity to quantify the closeness between new ideas and established knowledge. These checks focus on identifying close matches to validated concepts and flagging potential outliers, ensuring that the system not only finds similarities but also highlights discrepancies that might indicate the need for deeper scrutiny or re-evaluation.

- FAISS performs nearest-neighbor searches over high-dimensional vectors, helping to identify similar concepts based on predefined thresholds.

- Parquet files serve as version-aware memory, storing validated items, like legal precedent.

- In financial analysis, when a generative model suggests a correlation or metric, human analysts review it against historical data, verify the logic, and confirm its validity. This friction ensures errors do not become accepted facts.

This process embodies the Courtroom Model: association meets structure, refined and verified by friction. This perpetual validation loop not only updates the knowledge base as new data arrive; it also prioritises which claims to re-examine—triggered by real-world data, symbolic checks, human oversight, or similarity signals (e.g., cosine/L_2 distances)—so that older beliefs face deeper scrutiny whenever fresh evidence challenges them. In practice, the model could leverage dynamic prioritization algorithms to determine which associations or precedents are most in need of re-evaluation. This approach would enable the model to scale effectively to large datasets, focusing scrutiny on the most impactful claims and reducing computational load while maintaining rigorous validation. While this is not yet implemented, it represents a key area for future development, enhancing the system's efficiency and scalability.

4 Are LLMs a Dead End? Converging Error

Critics such as LeCun (LeCun, 2023) argue that purely autoregressive LLMs will always accumulate error: if the per-token slip rate is e, the probability an n-token answer is entirely correct decays as $P_{\text{correct}} = (1 - e)^n$. LeCun also notes the inability of LLMs to represent continuous high-dimensional spaces that characterize the real world, potentially limiting their ability to deal with complex, nonlinear data.

However, this does not mean we must abandon LLMs. Rather than expecting LLMs to handle everything on their own, the Courtroom Model allows them to excel at generating ideas and associations, while supplementing them with additional systems that anchor their outputs and guide them toward more reliable outcomes. By integrating persistent memory (like legal precedent) and real-world constraints (courtroom debate), the model offers a framework for continuously verifying and refining LLM-generated associations, ensuring they align with established facts and real-world data. This helps overcome challenges posed by complex, high-dimensional data, and can mitigate error propagation.

Meanwhile, the question "Are LLMs a dead end?" continues to evolve as part of the broader debate on AI's future. Some initially speculated that superintelligence was near, while others, like LeCun, pushed back, suggesting that LLMs' potential was overstated. The debate remains open—and that is exactly the point: we need friction to continue exploring ideas and improving models. The Courtroom Model offers one possible way forward by leveraging LLMs in combination with other systems, ensuring that we don't discard their value but integrate them into a broader framework that allows them to evolve meaningfully. While critics argue that LLMs have stalled AI progress, it may be more accurate to say that we're just learning to build one side of the brain, and as we expand our approach, we must continue building the other side as well.

5 Multi-LLM Dialogues: An Interim "Court"

Several approaches have attempted to improve LLM reasoning through collaboration or debate among multiple agents. Systems like ReConcile (J. Chen et al., 2024) and Self-Refine (Madaan et al., 2023) demonstrate that critical dialogue and iterative refinement between LLMs can enhance factual consistency and reasoning depth. However, these methods typically operate in a single-session paradigm, stabilizing on a consensus answer without maintaining a persistent knowledge

base. In contrast, the Courtroom Model incorporates continuous precedent management, ensuring that knowledge evolves dynamically over time rather than resetting between tasks.

6 Architectural Analogues

Structured exploration, exemplified by AlphaGo's policy-value search (Silver et al., 2016) and more recently by Tree-of-Thoughts planning (Yao et al., 2023), offers a compelling model for integrating creativity with systematic validation. In AlphaGo, a policy network generates moves, a value network evaluates outcomes, and self-play iteratively refines strategy. The Courtroom Model generalizes this principle to open-ended domains, applying continuous validation not only internally but also against a persistent external memory and, when needed, human oversight. This convergence supports the idea that the Courtroom Model is not prescriptive, but descriptive — a lens for recognizing effective reasoning architectures wherever they emerge. It offers a retrospective framework to describe such structures — and perhaps guide the design of future systems that balance creativity with constraint.

In formal mathematics, frameworks such as Lean (mathlib Community, 2020) and multi-agent proof systems like MA-LoT (Wang et al., 2025) demonstrate that combining generative reasoning with strict verification pipelines can achieve rigorous correctness. These systems ensure that every step adheres to formal logic before being accepted. Inspired by these principles, the Courtroom Model aims to bring analogous rigor to less formally structured domains, using semantic similarity, precedent management, and a friction-priced external judge—human or automated—to prevent the unchecked propagation of errors.

Both AlphaGo and Lean + LLM workflows demonstrate that progress often arises from structured friction, not just creative leaps. In the same way, the Courtroom Model allows for creativity (via LLMs) to drive idea generation, but that creativity is continually tested and refined through structured validation and external challenge, ensuring only robust, verified ideas become part of the model's evolving knowledge.

7 Open Questions

1. Is abstraction a distinct step in hybrid reasoning? When no exact precedent exists, does the model create a new one by compressing multiple associations? Should pattern recognition and abstraction be formal components in systems like this one?

2. Can LLMs + real-world observation produce entirely new theories? If so, how do we detect genuine novelty, and how do we measure validity?

3. Can the Courtroom Model generalize to other real-world domains and AI integrations? Healthcare, policy, engineering — any area that needs scale, creativity, and robust checks. Could this framework point toward a general pattern for hybrid AI?

4. How can we achieve Lean-level formal rigor in open-text domains without relying solely on large LLM ensembles, whose granularity still falls short in many fields?

5. How should the precedent store be schematised and evolved? Even in a well-defined vertical like finance, designing the descriptor schema (e.g. '(ticker, canonical_metric, period, unit, GAAP_flag, source_tag, . . .)') remains an art: too coarse results in semantic collisions; too fine results in unlimited vector growth. Can the system learn an optimal schema online—merging rarely-queried fields, splitting ambiguous ones, and doing so without breaking

L_2-distance semantics? What signals (query regret, type-confusion rate, index rebuild cost) best drive such schema evolution?

6. How do we maintain the tension between speed and scrutiny? Associative models want to generate rapidly; symbolic structures demand thorough verification. Speed wins until it fails. Scrutiny survives because it must.

7. What does success look like in a system that's always evolving? As new precedents form, do we prune old ones? How do we adapt long-standing truths? The answer may lie in how far our energy limitations let us revisit and revise what we already hold to be true.

8. Which self-play or adversarial-debate schemes could fill the V_θ slot while keeping κ_{compute} at or below the current human-review baseline?

> **Why a concept note?**
>
> The field currently optimises ever-larger generators without a matching theory of validation. We publish this abstraction before a prototype for two reasons: (i) to surface hidden degrees of freedom that can silently invalidate benchmarks, and (ii) to recruit empirical collaborators—open-sourcing code is planned for v2.

8   Toward Hybrid Intelligence

Purely neural stacks still stumble on reliability. Recent work (Marcus, 2020; Dawid and LeCun, 2024) shows that scale alone cannot guarantee robust reasoning. This has revived calls for neuro-symbolic fusion: pair an associative generator with a structured verifier. The Courtroom Challenge operationalises that philosophy. It links:

- an LLM's creative search,
- a persistent precedent memory, and
- an external judge (self-play agent, formal prover, ensemble LLM, or human),

into a single friction-priced loop that can correct itself over time. Reasoning at scale demands all three roles:

- Generate & explore — propose novel chains of thought.
- Validate & prune — confront each claim with precedent, logic, and counter-evidence.
- Evolve the canon — admit only those claims that survive scrutiny, while logging the cost of that scrutiny.

Without this three-way handshake, knowledge either drifts (if validation is cheapened) or ossifies (if exploration is throttled). The Courtroom specification is therefore a minimal test-bed for hybrid intelligence: break the conjecture, or show how your system pays the validation bill more efficiently.

9 Conclusion: A New Frontier

Progress in reasoning systems emerges when every new idea must survive structured challenge. The Courtroom Model unites neural creativity, symbolic memory, and real-world friction in a single, continuously verifiable loop. By insisting that accepted knowledge remains open to appeal whenever fresh evidence appears, it transforms static retrieval or one-off debate into a living jurisprudence of facts. Future work will automate domain-specific validators and refine escalation thresholds, but the central thesis stands: reliable AI will look less like a confident oracle and more like an evolving court—where claims, evidence, and precedent are in constant, accountable dialogue. If Conjecture 1 holds, genuinely autonomous, self-improving AI will require breakthroughs in scalable validation—pushing practical ASI timelines beyond what generator-only scaling curves imply.

10 Minimal State Machine

Formal state $S_t = (q_t, h_t, a_t, M_t)$, validator V_θ, memory update U, cost ledger $\text{WriteCost} = \sum_i \kappa_i c_i$. Validator actions: $\{\text{ACCEPT}, \text{REJECT}, \text{ESCALATE}, \text{DEFER}\}$.

11 Formal Framework

11.1 State Spaces

- Query space Q

- Hypothesis space H (outputs of the LLM generator G)

- Answer space A

- Memory space $M \subseteq \mathbb{R}^d$ (FAISS/Parquet vectors with payload)

As defined in Section 10, state

$$S_t = (q_t,\ h_t,\ a_t,\ M_t), \qquad M_t \text{ finite.}$$

11.2 Components

- **Generator.** $G : Q \times M \to H$ is an associative engine (e.g. an LLM) that proposes a hypothesis h_t. G may issue lightweight helper calls to numeric or symbolic libraries during exploration; such calls accrue κ_{compute} in the cost ledger but do not decide final truth.

- **Validator family.** $V_\theta : (Q, H, M) \to \{\text{ACCEPT}, \text{REJECT}, \text{ESCALATE}, \text{DEFER}\}$, with a tunable validation-preference dial $\theta \in \Theta$. V_θ may be (i) a stronger ensemble LLM, (ii) a formal-proof kernel, (iii) a deterministic numeric or symbolic tool-chain that certifies or refutes the claim, or (iv) a human panel as fallback.[1] DEFER means "skip this round": no answer is returned, no escalation cost is paid, and the generator proceeds to the next query.

- **Memory update rule.** $U : (M, q, h, a) \to M'$ records (or merges) a precedent whenever a claim is accepted. Whereas vanilla RAG treats retrieved passages as a transient context window (Lewis et al., 2020)—and even iterative or self-reflective extensions still validate only

---

[1] *Self-play precedent.* AlphaZero's value network functions as an automated judge for Monte-Carlo rollouts. The Courtroom validator is its open-domain analogue: any module that can reliably downgrade faulty claims can occupy the judge slot.

the current answer (Shao et al., 2023; Asai et al., 2024)—the Courtroom store is persistent: it survives across sessions and every entry can later be re-audited for contradiction or obsolescence.

Provisionality of verdicts. Even the external Judge's verdicts are provisional: they may be rescinded if subsequent audits or new primary filings contradict earlier conclusions. Accordingly, Judge-validated precedents can be re-queued, incur additional WriteCost, and overwrite prior entries via the usual supersession mechanism.

11.3 One–Step Transition

Given $S_t = (q_t, h_t, a_t, M_t)$:

1. $h_t \leftarrow G(q_t, M_t)$

2. $d_t \leftarrow V_\theta(q_t, h_t, M_t)$

   - If $d_t = \text{ACCEPT}$ set $a_t = h_t$
   - Else obtain corrected answer $a_t = C(q_t, h_t)$ (oracle C may call a human; cost is logged)
   - If $d_t = \text{DEFER}$ set $a_t = \bot$ and skip Steps 3-4 below.

3. $M_{t+1} \leftarrow U(M_t, q_t, h_t, a_t)$

4. External stream provides q_{t+1}

11.4 Cost Ledger

Let $c_{i,t}$ be the measured consumption of resource i at step t (bytes written, FAISS rebuild time, human minutes, etc.). Fix weights κ_i once per benchmark—should the conjecture survive first contact. The per-write cost is

$$\text{WriteCost}_t = \sum_i \kappa_i\, c_{i,t}.$$

Cumulative friction budget ... $F_T = \sum_{t=0}^{T-1} \text{WriteCost}_t.$[2]

11.5 Performance Measures

Instant error

$$e_t = \mathbb{I}\big[a_t \neq \text{ground}(q_t)\big].$$

Prequential regret after budget B:

$$R(B) = \sum_{t=0}^{T(B)-1} e_t, \quad T(B) = \min\{\,T : F_T \geq B\,\}.$$

11.6 Setup

The definitions in §11 merely fix notation $(T, \kappa, \text{Err}, \theta, X)$; they do not restrict the validator's internal design. We now state Lemma 1, an information-theoretic lower bound that applies to any interactive validator whose transcript carries at most I bits of information about the hidden label. The inequality therefore functions as motivation for the Courtroom architecture rather than a theorem that assumes it.

---

[2]DEFER incurs zero cost, so the corresponding WriteCost_t entry is 0.

Per-round notation. Fix a margin parameter $\delta \in (0, 1]$. For every run ω and round $t = 1, \ldots, T$ set

$$
\begin{aligned}
y_t &:= y_t(\theta) && \text{(ground-truth label)}, \\
h_t &:= h_t(\omega) && \text{(validator's automatic prediction)}, \\
\mathrm{Esc}_t &:= \mathbf{1}_{\{\text{round } t \text{ is escalated}\}}.
\end{aligned}
$$

Define the *leverage rounds*

$$
\mathcal{L}(\omega) := \Big\{ t : \mathrm{Esc}_t = 0 \text{ and } |y_t - h_t| \geq \delta \Big\}, \qquad \kappa(\omega) := \frac{|\mathcal{L}(\omega)|}{T}. \tag{1}
$$

The cumulative label error on those rounds is

$$
\mathrm{Err}(\omega) := \sum_{t \in \mathcal{L}(\omega)} \mathbf{1}_{[h_t \neq y_t]}. \tag{2}
$$

For any hypothetical label index m and transcript x we write

$$
\mathrm{Err}_m(x) := \sum_{t \in \mathcal{L}(\omega)} \mathbf{1}_{[h_t(x) \neq y_t(m)]},
$$

i.e. the cumulative error that would occur on the same leverage rounds if the hidden parameter were m. Define the global means:

$$
\bar{\kappa} := \mathbb{E}[\kappa], \qquad \mathrm{Err}_{\mathrm{bar}} := \mathbb{E}[\mathrm{Err}].
$$

11.7   Information-theoretic lower bound

Lemma 1 (Interactive-Fano corollary, general prior)**.** *Let V be any validator run for T rounds. With $\kappa(\omega)$ and $\mathrm{Err}(\omega)$ defined in (1)-(2), write $\mathrm{Err}_{\mathrm{bar}} := \mathbb{E}[\mathrm{Err}]$. Assume the transcript X obeys the information budget $I(\theta; X) \leq I$ under an arbitrary prior μ on the M-point label set $(M \geq 2)$. Define*

$$
H_\star := \begin{cases} H(\theta) & \text{(Shannon entropy)} \\ -\log p_{\max} & \text{(min-entropy)}, \end{cases} \qquad p_{\max} := \max_m \mu(m).
$$

Then

$$
\boxed{\mathrm{Err}_{\mathrm{bar}} \; \geq \; \frac{\delta\, T}{4}\, \kappa_\star \left(1 - \tfrac{I + \log 2}{H_\star} \right), \quad \kappa_\star := \mathbb{E}\big[\kappa \,|\, \mathrm{Err} > \tfrac{1}{2} \delta \kappa T \big]} . \tag{3}
$$

Remark (Entropy convention)**.** *Throughout we write $H_\star$ for whichever entropy measure is appropriate in context—Shannon entropy when a prior is specified, min-entropy when no distributional knowledge is assumed. Uniform priors maximise $H_\star$, producing the strongest (largest) version of the bound, while skewed priors reduce $H_\star$ and weaken the barrier proportionally.*

Proof sketch. Step 1 (separation). Fix $m \neq m'$ and transcript x. On each leverage round $t \in \mathcal{L}(\omega)$ we have $|y_t(\theta) - h_t| \geq \delta$. Hence h_t can coincide with at most one of $y_t(m), y_t(m')$, so

$$
\mathrm{Err}_m(x) + \mathrm{Err}_{m'}(x) \; \geq \; \delta\, \kappa(\omega)\, T . \tag{$*$}
$$

Step 2 (interactive Fano). Set the path-dependent loss

$$
L^\star(m, x) := \mathbf{1}_{\{\mathrm{Err}_m(x) > \frac{1}{2} \delta\, \kappa(\omega)\, T\}}
$$

and $\Delta = \frac{1}{2}$ in Proposition 3 of (F. Chen et al., 2024). Inequality $(*)$ fulfils the separation requirement (F. Chen et al., 2024, Eq. (5)). Define

$$\gamma := \frac{1}{2}\left(1 - \frac{I + \log 2}{H_\star}\right) \in (0, \tfrac{1}{2}].$$

Then

$$\Pr[\mathrm{Err}(\theta, X) > \tfrac{1}{2}\delta\,\kappa(\omega)T] \;\geq\; \gamma. \tag{$\dagger$}$$

Step 3 (tail $\Rightarrow$ expectation, expanded). With

$$\mathcal{E} := \left\{\mathrm{Err} > \tfrac{1}{2}\delta\,\kappa T\right\}, \qquad \kappa_* := \mathbb{E}[\kappa \mid \mathcal{E}],$$

we have:

$$\begin{aligned}
\mathrm{Err}_{\mathrm{bar}} &= \mathbb{E}[\mathrm{Err}] \\
&\geq \mathbb{E}\!\left[\tfrac{1}{2}\delta\,\kappa T\,\mathbf{1}_{\{\mathrm{Err} > \frac{1}{2}\delta\,\kappa T\}}\right] \\
&= \frac{\delta T}{2}\,\mathbb{E}[\kappa\,\mathbf{1}_{\mathcal{E}}] \\
&= \frac{\delta T}{2}\,\kappa_*\,\Pr(\mathcal{E}) \\
&\geq \frac{\delta T}{4}\,\kappa_*\left(1 - \frac{I + \log 2}{H_\star}\right).
\end{aligned}$$

$\square$

Remark (relating κ_* to $\bar{\kappa}$). Write $p_{\mathrm{bad}} := \Pr(\mathcal{E})$ with $\mathcal{E} = \{\mathrm{Err} > \tfrac{1}{2}\delta\kappa T\}$. Since $0 \leq \kappa \leq 1$ one has $\bar{\kappa} = p_{\mathrm{bad}}\kappa_* + (1 - p_{\mathrm{bad}})\kappa_*^{\complement} \leq p_{\mathrm{bad}}\kappa_* + 1 - p_{\mathrm{bad}}$. Hence $\kappa_* \geq \bar{\kappa} + p_{\mathrm{bad}} - 1 \geq \bar{\kappa} + \gamma - 1$, where γ is the Fano factor. Therefore $\kappa_* \geq \max\{0, \bar{\kappa} + \gamma - 1\}$ as used in the following discussion.

Discussion. Write the bound as $\mathrm{Err}_{\mathrm{bar}} \geq (\tfrac{1}{2}\delta T)\max\{0, \bar{\kappa} + \gamma - 1\}$. Because $\gamma > 0$ whenever the transcript leaks fewer than $(H_\star - \log 2)$ bits, any validator that handles a non-trivial share of rounds automatically $(\bar{\kappa} > 1 - \gamma)$ faces a linear-in-T unavoidable error floor. Escalating a larger fraction of rounds (making $\bar{\kappa}$ smaller) or relaxing the margin δ are the only ways to suppress this growth, precisely matching the "friction buys accuracy" narrative.

11.8 Sparse-memory bounds and tightness

We bound how many precedents a validator stores when it keeps only queries that are at least a distance τ apart. The first lemma gives a worst-case upper bound; the second shows that bound is tight under mild stochastic traffic.

Lemma 2 (Sparse-memory upper bound). *Let $\mathcal{Q} \subset \mathbb{R}^d$ be bounded with $\mathrm{diam}\,\mathcal{Q} = D < \infty$. Suppose the validator maintains a set M_t that starts empty and, at each round t, inserts the query q_t only if $\min_{m \in M_{t-1}}\|q_t - m\|_2 > \tau$. Then for all $T \geq 0$*

$$|M_T| \;\leq\; \left(\frac{\mathrm{diam}\,\mathcal{Q}}{\tau}\right)^d.$$

Hence total storage cost is $O(\tau^{-d})$.

Proof sketch. The rule ensures M_T is a τ-packing of $\mathcal{Q}$. By a standard ε-net/volume argument, any τ-packing inside a D-diameter set has size at most $(D/\tau)^d$. (In the worst-case geometry one needs an extra factor 2^d; we omit it because all subsequent results depend only on the τ^{-d} scaling.) $\square$

Lemma 3 (Stochastic tightness of τ-packing)**.** *Let the setting of Lemma 2 hold, and assume in addition that the queries $(q_t)_{t\geq 1}$ are i.i.d. with a density $f \geq \lambda > 0$ on $\mathcal{Q}$. Define*

$$T_0(\tau,\delta) \;=\; \frac{2^d}{\lambda\,\nu_d}\,\tau^{-d}\Big[d\log(D/\tau) + \log(1/\delta)\Big], \qquad \nu_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

Then, for any confidence level $0 < \delta < 1$ and all $T \geq T_0$,

$$\Pr\Big[\,|M_T| \;\geq\; \tfrac{1}{2}\nu_d^{-1}D^{-d}\tau^{-d}\Big] \;\geq\; 1-\delta.$$

In particular, under full-support i.i.d. traffic the memory size scales $\Theta(\tau^{-d})$ with high probability.

Sketch. 1. Maximal 3τ-packing.[3] Let $\{c_i\}_{i=1}^m \subset \mathcal{Q}$ be maximal with pairwise distances $> 3\tau$. Axis-aligned cubes of side $3\tau/\sqrt{d}$ give $m \geq \alpha_d(D/\tau)^d$ with $\alpha_d = 3^{-d}2^{-d}$.
2. Disjoint balls and guaranteed insertion. Set $B_i = B(c_i, \frac{\tau}{2})$. For any query $q_t \in B_i$ and any prior memory point $m \in B_j$ $(j \neq i)$,

$$\|q_t - m\| \;>\; 3\tau - \tfrac{\tau}{2} - \tfrac{\tau}{2} \;=\; 2\tau > \tau,$$

so the insert-iff-distance $> \tau$ rule is automatically satisfied. The balls are disjoint and lie inside $\mathcal{Q}$.
3. Uniform hit probability. By the density lower bound, each ball has $p_\tau = \lambda\nu_d 2^{-d}\tau^d$ probability mass; the value is the same for every i.
4. Empty-ball tail. $\Pr(B_i \text{ empty}) = (1-p_\tau)^T \leq e^{-p_\tau T}$ (using $\log(1-x) \leq -x$). Choosing $T \geq T_0$ forces $e^{-p_\tau T} \leq \delta/m$.
5. Union bound and memory lower bound. With probability $\geq 1-\delta$ every B_i is hit at least once. If the first hit is accepted, it is written to memory; if it is rejected, the blocking memory point m satisfies $\|m - c_i\| < 1.5\tau$ and therefore lies in the expanded ball $\tilde{B}_i := B(c_i, 1.5\tau)$. Because the $\tilde{B}_i$ are disjoint (centre separation $> 3\tau$), each centre still accounts for one distinct stored point, so $|M_T| \geq m \geq \alpha_d(D/\tau)^d$. $\square$

Discussion — from tightness to the trilemma. Lemma 2 caps memory by $O(\tau^{-d})$ for any query sequence. Lemma 3 shows that, under full-support i.i.d. traffic, the same $\Theta(\tau^{-d})$ rate is unavoidable. Together with the Fano-style information bound in Lemma 1 we now have two orthogonal "laws of friction":

- Information ledger. To keep transcripts short, the validator must tolerate a Fano error floor.

- Memory ledger. At granularity τ it must provision $\Theta(\tau^{-d})$ precedent cells—no compression trick beats this when queries explore the space fully.

What remains open—and is precisely what the next two conjectures tackle—is a joint lower bound that couples both ledgers plus the escalation (human-minutes) ledger when queries arrive in an open-world, potentially adaptive fashion. The conjectures formulate that missing link.

---

[3]Using 3τ rather than τ only shrinks the packing constant by 3^{-d}; the $\Theta(\tau^{-d})$ scaling is unaffected.

11.9 Core Conjectures

Conjecture 1 (Validation trilemma)**.** *Let $\mathcal{S} = \langle s_1, \ldots, s_T \rangle$ be an open-world data stream whose source has Shannon entropy rate at least h bits per symbol (equivalently, $H(\mathcal{S}_{1:T}) \geq hT$). For any system that*

> *(i) stores at most B bytes of persistent precedent and*
> *(ii) budgets at most κ bits of external validation per N symbols,*

the prequential regret after it consumes B bytes satisfies

$$R(B) \quad \geq \quad f(h, \kappa, B),$$

for some monotone f (non-increasing in κ, non-decreasing in h).

Conjecture 2 (No-free-lunch verifier)**.** *Under the same entropy assumption, any fully automatic validator V_θ that never escalates to humans must either*

> *(i) mis-accept adversarially crafted claims with probability at least ε, or*

> *(ii) write an unbounded amount of new precedent as $T \to \infty$.*

Falsify-Us

Build an agent that—under the protocol of Section 11—keeps

- cumulative error $\leq \varepsilon |Q|$

- total precedent $\leq \kappa |Q|$

- total human-escalation cost $\leq \beta |Q|$

for arbitrarily long open-world data streams S.

---

12 Future Work

Breaking Conjecture 2 requires automating V_θ without humans. Another open direction is a principled pruning policy that minimises regret while keeping κ_{rebuild} finite.

A Limitations, Testability & Anticipated Objections

A.1 Degrees-of-Freedom That Threaten Falsifiability

Control knob	Risk if left unrestricted
`validation_preference` (accuracy↔speed)	Post-hoc tuning can guarantee perfect scores by cranking friction to ∞.
`confidence_threshold`	Dropping the bar lets the system refuse hard queries instead of failing.
Human override	Oracle patches hide systematic weaknesses; variance unbounded.
Evolving memory store	If the FAISS/Parquet index grows during test, the model can memorise the benchmark.

Unless these knobs are frozen, the framework is empirically unfalsifiable.

A.2 Benchmark Protocol

Because precedent management is the model's hallmark, a static snapshot is not enough; a prequential (streaming) phase measures learning velocity under a cost cap.

Phase	Frozen	Allowed	Metric
Phase 1: Cold-start snapshot	Validation policy confidence & escalation rules initial memory M_0	No index writes	Accuracy$_0$, latency$_0$ at cost$= 0$
Phase 2: Streaming (N batches)	Policy still fixed	System may add / merge precedents; every write logged and charged (bytes + rebuild + human-minutes, $\kappa_{\text{human}}=0$ if fully automated)	Area under curve: error vs. cumulative friction cost

Rules.

1. **Cost ledger:** each write incurs

$$\text{WriteCost}_t \;=\; \sum_i \kappa_i\, c_{i,t}, \qquad \text{unused terms may take } \kappa_i = 0.$$

2. **Delayed grading:** batch k is scored before the system sees batch $k+1$.

3. **Full disclosure:** publish the write log, cost ledger and final memory snapshot so any team can replay the run.

Phase 1 provides a reproducible baseline; Phase 2 tests what is novel: when to file a new precedent, when to reuse, and how fast error falls given a finite friction budget.

A.3 Comparative Landscape

Axis	End-to-End Brain	Neuro-symbolic	Courtroom Model
Core ambition	Single monolith handles everything	NN + logic engine	LLM generates; external friction validates
Strength	No human loop	Explainability, deductive power	Halts hallucination without retraining
Weakness	Brittle world-model, hallucination	Heavy logic engineering	Testability collapses if knobs are free
Ideal use case	Simulated-world agents	Formal maths, compliance	High-risk domains where trust outranks latency

A.4 Residual Limitations

Even under a frozen policy:

- Long-tail error modes can slip through if the cost budget forces approximate checks.

- The framework gives no universal recipe for setting `validation_preference`; domain tuning remains open.

- Human reviewers inject bias; mitigation (diverse panels, blind review) is future work.

A.5 Anticipated Objections

- *"Show us the leaderboard!"* The conjecture precedes the benchmark. Until someone beats or disproves Conjecture 1, a public ranking would optimise for the wrong signal (grading leniency, not epistemic risk).

- *"Nothing fundamentally new."* Yet the missing piece in public discourse is how these parts inter-lock at run time. The Courtroom meta-configuration specifies that interface for the first time. The ingredients exist; what is novel is (i) treating both the friction dial θ and the retention dial τ as first-class, measurable hyper-parameters, and (ii) publishing a meta-configuration file so any lab can replay, ablate, or retune the system.

- *"Without a public benchmark the Courtroom Model is just arm-chair philosophy."* A benchmark that mixes automated checks with open-world human escalation is inherently non-canonical: reviewers differ in domain expertise, risk tolerance, and even sleep. Two laboratories can "reproduce" the same run-log yet obtain different adjudications simply because their reviewers notice different corner-cases. The resulting variance swamps any signal from the wrapper itself. In other words, the benchmark would be measuring humans, not the architecture. Publishing a noisy leaderboard now would mislead the community into optimising for the wrong axis: humour the graders rather than reduce epistemic risk.

- *"Just freeze the humans, then!"* Freezing humans defeats the point. The Courtroom cycle exists precisely because some domains contain irreducible judgment calls (e.g., fraud indicators in SEC filings). Replacing humans with a static rubric re-introduces brittle heuristics— the very failure mode we are trying to avoid.

- *"But without numbers you cannot prove progress."* The claim is architectural, not empirical: CLAIM A: persistent precedent + adaptive friction is necessary for bounded-error knowledge bases in an open world; CLAIM B: current "bigger-is-better" pipelines violate CLAIM A. Both claims are falsifiable: show a practical system that achieves low hallucination without any external precedent store or escalation dial, and the Courtroom premise collapses. Until such a counter-example exists, producing another decimal improvement on the latest benchmark leaderboard is beside the point.

- *"Isn't this just a dressed-up retrieval-augmented loop?"* Retrieval-augmented generation (RAG) treats the store as a context-window scoped evidence buffer flushed between prompts. The Courtroom store is jurisprudential—each write becomes precedent and must pay rent (friction) forever. That difference, though conceptually small, flips the optimisation target from "accuracy at $k = 1$" to "prequential regret under a cost budget." No existing RAG benchmark measures that regime.

- *"Validator still needs humans—won't scale."* Baseline validator uses humans today; our conjecture invites automated judges that meet the same friction budget. Section A.2 logs $\kappa_{\text{human}} > 0$ until automated checkers emerge.

- *"Unfalsifiable if the dial is tuned post-hoc."* The evaluation protocol (Section A.2) freezes the validation policy before scoring and logs every write, so any attempt to retune after seeing the test set is detectable.

- *"Memory blow-up defeats latency budgets."* Section 11.5 already prices FAISS rebuild time. Setting κ_{rebuild} higher turns slow writes into economic friction that discourages runaway growth.

Acknowledgments

of OpenAI ChatGPT (model o3, June 2025). All ideas, experiments, and final text were reviewed, verified, and are the sole responsibility of the author.

References

Marcus, Gary (2020). "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence". In: *arXiv preprint.* eprint: arXiv:2002.06177. URL: https://arxiv.org/abs/2002.06177.

Shinn, Noah, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao (2023). "Reflexion: Language Agents with Verbal Reinforcement Learning". In: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 8634–8652. eprint: arXiv:2303.11366. URL: https://arxiv.org/abs/2303.11366.

Lewis, Patrick et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).* eprint: arXiv:2005.11401. URL: https://arxiv.org/abs/2005.11401.

Fan, Wenqi et al. (2024). "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models". In: *arXiv preprint.* eprint: arXiv:2405.06211. URL: https://arxiv.org/abs/2405.06211.

Asai, Akari, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi (2024). "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection". In: *Proceedings of the 12th International Conference on Learning Representations (ICLR).* Originally posted as arXiv:2310.11511. URL: https://openreview.net/pdf?id=hSyW5go0v8.

Shao, Zhihong, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen (2023). "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy". In: *Findings of ACL.* URL: https://arxiv.org/abs/2305.15294.

Qian, Hongjin et al. (2024). "MemoRAG: Boosting Long Context Processing with Global Memory-Enhanced Retrieval Augmentation". In: *arXiv preprint.* URL: https://arxiv.org/abs/2409.05591.

Park, Sungjune, Heehwan Kim, Haehyun Cho, and Daeseon Choi (2024). *Judgment-of-Thought Prompting: A Courtroom-Inspired Framework for Binary Logical Reasoning with Large Language Models.* eprint: 2409.16635. URL: https://arxiv.org/abs/2409.16635.

He, Bolei et al. (2024). "Retrieving, Rethinking and Revising: The Chain-of-Verification Can Improve Retrieval-Augmented Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2024.* Earlier pre-print available at arXiv:2410.05801, pp. 10371–10393. URL: https://aclanthology.org/2024.findings-emnlp.607.pdf.

Bai, Yuntao et al. (2022). "Constitutional AI: Harmlessness from AI Feedback". In: *arXiv preprint.* eprint: arXiv:2212.08073. URL: https://arxiv.org/abs/2212.08073.

Chen, Justin, Swarnadeep Saha, and Mohit Bansal (2024). "ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085. DOI: `10.18653/v1/2024.acl-long.381`. eprint: `arXiv:2309.13007`. URL: `https://aclanthology.org/2024.acl-long.381`.

Madaan, Aman et al. (2023). "Self-Refine: Iterative Refinement with Self-Feedback". In: *arXiv preprint*. eprint: `arXiv:2303.17651`. URL: `https://arxiv.org/abs/2303.17651`.

Wang, Ruida et al. (2025). "MA-LoT: Multi-Agent Lean-based Long Chain-of-Thought Reasoning Enhances Formal Theorem Proving". In: *arXiv preprint*. eprint: `arXiv:2503.03205`. URL: `https://arxiv.org/abs/2503.03205`.

LeCun, Yann (Apr. 2023). *Unpopular Opinion about AR-LLMs (slide 14, Santa Fe Institute talk)*. `https://www.slideshare.net/slideshow/yann-lecun-20230424-santa-fe-institute-pdf/269726578`.

Silver, David et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587, pp. 484–489. DOI: `10.1038/nature16961`. URL: `https://doi.org/10.1038/nature16961`.

Yao, Shunyu et al. (2023). "Tree of Thoughts: Deliberate Problem Solving with Large Language Models". In: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. eprint: `arXiv:2305.10601`. URL: `https://arxiv.org/abs/2305.10601`.

mathlib Community (2020). "The Lean Mathematical Library". In: *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP '20)*. DOI: `10.1145/3372885.3373824`. URL: `https://arxiv.org/abs/1910.09336`.

Dawid, Anna and Yann LeCun (2024). "Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence". In: *Journal of Statistical Mechanics: Theory and Experiment*, p. 104011. DOI: `10.1088/1742-5468/ad292b`. URL: `https://arxiv.org/abs/2306.02572`.

Chen, Fan, Dylan J. Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu (2024). "Assouad, Fano, and Le Cam with Interaction: A Unifying Lower Bound Framework and Characterization for Bandit Learnability". In: *arXiv preprint*. eprint: `arXiv:2410.05117`. URL: `https://arxiv.org/abs/2410.05117`.