Bits Beat Tokens: A Regret Rate Distortion Theory for Large Language Model Agents

Wanrong Yang[1]* **Lingfang Li**[1]* **Dominik Wojtczak**[1]
Yihang Zhou[2] **Danli Shi**[3] **Yalin Zheng**[4]†

[1]School of Computer Science and Informatics, University of Liverpool
[2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3]School of Optometry, The Hong Kong Polytechnic University
[4]Department of Eye and Vision Sciences, University of Liverpool

Abstract

A language model agent chooses actions from a context built by retrieval, summaries, prompt templates, and memory. Yet token counts do not define how much task information this context must preserve to support a given decision quality. The structural problem is to give context adequacy a decision-based budget that separates information present in the prompt from information used by the decoder. We formalise this problem through the Markov chain $H \to C \to A$ and a regret rate distortion function $\mathcal{R}_\ell(\varepsilon)$, using Bayes regret to measure the value lost by an action and treating actions with equal value alike. Classical data processing gives a necessary budget at each context cut that adds no fresh information, and a random recommendation with a copy decoder attains this budget for uniform zero-one decisions, requiring about 3.14 bits for 16 tools at regret 0.1. The empirical floor check audits the full measurement pipeline, while the observed margins and clustering describe tested agent behaviour. Across 441 balanced, single-step, 16-way tool-selection cells from three models and three benchmarks, the median action-side margin above the exact frontier is 1.07 bits, 97.5% of 629 matched-rate pairs have overlapping regret confidence intervals, and input tokens per action-side bit reach about 1.7×10^6 near saturation. These results make context adequacy a quantitative design problem, separating a necessary mutual-information budget from observed information use and token count, with an exact achievable frontier for uniform zero-one decisions under stochastic context encoding.

1 Introduction

Modern large language model (LLM) agents are built around in-context decisions [31]. The history of a task, including prior observations, retrieved documents, tool outputs, and earlier actions, is compressed into a finite prompt, and the model emits an action such as a tool call [34, 39], a code patch [9, 19], or a structured answer [49, 54]. What practitioners actually control is rarely the model itself. The lever is the upstream compression (prompt templates, retrieval policies, summarisation, scratchpads, and key-value caches [12]), together with the wrapper that orchestrates them, an artefact that the deployment community now calls the agent's harness [2, 3]. As context windows have stretched into the million-token regime [13, 25], the harness has become load-bearing. A recent evaluation survey [56] warns that current benchmarks conflate the backbone model with its surrounding scaffold, and Anthropic's deployment guidance [3] reports that long-running agents fail or succeed primarily as a function of how their context is managed across tasks for many hours or even several days.

*Equal contribution. †Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This raises a basic and largely unanswered question: how much decision-relevant information about the task history must the context preserve before an in-context agent can act with low regret? The question is operational. Engineers tune retrieval depth, summary length, and prompt layout under uncertainty about whether the cause of a failure is the model, the wrapper, or an information bottleneck strict enough that no model or wrapper choice can repair it. We want an answer that does not depend on the model, decoder, or scaffold to this question, expressed in bits rather than tokens, so that a practitioner can compare a target regret with the information preserved by a fixed context pipeline. One empirical pattern has solidified across long-context studies but needs a common decision-based budget. Tokens are not interchangeable with information. Relevant evidence is "lost in the middle" of long contexts [5, 24], LLMs are easily distracted by irrelevant material in their input [43], apparently lossless prompt compression silently degrades downstream behaviour [17, 18], and stress-test benchmarks reveal that nominal context length exceeds usable length by a wide margin [15, 58]. Engineers debug these failures case by case. The harness community treats the wrapper as an empirical optimisation target and has produced a large design space of scaffolds, but we are not aware of a model-agnostic theory that links these effects to a regret budget while keeping prompt information separate from the information used by the model.

We approach the question through rate-distortion theory. Modelling the agent as a Markov chain *history* $\to$ *context* $\to$ *action* and treating Bayes regret (the value gap between the chosen and the optimal action) as the distortion measure, we define a regret rate distortion (RRD) function $\mathcal{R}_\ell(\varepsilon)$ and obtain its converse as a one-line consequence of the data processing inequality (DPI) [10], where a context whose capacity falls short of $\mathcal{R}_\ell(\varepsilon)$ cannot drive expected regret below ε, regardless of model scale, decoder, or scaffold. To put the bound in contact with practice we instantiate it on a controlled 16-way single-step tool-selection task across three open-source LLMs (Qwen3-32B [35], Gemma3-27B [14], GLM-4-32B [46]) and three public benchmarks (API-BANK [23], τ-bench [55], ToolBench [34]), sweeping catalogue size, insertion position, description compression, and a chain-of-thought (CoT) [50] toggle. The gold tool index $X = f(H)$ is a function of the history. Thus the true action-side rate satisfies $I(X, A) \le I(H, A) \le I(H, C)$. We estimate $I(X, A)$ to study information use. This lower bound cannot by itself identify missing context information. Figure 2 previews the margin distribution.

Contributions. (i) We formalise the context adequacy problem for LLM agents through history, context, and action, with Bayes regret as distortion (Section 3). This is the core contribution. It places position, compression, and tool overload in one decision-based framework and treats actions with equal value alike. (ii) We use classical rate distortion and data processing to derive a necessary bit budget (Theorem 5, Corollary 6). For uniform zero-one decisions the budget is exact, with a simple context construction and copy decoder (Appendix 3). We also state finite-problem achievability and the link to rational inattention with proofs. (iii) On 441 balanced single-step cells, Experiment 1 audits measurement, Experiment 2 measures matched-rate regret clustering, and Experiment 3 measures CoT shifts (Section 5). We also report input tokens per action-side bit. These experiments cover 16-way tool selection and do not establish a diagnosis for full agent trajectories.

2 Related Work

RRD formalises context adequacy for deployed agents with classical information theory. The contribution lies in the problem and its measurement, rather than a new information inequality. *Information-theoretic decision making.* Mutual information (MI) has long driven exploration in bandits and reinforcement learning [8, 26, 36, 37]. Rate-distortion has been used to pick what a capacity-limited learner should fit [4], and probably approximately correct (PAC)-Bayes plus MI bounds control generalisation [52]. RRD borrows the classical machinery [10, 40, 41, 47] but shifts the bound from what an algorithm learns to what a deployed prompt transmits. *Theory of in-context learning.* A Bayesian thread treats in-context learning as approximate posterior inference [11, 16, 51, 59]. Reasoning scaffolds [28, 44, 50, 54] act as decoders that improve achievability without lowering the floor. We measure their effect on one CoT slice in Section 5. *Long-context use and context engineering.* "Lost-in-the-middle" [5, 24, 38] and usable-context benchmarks [7, 15, 25, 58] show that nominal context overstates usable context. Deployed knobs [retrieval-augmented generation (RAG) [12, 21, 30], prompt compression [17, 18], and context engineering [2, 57]] are exercised by recent benchmarks [23, 34, 53, 55, 60]. Each studies one knob, while RRD groups them under a single necessary condition that any deployed prompt must satisfy. *Agent harness research.* The

wrapper around the model is now load-bearing for real deployments [3, 57], and recent surveys warn that benchmarks conflate backbone with scaffold [56]. Harness designers chase *achievability*. Our converse caps it from below, and by Corollary 6 a composed harness inherits the *tightest* stage's rate, so a sub-floor retrieval stage cannot be rescued by later planners or verifiers. *MI estimation.* Miller-Madow (MM) de-biases plug-in entropy on finite alphabets [29, 32]. High-dimensional actions are covered by mutual information neural estimation (MINE) [6], the InfoNCE contrastive objective [48], and variational MI [33]. Combining MM-corrected entropies is not exact debiasing of plug-in MI. Appendix 4 characterises the residual and shows our Experiment 1 audit at the point estimate and after clipping. *Rational inattention and achievability.* Sims [45] studies decisions with an MI cost. Matějka and McKay [27] give the optimal choice rule with an endogenous action prior. RRD is the inverse budget form of this value problem. Appendix 3 states the relation and adapts the strong functional representation lemma of Li and El Gamal [22] to context coding with shared randomness.

3 In-Context Agents as Information Channels

The starting observation of this paper is that a language-model agent never acts on the raw task history. Between the environment and the emitted action sits a context that has been assembled by retrieval, summarisation, prompt templates, and memory, and every quantity the decoder can use is mediated through that context. We model a single in-context decision step as a three-letter Markov chain whose middle letter, the context, is the only information channel through which the history reaches the action. The remainder of this section fixes the notation we need to make this picture precise, records the measurability conditions under which the chain is well-posed, and defines the loss, regret, against which we will later constrain the channel.

We adopt the convention that capital letters denote random variables and lowercase letters denote realisations, and we write $I(\cdot,\cdot)$ for MI, always evaluated in bits ($\log_2$) unless explicitly stated otherwise. Let $H \sim P_H$ be the *task history*, taking values in a measurable space $\mathcal{H}$ that contains everything potentially relevant to the current decision, including prior observations, retrieved documents, tool outputs, memory states, and previous actions. Let $C \in \mathcal{C}$ be the *context*, the prompt that the decoder actually reads, and let $A \in \mathcal{A}$ be the emitted *action*. The agent factorises as a (possibly stochastic) compression $C = \psi(H)$ and a decoder $A \sim \pi(\cdot \mid C)$, which together induce the Markov chain $H \to C \to A$ depicted in Figure 1, so that $I(H, A \mid C) = 0$. The action space $\mathcal{A}$ may be finite (a multiple-choice answer), countably infinite (the set of strings $\Sigma^\star$ over a token alphabet Σ), or continuous ($\mathbb{R}^d$ for control). On the experimental task of Section 5, the decision label is the gold tool index $X = f(H) \in \{1, \ldots, 16\}$. The history also includes query and catalogue details. The zero-one decision problem reduces to X, but H need not equal X. We write ε for the regret budget that appears in the theory and $\widehat{\varepsilon}$ for its empirical estimate from finite samples, and we use $\widehat{I}_{\mathrm{MM}}$ for the MM plug-in estimate of $I(X, A)$ that we deploy in the experiments.

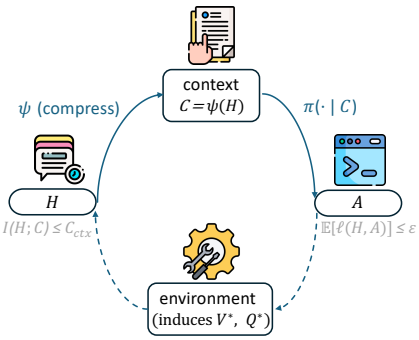

Figure 1. In-context agent as the Markov chain $H \to C \to A$. The compression ψ folds the history H into the context C, the decoder $\pi(\cdot \mid C)$ emits the action A, and the DPI yields $I(H, A) \le I(H, C)$, which is the basis of Theorem 5.

The main text uses finite alphabets for exposition, but each statement extends to the non-finite case under the standard-Borel regularity conditions developed in Appendix 2, in which $\mathcal{H}, \mathcal{C}, \mathcal{A}$ are standard Borel, $\ell \in \{g \mid g \text{ maps } \mathcal{H} \times \mathcal{A} \text{ to } \mathbb{R}_{\ge 0}\}$ is jointly measurable, and $\ell(H, A)$ is integrable. Under those conditions the DPI proof of Theorem 5 applies verbatim, so the reader can treat the finite presentation in the main text as a notational convenience. The chain $H \to C \to A$ likewise covers a single in-context decision step, where C is a deterministic-or-stochastic compression of H and A depends on H only through C. Multi-stage scaffolds such as ReAct, Reflexion, and self-RAG, which issue new reads of the environment between stages, are covered by the staged extension $H \to C_1 \to \cdots \to C_k \to A$ of Corollary 6, provided each C_{i+1} is computed from C_i alone. Agent loops that splice in fresh observations at each step are handled by re-applying the converse decision by decision rather than by collapsing the loop into a single chain.

The experiments vary insertion position, pool size, description compression, and CoT. The first three change the context construction. CoT changes the decoder, with a small instruction change to the

prompt. All four change the induced history-to-action channel. We estimate the action-side rate $\widehat{I}_{\mathrm{MM}}(X, A)$ because its alphabet is small. We now turn to the loss that the channel is asked to support. We are interested in a distortion measure that is faithful to generative agents, whose surface forms are not unique. Two strings that induce the same downstream value should receive the same penalty.

Definition 1 (Regret distortion)**.** Let $Q^\star(h, a)$ denote the Bayes-optimal value of taking action a after history h, and let $V^\star(h) \triangleq \sup_{a \in \mathcal{A}} Q^\star(h, a)$ denote the Bayes-optimal value at h. The *regret distortion* is

$$\ell(h, a) \;\triangleq\; V^\star(h) \;-\; Q^\star(h, a) \;\geq\; 0.$$

The regret distortion is zero whenever a is optimal under h, non-negative for every kernel $\kappa(a \mid h)$, and invariant on the $Q^\star$*-equivalence classes* defined by $a \sim a' \iff Q^\star(\cdot, a) = Q^\star(\cdot, a')$, so that two surface forms that induce the same value are charged the same distortion. We assume $\ell(H, A)$ is integrable, which holds automatically in finite-horizon or discounted decision problems with bounded rewards. The information that the prompt actually carries about the history is $I(H, C)$, and the largest such information that a deployment can realise under its operational constraints is the *context capacity*

$$C_{\mathrm{ctx}} \;\triangleq\; \sup_{\psi} I(H, C),$$

where the supremum is over context maps allowed by the deployment budget under the fixed source distribution P_H. Each realised map satisfies $I(H, C) \leq C_{\mathrm{ctx}}$. This quantity characterises the maximum history-context mutual information available within the admissible class of context constructions, rather than the physical code length of a particular prompt. A model's ability to use the resulting information is a separate decoder property.

Information present and information used. Syntactic context information $I(X, C)$ measures what the actual prompt preserves about the target. The action-side rate $I(X, A)$ measures what reaches the output under the chosen decoder. Data processing gives $I(X, A) \leq I(X, C) \leq I(H, C) \leq C_{\mathrm{ctx}}$. The finite-sample estimate $\widehat{I}_{\mathrm{MM}}$ can lie outside this range through estimation error. If the query determines X and is fully retained, then $I(X, C) = H(X) = 4$ bits in our balanced task. A retained catalogue also identifies X in this controlled construction because its only core tool is the gold tool (Appendix 5). Under either verified recovery condition, $4 - \widehat{I}_{\mathrm{MM}}(X, A)$ estimates context-to-action information loss. It does not measure general $I(H, C)$ or C_{ctx}. We do not assume the recovery condition for capped prompts. Low action-side rate alone cannot distinguish missing prompt information from available information that the decoder does not use.

4 The Regret Rate Distortion Theory

The regret rate distortion function maps a target regret to the least information a decision rule must retain. Classical data processing then places this budget at the context cut. We give the exact uniform zero-one case below. Appendix 3 supplies context constructions and general finite-problem results.

Definition 2 (Regret rate distortion function)**.** For any $\varepsilon \geq 0$, the *regret rate distortion function* of an in-context decision problem with joint law $P_{H, V^\star, Q^\star}$ is

$$\mathcal{R}_\ell(\varepsilon) \;\triangleq\; \inf_{\kappa(a|h) \text{ with } \mathbb{E}[\ell(H,A)] \leq \varepsilon} I(H, A), \tag{1}$$

where $A \sim \kappa(\cdot \mid H)$ and the infimum is over all stochastic decision kernels from $\mathcal{H}$ to $\mathcal{A}$.

The quantity $\mathcal{R}_\ell(\varepsilon)$ depends only on the joint law of the decision problem, not on the model, decoder, or scaffold, and is unchanged by any reparametrisation of the agent. Three structural properties follow from convex-analytic arguments on (1), and we record them before turning to the converse, since the saturation clause is also needed in the closed-form derivation below.

Proposition 3 (Basic properties)**.** *The function $\mathcal{R}_\ell$ on $\mathbb{R}_{\geq 0}$ is non-increasing and convex in ε, and saturates to $\mathcal{R}_\ell(\varepsilon) = 0$ for every $\varepsilon \geq \varepsilon^\circ$ when a best constant action exists, where $\varepsilon^\circ \triangleq \inf_{a \in \mathcal{A}} \mathbb{E}[\ell(H, a)]$ is the expected regret of the best constant action.*

Proof. Enlarging the regret budget enlarges the feasible set of decision kernels, so the infimum is non-increasing in ε. A mixture of feasible kernels is feasible by the convexity of expectation, and

$I(H, A)$ is itself convex in $\kappa(a \mid h)$ for fixed P_H [10], so the infimum of a convex function over a convex set is convex in the constraint level. Saturation follows from the deterministic constant kernel κ° with $A \equiv a^\circ$ for $a^\circ \in \arg\min_a \mathbb{E}[\ell(H, a)]$, which has $I(H, A) = 0$ and $\mathbb{E}[\ell(H, A)] = \varepsilon^\circ$. □

We now bring in the channel side. The Markov structure $H \to C \to A$ from Section 3 produces a DPI on each transition, and chaining those inequalities with Definition 2 gives the main converse. We first isolate the context-side step as a lemma so that the theorem itself reduces to a single line.

Lemma 4 (Context-side DPI)**.** *Under the Markov chain $H \to C \to A$ and any decoder π, $I(H, A) \le I(H, C) \le C_{\mathrm{ctx}}$.*

Proof. The DPI [10] applied to $H \to C \to A$ yields $I(H, A) \le I(H, C)$, and the second inequality is the definition of C_{ctx}. □

Combining Definition 2 with Lemma 4 gives the main converse. Any kernel $P(A \mid H)$ that the agent realises is feasible in (1), and feasibility forces its MI with the history to be at least $\mathcal{R}_\ell(\varepsilon)$, while data processing forces that same MI to be at most C_{ctx}. Stringing the two inequalities together gives the bound we want.

Theorem 5 (Context-capacity regret rate distortion converse)**.** *Let $H \to C \to A$ be an in-context decision process with $I(H, C) \le C_{\mathrm{ctx}}$. If the agent achieves expected regret at most ε, that is, $\mathbb{E}[\ell(H, A)] \le \varepsilon$, then*

$$C_{\mathrm{ctx}} \;\ge\; \mathcal{R}_\ell(\varepsilon). \tag{2}$$

Equivalently, $C_{\mathrm{ctx}} < \mathcal{R}_\ell(\varepsilon)$ implies $\mathbb{E}[\ell(H, A)] > \varepsilon$.

Proof. Since the agent achieves $\mathbb{E}[\ell(H, A)] \le \varepsilon$, the induced kernel $P(A \mid H)$ is feasible in the optimisation (1), so $\mathcal{R}_\ell(\varepsilon) \le I(H, A)$. Combining with Lemma 4, $\mathcal{R}_\ell(\varepsilon) \le I(H, A) \le I(H, C) \le C_{\mathrm{ctx}}$. □

The theorem gives a necessary condition. It does not promise that a given LLM will use the available information well. In the uniform M-way setting, a context that carries a random tool recommendation attains the bound with a copy decoder (Proposition 13). Thus the frontier is exact for this decoder class. A positive empirical margin measures distance from that frontier at the achieved regret. It does not alone locate the source of failure.

The same necessary budget applies at every stage of a pipeline that adds no fresh information.

Corollary 6 (Multi-stage converse)**.** *If the context is constructed by a pipeline $H \to C_1 \to C_2 \to \cdots \to C_k \to A$ in which each C_{i+1} is computed from C_i alone, and $\mathbb{E}[\ell(H, A)] \le \varepsilon$, then $\min_i I(H, C_i) \ge \mathcal{R}_\ell(\varepsilon)$.*

Proof. Apply the DPI stage by stage and combine with Theorem 5. Details are given in Appendix 2. □

Remark 7 (Recovering classical rate distortion)**.** Choosing $\mathcal{A} = \mathcal{H}$ and $\ell(h, a) = d(h, a)$ for a metric d recovers classical (reconstruction) rate distortion [41], and choosing $\ell(h, a) = \mathbf{1}[a \ne f(h)]$ for a labelling f recovers the Hamming converse. Regret rate distortion subsumes both as special cases.

For uniform M-way zero-one decisions, the classical Hamming rate distortion function gives an exact budget [10].

Theorem 8 (M-ary Hamming floor)**.** *For integer $M \ge 2$, let $H = X \sim \mathrm{Unif}\{1, \dots, M\}$ and $A \in \{1, \dots, M\}$ with optimal action $A = X$, and let $\ell(X, A) \triangleq \mathbf{1}[A \ne X]$, so that $\mathbb{E}[\ell(X, A)] = \mathbb{P}(A \ne X)$. Then for every $0 \le \varepsilon \le (M - 1)/M$,*

$$\mathcal{R}_\ell(\varepsilon) \;=\; \log_2 M \;-\; h_2(\varepsilon) \;-\; \varepsilon \log_2(M - 1), \tag{3}$$

and $\mathcal{R}_\ell(\varepsilon) = 0$ for $\varepsilon \ge (M - 1)/M$, where $h_2(p) \triangleq -p \log_2 p - (1 - p) \log_2(1 - p)$ is the binary entropy.

Proof. Write $P_e \triangleq \mathbb{P}(A \neq X)$ for any feasible kernel and decompose the MI as $I(X,A) = H(X) - H(X \mid A)$. Since X is uniform on an M-element alphabet, $H(X) = \log_2 M$, so minimising $I(X,A)$ subject to $P_e \leq \varepsilon$ is equivalent to maximising $H(X \mid A)$ under the same constraint. Conditional on $A = a$, the distribution of X places mass $1 - q_a$ on the value a (the event $X = a$) and total mass q_a on the remaining $M - 1$ atoms, where $q_a \triangleq \mathbb{P}(X \neq a \mid A = a)$. Splitting $H(X \mid A = a)$ into the binary uncertainty between $\{a\}$ and its complement and the residual uncertainty among the $M - 1$ off-diagonal atoms gives

$$H(X \mid A = a) \;=\; h_2(q_a) \;+\; q_a\, H\big(X \mid A = a,\, X \neq a\big) \;\leq\; h_2(q_a) \;+\; q_a \log_2(M-1),$$

with equality precisely when $X \mid (A = a,\, X \neq a)$ is uniform on the $M - 1$ off-diagonal atoms. Taking the expectation over A and applying Jensen's inequality with the concavity of h_2 on $[0,1]$,

$$H(X \mid A) \;\leq\; \mathbb{E}[h_2(q_A)] + \mathbb{E}[q_A]\log_2(M-1) \;\leq\; h_2\big(\mathbb{E}[q_A]\big) + \mathbb{E}[q_A]\log_2(M-1).$$

By Bayes' rule, $\mathbb{E}[q_A] = \sum_a \mathbb{P}(A = a)\mathbb{P}(X \neq a \mid A = a) = \mathbb{P}(A \neq X) = P_e$. The map $p \mapsto h_2(p) + p\log_2(M-1)$ is strictly increasing on $[0, (M-1)/M]$ and attains its maximum value $\log_2 M$ at $p = (M-1)/M$, so for every feasible kernel with $P_e \leq \varepsilon \leq (M-1)/M$,

$$H(X \mid A) \;\leq\; h_2(\varepsilon) + \varepsilon \log_2(M-1),$$

which gives $I(X,A) \geq \log_2 M - h_2(\varepsilon) - \varepsilon\log_2(M-1)$, and therefore $\mathcal{R}_\ell(\varepsilon) \geq \log_2 M - h_2(\varepsilon) - \varepsilon\log_2(M-1)$.

To match the lower bound, consider the *M-ary symmetric channel*

$$\kappa^\star(a \mid x) \;=\; (1-\varepsilon)\,\mathbf{1}[a = x] \;+\; \frac{\varepsilon}{M-1}\,\mathbf{1}[a \neq x].$$

Under uniform X, the marginal of A is uniform on $\{1,\dots,M\}$, so $\mathbb{P}(A = a) = 1/M$ for every a. The induced posterior is $\mathbb{P}(X = a \mid A = a) = 1 - \varepsilon$ and $\mathbb{P}(X = x \mid A = a) = \varepsilon/(M-1)$ for every $x \neq a$, so $\mathbb{P}(A \neq X) = \varepsilon$ exactly and $H(X \mid A = a) = h_2(\varepsilon) + \varepsilon\log_2(M-1)$ for every a. The two displayed inequalities above hold with equality, so $I(X,A) = \log_2 M - h_2(\varepsilon) - \varepsilon\log_2(M-1)$ and $\kappa^\star$ witnesses the matching upper bound.

For $\varepsilon \geq (M-1)/M$ the saturation clause of Proposition 3 applies, so the input-independent kernel $\kappa^\circ(a \mid x) = 1/M$ achieves $\mathbb{E}[\ell(X,A)] = \mathbb{P}(A \neq X) = (M-1)/M \leq \varepsilon$ with $I(X,A) = 0$. □

Remark 9 (Reduction to the decision label)**.** The Hamming formula also applies when $X = f(H)$ and the zero-one loss depends on H only through X. For any kernel $P(A \mid H)$, data processing gives $I(X,A) \leq I(H,A)$, while the induced joint law of (X,A) preserves the same expected loss. Conversely, any kernel $P(A \mid X)$ can be implemented through $H \to X \to A$, for which $I(H,A) = I(X,A)$. Hence the rate-distortion optimisation over H has the same optimum as that over X. The corresponding floor also remains valid when additional action labels represent errors, since mapping these labels to a fixed core action cannot increase either mutual information or zero-one loss.

Three numerical anchors illustrate the magnitude of the floor. For M=4 and ε=0.1, $\mathcal{R}_\ell(0.1) = 2 - h_2(0.1) - 0.1\log_2 3 \approx 1.37$ bits, and at M=16 (the operating point of Section 5) the same tolerance requires ≈ 3.14 bits, and at M=256 it requires ≈ 6.73 bits. These are hard floors in the sense that any context transmitting less information cannot support expected regret below ε, regardless of model or scaffold, so the bound becomes a deployment diagnostic. Given a target ε, compare the required floor with a separately measured rate. An action-side rate below that target floor rules out the target for the current configuration. It does not identify which part must change. Appendix 3 renders the floors for $M \in \{4, 16, 64, 256\}$ in Figure 5 and reports the numerical values at practitioner-relevant operating points in Table 4, and discusses heterogeneous action costs and non-uniform X.

The general converse does not require $\mathcal{A}$ to be finite. Language agents emit strings $\Sigma^\star$, continuous controls $\mathbb{R}^d$, or structured tool calls, and the regret distortion is well-defined under the stated measurability and integrability assumptions on $\ell(H,A)$, with Theorem 5 holding verbatim. The equivalence-class collapse on actions with identical $Q^\star$ preserves the optimum, so the rate is governed by the number of behaviourally distinct actions, not by $|\Sigma^\star|$ (Appendix 2).

5 Experiments

We study balanced, single-step, 16-way zero-one tool selection ($n = 320$ trials per cell, $H(X) = 4$ bits), with the exact frontier of Remark 9. Catalogue size counts entries, $N \in \{16, 64, 128, 512\}$; gold-entry position $p \in \{\text{BEG}, \text{MID}, \text{END}\}$ denotes beginning, middle, or end. Description codes τ are T0 (verbatim), T1/T2 (first 50%/25% of description characters), and C1 (supplied paraphrase, otherwise T1); tool names are retained.

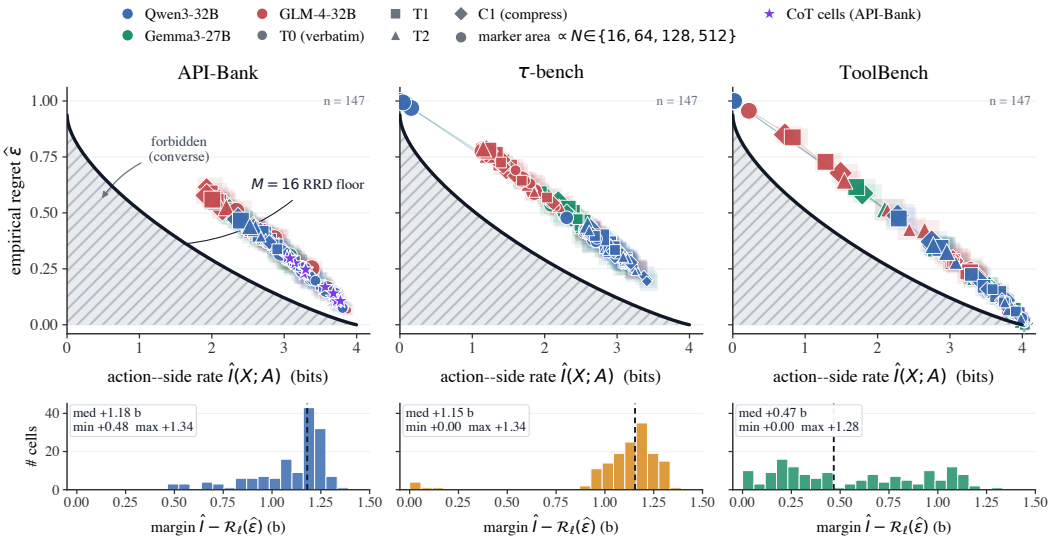

Figure 2. The RRD frontier across three models and three benchmarks. *Top.* per-cell $(\widehat{I}_{\text{MM}}, \widehat{\varepsilon})$ markers with bootstrap confidence interval (CI) rectangles (B=400, 95%). Hue is model, shape is compression τ, marker area is pool size N, and translucent traces connect the T0→T1→T2→C1 sweep at fixed (model, N, MID). The black curve is the analytical M=16 Hamming floor (Equation 3). The hatched region is forbidden by Theorem 5. Stars on the API-Bank panel mark the 12 CoT cells of Figure 3. *Bottom.* per-benchmark histograms of the cell margin $\widehat{I}_{\text{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon})$, with median dashed.

We report three experiments. Experiment 1 audits measurement: balance, parsing, label merging, MM correction, and bootstrap. Every cell has exactly 20 trials for each gold tool. Regret and plug-in MI come from the same confusion matrix, so Fano already guarantees the plug-in floor check. This audit is a prerequisite for using the measurements, rather than independent evidence for the theorem. Experiment 2 measures matched-rate regret clustering. Experiment 3 measures CoT shifts and proximity to the frontier. These empirical patterns are falsifiable and are not consequences of the converse. A deployment comparison instead fixes a target regret in advance and measures the rate separately.

Experiment Setup. We evaluate three open-source LLMs (Qwen3-32B, Gemma3-27B, GLM-4-32B) on three public tool-use benchmarks (API-Bank, τ-bench, ToolBench). From each benchmark we select a fixed M=16 *core pool* of tools and uniformly draw a query mapped to a gold tool $X \in \{1, \ldots, 16\}$. Distractors exclude all core tools, so the catalogue contains only one core identifier. The model emits a tool call, parsed into a label A over the 16 core tools plus two absorbing labels (OTHER_DISTRACTOR and INVALID). The (N, p, τ) grid gives 48 no-CoT settings per (model, benchmark). One CoT setting at (N=128, MID, T0) per pair gives $432 + 9 = 441$ main cells. A separate 12-cell CoT supplement covers the (Qwen3, API-BANK, T0) slice. Each cell uses 320 trials. The analytical M=16 floor is $\mathcal{R}_\ell(\widehat{\varepsilon}) = 4 - h_2(\widehat{\varepsilon}) - \widehat{\varepsilon}\log_2 15$ (Equation 3), with its zero-rate branch for $\widehat{\varepsilon} \geq 15/16$. The 18-valued variant is the weaker test because the two absorbing labels can retain information about X. We therefore report both a 16-valued collapse and a 17-valued merge in Appendix 6. We estimate $\widehat{I}_{\text{MM}}$ from the joint counts and use gold-stratified bootstrap CIs ($B = 400$). The paper analysis and label-collapse audit use this protocol. Ordinary resampling can give different CIs, as documented in Appendix 4. For plug-in counts, mapping absorbing labels to a fixed core tool cannot increase error. Applying Fano to that mapped action and then DPI gives the same lower floor at the original regret. MM correction need not preserve this guarantee. Apparent violations would first require an estimator audit. For $\widehat{\varepsilon} < 15/16$, the smallest margin is $+0.034$ bits before clipping $\widehat{I}_{\text{MM}}$ to 4 bits and 0.000 after clipping. Appendix 4 gives the settings and bias checks. The

	loose. $\lvert\Delta\widehat{I}\rvert \le 0.10$ b				strict. $\lvert\Delta\widehat{I}\rvert \le 0.05$ b			
	n_{pairs}	$n_{\text{CI}\cap}$	overlap	median $\lvert\Delta\widehat{\varepsilon}\rvert$	n_{pairs}	$n_{\text{CI}\cap}$	overlap	median $\lvert\Delta\widehat{\varepsilon}\rvert$
N=16	275	262	95.3%	0.0094	185	178	96.2%	0.0063
N=64	129	126	97.7%	0.0125	72	69	95.8%	0.0094
N=128	117	117	100.0%	0.0156	59	59	100.0%	0.0063
N=512	108	108	100.0%	0.0125	55	55	100.0%	0.0063
all N	629	613	97.5%	0.0125	371	361	97.3%	0.0063

Table 1. Experiment 2: Matched-rate regret clustering. Each row counts unordered no-CoT cell pairs within a fixed (model, bench, N) slice that differ in at least one of {intervention, position}. $n_{\text{CI}\cap}$ counts pairs with overlapping bootstrap regret CIs. Median $\lvert\Delta\widehat{\varepsilon}\rvert$ shrinks monotonically as the rate-matching threshold tightens.

study used about 4,000 A100 graphics processing unit (GPU) hours with a 32,768-token input cap (Appendix 15).

Experiment 1: Measurement audit. Figure 2 reports a margin $\mu = \widehat{I}_{\text{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon})$ with minimum, median, and maximum of 0.000, 1.073, and 1.343 bits. Zero-margin cells occur in the saturated regime where both rate and floor approach zero. Tight cells also occur at the high-information corner on ToolBench. This distribution describes where the tested policies sit relative to the exact frontier. Its sign is an implementation and estimation check. The 16-valued collapse lowers the median margin to 0.949 bits. Full results and scoring rules are in Appendix 6.

Experiment 2: Matched-rate regret clustering. We ask whether similar action-side rates track similar regret. The one-sided converse does not imply this. We enumerate unordered no-CoT cell pairs within each (model, bench, N) slice that differ in at least one of {intervention, position} and report $\lvert\Delta\widehat{I}_{\text{MM}}\rvert$, $\lvert\Delta\widehat{\varepsilon}\rvert$, and bootstrap-CI overlap on $\widehat{\varepsilon}$ in Table 1. Of 629 pairs with $\lvert\Delta\widehat{I}_{\text{MM}}\rvert \le 0.10$ b, 613 (97.5%) share overlapping regret CIs and the median $\lvert\Delta\widehat{\varepsilon}\rvert$ is 0.0125. Tightening the rate threshold to 0.05 b cuts the median to 0.0063 while keeping 97.3% overlap. At a rate of 2 bits, uniform 16-way confusion matrices can have regret from about 0.29 to 1. Thus classification mechanics allow a gap of about 0.7, far larger than the observed median 0.0125. Appendix 7 gives this null comparison. CI overlap is a conservative consistency check rather than a formal equivalence test, so $\lvert\Delta\widehat{\varepsilon}\rvert$ is the more decision-relevant statistic. The per-pair scatter is in Figure 9.

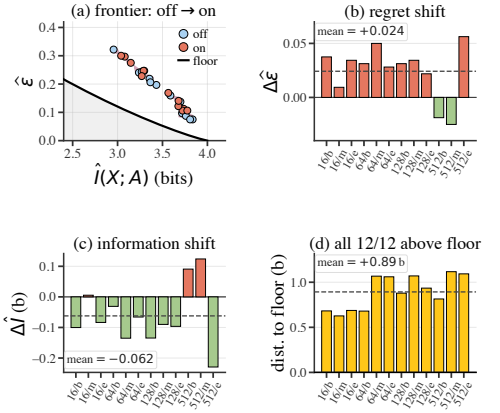

Figure 3. Experiment 3: CoT operating points and measured shifts on 12 cells. Mean regret change is +0.024 and mean action-side rate change is −0.062 bits. The mean margin is 0.89 bits. Staying above the floor is part of the Experiment 1 audit.

Experiment 3: CoT shifts. A decoder that reasons only from C still obeys the DPI. This does not predict how its rate or regret will change. We compare CoT on and off for 12 Qwen3, API-Bank, T0 cells. The mean regret change is +0.024 and the mean action-side rate change is −0.062 bits (Figure 3). These shifts, rather than the absence of floor crossings, are the empirical content of Experiment 3. The mean margin is 0.89 bits. This slice was chosen under the compute budget because its prompts fit the window and retain the full catalogue. We do not infer the same effect on the other benchmarks. Appendix 8 gives the per-cell results. Table 2 reports $\widehat{\varepsilon}$ at (T0, MID) across $N \in \{16, 64, 128, 512\}$, the ($N$=128, MID) compression sweep, and the slope of $\widehat{\varepsilon}$ on $\log_2 N$. Three patterns recur across all three models. (i) The fitted slope of regret on $\log_2 N$ is positive, ranging from +0.045 to +0.195 per doubling. (ii) ToolBench saturates at $\widehat{\varepsilon}$=1.0 at N=512 once the verbatim T0 prompt exceeds the 32,768-token window and the parser returns mostly INVALID. (iii) The compression interventions T1, T2, C1 typically *worsen* regret relative to T0 even though token counts shrink. Compression has two regimes. In uncapped API-Bank cells, it lowers the action-side rate and raises regret. When T0 hits the input cap, compression can instead raise the rate and lower regret by keeping more of the catalogue in the prompt. These are measured changes, rather than directions implied by the converse (Appendix 11).

Model	Benchmark	$\widehat{\varepsilon}$ at (T0, MID, N)				$\widehat{\varepsilon}$ at (N=128, MID, τ)			Slope $\partial\widehat{\varepsilon}/\partial \log_2 N$
		N=16	N=64	N=128	N=512	T1	T2	C1	
Qwen3-32B	API-BANK	0.10	0.20	0.24	0.32	0.42	0.42	0.38	+0.045
	τ-bench	0.31	0.38	0.42	0.99	0.42	0.42	0.42	+0.132
	ToolBench	0.02	0.11	0.16	1.00	0.17	0.15	0.15	+0.191
Gemma3-27B	API-BANK	0.14	0.29	0.39	0.43	0.55	0.53	0.50	+0.059
	τ-bench	0.40	0.54	0.62	1.00	0.59	0.52	0.57	+0.118
	ToolBench	0.01	0.12	0.25	1.00	0.35	0.21	0.35	+0.195
GLM-4-32B	API-BANK	0.12	0.26	0.37	0.51	0.55	0.54	0.57	+0.080
	τ-bench	0.53	0.69	0.78	0.78	0.78	0.69	0.76	+0.051
	ToolBench	0.02	0.10	0.25	1.00	0.29	0.42	0.29	+0.194

Table 2. Headline empirical results across three models and three benchmarks. *(Left)* regret $\widehat{\varepsilon}$ at (T0, MID) for $N \in \{16, 64, 128, 512\}$. ToolBench saturates at $\widehat{\varepsilon}$=1.0 once the verbatim T0 prompt exceeds the 32,768-token window. *(Middle)* regret under compressions {T1, T2, C1} at (N=128, MID). *(Right)* slope of $\widehat{\varepsilon}$ on $\log_2 N$ at (T0, MID) from a least-squares fit across the four N values, positive in all 9 (model, bench) pairs.

Input tokens and action-side bits. Figure 4 compares input token counts with $\widehat{I}_{\mathrm{MM}}(X, A)$. The ratio reaches about 5,800 tokens per bit on API-Bank at $N = 512$, about 4×10^5 on τ-bench, and about 1.7×10^6 on ToolBench near saturation. The median on this slice is about 1,787 tokens per action-side bit. The saturated badge marks $\widehat{I}_{\mathrm{MM}} < 0.05$ bits, where a ratio with a near-zero denominator is sensitive to estimation noise. These numbers measure information reaching the action. They do not measure prompt capacity. A low action-side rate can coexist with a fully informative prompt, as the retained-context audit in Appendix 5 makes explicit.

6 Discussion

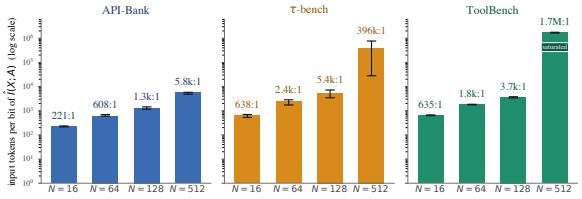

Figure 4. Input tokens per action-side bit on the (T0, MID, no-CoT) slice. Per-benchmark towers of input tokens per bit of $\widehat{I}_{\mathrm{MM}}(X, A)$ on a shared log y-axis. Bar height is cross-model median, and whiskers are min/max. The *saturated* badge marks cells where every model's $\widehat{I}_{\mathrm{MM}}$ falls below the 0.05 b numerical-noise floor.

Context budgets and diagnostic scope. The main contribution is to ask what decision-relevant information a context must preserve for a target regret. Regret distortion respects value-equivalent actions, focusing the budget on downstream decisions rather than history reconstruction. The context cut separates prompt information from policy use. Data processing gives a common budget for position, compression, retrieval, and tool overload. Their effects remain empirical: shorter prompts must preserve information needed for the chosen loss. A context rate below the target floor rules out that target for every decoder at the cut; a below-floor action-side rate rules it out for the current configuration. Both require uncertainty and bias checks. An action-side deficit cannot distinguish missing from unused information, and an above-floor action-side rate does not certify prompt adequacy at lower regret. Our median margin of 1.07 bits is distance from an attainable frontier at the achieved regret. Context audits are needed before assigning it to decoder failure, especially for capped inputs with unverified recovery (Appendix 5).

Inference-time scaling. CoT, self-consistency, and verification remain decoders of C when they add no external information. They may improve information use, but their action-side rate remains bounded by $I(H, C)$. Extra computation can change an operating point without changing the context budget. Our CoT slice measures one such change, with small mean shifts and substantial reasoning cost (Appendix 8); it does not establish effects for other reasoning methods. Verification using an external source changes the available information and requires a new context cut.

Practical workflow and per-stage profiling. Choose a target regret before collecting pilot data and compute its floor from the task prior and loss. Estimate action-side rate with uncertainty, then audit gold-entry presence, identifiability, and reference recovery from the actual context. When retention fails, change retrieval or compression; when adequate decision information is certified but outputs use little of it, compare decoders. Recheck retention and regret after intervention: meeting the rate

floor remains only a necessary condition. This workflow has not been validated in a deployment pilot. Log retrieval, assembly, compression, and routing outputs C_i, including visible side inputs and shared state. Without fresh observations, Corollary 6 requires $I(H, C_i) \geq \mathcal{R}_\ell(\varepsilon)$ at every cut for final regret at most ε. Fresh reads require enlarging the history and reapplying the converse. For discrete retrieval identifiers (IDs), tools, and routes, use target-label joint counts, plug-in and MM estimates, sample-support checks, and gold-stratified bootstrap under balanced sampling. Since $X = f(H)$, $I(X, C_i)$ lower-bounds $I(H, C_i)$; a small value alone cannot certify a full-history information deficit. A budget test needs the complete decision variable or a separate upper bound, together with retention and recovery audits.

Measuring text intermediates. A finite joint table is usually too sparse for text contexts. MINE or InfoNCE provide variational lower bounds [6, 33, 48], making estimator expressiveness and sample support part of the diagnosis. Train and evaluate on separate samples, report the critic class, negative-sample scheme, and uncertainty, and calibrate on a discrete stage with known counts. A low variational lower bound cannot certify an information deficit: the estimator may miss information the context preserves. A below-floor upper bound on context information can support such an impossibility claim. No high-dimensional or full multi-stage measurement is completed here. PAC-Bayes and MI analyses relate generalisation to information shared by data and a learned hypothesis [52]; the information bottleneck (IB) trades compression for prediction [47], and rate-distortion analyses of decision making pivot the inequality on the learner [4, 36, 37]. Rational inattention [27, 45] gives the same decision problem in value-versus-rate form. Appendix 3 proves the inverse relation and states the optimal policy. Task-oriented rate distortion also studies downstream decisions [20, 42]. The deployed context cut makes retained information, decoder use, and measurement limits explicit.

Empirical scope. The instantiation covers balanced, single-step $M{=}16$ tool selection across three open-source LLMs and three benchmarks. We do not evaluate multi-turn dialogue, code generation, planning, or continuous control, or directly estimate $I(H, C)$. Experiment 1 audits measurement. Experiments 2 and 3 describe model behaviour; none tests the converse's full generality. Extension requires a task-specific prior and loss, a defensible regret reference, and measurements suited to the new action space. Appendix 14 details these assumptions and the limits of estimators, the action quotient, and achievability constructions.

Future directions and responsible use. A retrieve-then-select pilot would test the proposed workflow by pairing intermediate-context audits with controlled decoder comparisons. CoT on other benchmarks and sample-size or estimator comparisons would test the scope and stability of the observed patterns. Rate-aware training could optimise $\mathbb{E}[\ell(H, A)] + \beta I(H, A)$, with $\beta \geq 0$, while benchmark reports could pair action-side MI with accuracy and context retention. These are proposed studies, not established results. Decision-relevant bit budgets can guide retrieval, compression, and decoding experiments. The input-tokens-per-action-bit gap motivates testing shorter prompts and better information use; it does not establish that extra tokens are useless or that removal saves compute without changing regret. Those claims require direct intervention studies. This work uses existing benchmarks and introduces no new model training. Estimation error and incomplete context logs can misdirect changes, while treating a necessary budget as a performance guarantee can overstate capability. Reporting the target loss, measured channel, uncertainty, and context-audit limits is therefore part of responsible use.

7 Conclusion

A language-model agent can act reliably only when its assembled context preserves enough decision-relevant information for the target regret, yet token counts alone do not specify this requirement. We formalise context adequacy through $H \to C \to A$ and a regret rate-distortion function $\mathcal{R}_\ell(\varepsilon)$, deriving a necessary mutual-information budget at each context stage, exact for uniform zero-one decisions and attainable for finite problems under shared randomness. These results establish the feasibility of decision-based, model-agnostic context accounting that separates a necessary information budget from observed information use and token count while keeping retention audits distinct from action-side measurement, though deployment still requires pilots, estimator and retention validation, broader tasks and per-stage trajectory tracking, calibration of the regret reference, and computational-efficiency improvements.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant EP/X03688X/1. This work was also supported by the EPSRC Centre for Doctoral Training in Distributed Algorithms. The authors are grateful for the support of Stephen Pasteris at The Alan Turing Institute. The authors declare no competing interests.

References

[1] Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. *Anthropic Technical Report*, 2024.

[2] Anthropic. Effective context engineering for AI agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents, 2025. Anthropic Engineering Blog, 29 September 2025.

[3] Anthropic. Effective harnesses for long-running agents. https://www.anthropic.com/engineering/effective-harnesses-for-long-running-agents, 2025. Anthropic Engineering Blog, 26 November 2025.

[4] Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *International Conference on Machine Learning (ICML)*, 2021.

[5] George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E. Hunter, and Katharina von der Wense. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop QA. *arXiv preprint arXiv:2412.10079*, 2024.

[6] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.

[7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[8] Sébastien Bubeck and Mark Sellke. First-order Bayesian regret analysis of Thompson sampling. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT)*, volume 117 of *PMLR*, pages 196–233, 2020.

[9] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[10] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[11] Fabian Falck, Ziyu Wang, and Chris Holmes. Is in-context learning in large language models Bayesian? a martingale perspective. In *International Conference on Machine Learning (ICML)*, 2024.

[12] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[13] Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[14] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[15] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024.

[16] Hong Jun Jeon, Jason D. Lee, Qi Lei, and Benjamin Van Roy. An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*, 2024.

[17] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[18] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[19] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024.

[20] Victoria Kostina and Babak Hassibi. Rate-cost tradeoffs in control. *IEEE Transactions on Automatic Control*, 64(11):4525–4540, 2019.

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018. doi: 10.1109/TIT.2018.2865570.

[23] Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of EMNLP*, 2023.

[24] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[25] Xiaoran Liu, Ruixiao Li, Mianqiu Huang, Zhigeng Liu, Yuerong Song, Qipeng Guo, Siyang He, Qiqi Wang, Linlin Li, Qun Liu, Yaqian Zhou, Xuanjing Huang, and Xipeng Qiu. Thus spake long-context large language model. *arXiv preprint arXiv:2502.17129*, 2025.

[26] Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[27] 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.

[28] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey. *Transactions on Machine Learning Research (TMLR)*, 2023.

[29] George A. Miller. Note on the bias of information estimates. In Henry Quastler, editor, *Information Theory in Psychology: Problems and Methods*, pages 95–100. Free Press, Glencoe, IL, 1955. The Miller–Madow correction combines this estimator with W. G. Madow's earlier (1948) discussion; Madow is not a co-author of this paper.

[30] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, et al. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[31] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[32] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6): 1191–1253, 2003.

[33] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[34] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *Proceedings of ICLR*, 2024.

[35] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[36] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[37] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

[38] Nikolaus Salvatore, Hao Wang, and Qiong Zhang. Lost in the middle: An emergent property from information retrieval demands in LLMs. *arXiv preprint arXiv:2510.10276*, 2025.

[39] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

[40] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948. Published in two parts: vol. 27, no. 3 (Jul.), pp. 379–423; and vol. 27, no. 4 (Oct.), pp. 623–656.

[41] Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 7:142–163, 1959.

[42] Jiawei Shao, Yuyi Mao, and Jun Zhang. Learning task-oriented communication for edge inference: An information bottleneck approach. *IEEE Journal on Selected Areas in Communications*, 40(1):197–211, 2022.

[43] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. *International Conference on Machine Learning (ICML)*, 2023.

[44] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

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

[46] Team GLM. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[47] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[49] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. In *Transactions on Machine Learning Research (TMLR)*, 2024.

[50] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[51] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

[52] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[53] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, et al. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

[54] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[55] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *arXiv preprint arXiv:2406.12045*, 2024.

[56] Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of LLM-based agents. *arXiv preprint arXiv:2503.16416*, 2025.

[57] Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*, 2025.

[58] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[59] Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

[60] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Appendix Contents

1 Notation Glossary 16
2 The Converse in Full Generality 16
3 Derivation of the Hamming Rate-Distortion Formula 18
4 Estimation Protocol 20
5 Per-Cell Summary 21
6 Experiment 1: Measurement audit and margin structure 30
7 Experiment 2: Matched-rate regret clustering 32
8 Experiment 3: CoT shifts 34
9 Lost-in-the-Middle Heatmap 35
10 Tokens-Versus-Bits Gap, Full Panels 36
11 Compression Dose Response 38
12 Failure Modes and Confusion Matrices 39
13 Per-Knob Slope Visualisation 40
14 Limitations 41
15 Reproducibility 43

The appendix is organised as follows. Appendix 1 fixes notation. Appendix 2 proves the converse in full generality (countable, continuous, multi-stage), and Appendix 3 carries the remaining proofs. Appendices 4 to 5 document the experimental protocol and full per-cell numbers and context audit. Appendices 6 to 10 expand on Experiment 1, Experiment 2, Experiment 3, the lost-in-the-middle effect, and the tokens-versus-bits gap. Appendices 11 to 12 provide the compression dose response and failure-mode dissection. Appendix 13 provides per-knob slope visualisations. Appendix 14 details the limitations, and Appendix 15 records reproducibility details and compute resources. The deployment workflow, future directions, and broader-impact considerations are discussed in Section 6.

How figures and tables map to the claims. Experiment 1 audits the measurement pipeline and describes margin structure. Experiment 2 concerns matched-rate regret clustering. Experiment 3 concerns the measured CoT shifts on one slice. T denotes input tokens per action-side bit. U denotes the context audit needed to interpret those measurements. The per-cell grid and margin panels address Experiment 1. The pair scatter and breakdown address Experiment 2. CoT tables and cost panels address Experiment 3. Token and rate panels address T. Position, compression, and confusion panels describe action-side patterns that motivate U. Their shapes alone do not identify the cause of information loss.

1 Notation Glossary

2 The Converse in Full Generality

We restate Theorem 5 in a form that covers the non-finite action spaces relevant to language agents, and follow with the equivalence-class quotient and the multi-stage form. Each generalisation re-uses the same DPI argument. The proofs remain one line.

Countable and continuous actions.

Theorem 10 (General converse)**.** *Let $(\mathcal{H}, \Sigma_{\mathcal{H}}), (\mathcal{C}, \Sigma_{\mathcal{C}}), (\mathcal{A}, \Sigma_{\mathcal{A}})$ be standard Borel spaces, and let $H \to C \to A$ be a Markov chain with $I(H, C) \le C_{\mathrm{ctx}}$ and $\mathbb{E}[\ell(H, A)] \le \varepsilon$, where $\ell(\cdot, \cdot)$ is a non-negative measurable distortion with $\ell(H, A)$ integrable. Then $C_{\mathrm{ctx}} \ge \mathcal{R}_\ell(\varepsilon)$, with $\mathcal{R}_\ell$ as in Definition 2.*

Proof. MI for general random variables is defined via the supremum over finite measurable partitions. The DPI holds at this level of generality [10], so $I(H, A) \le I(H, C) \le C_{\mathrm{ctx}}$. The argument mirrors Lemma 4 and Theorem 5. The infimum in (1) is then over all measurable kernels $\kappa(\cdot \mid h)$ on $\Sigma_{\mathcal{A}}$. □

Equivalence-class collapse.

Lemma 11 (Quotient action space)**.** *Let $\sim$ be the equivalence relation on $\mathcal{A}$ defined by $a \sim a' \iff Q^\star(\cdot, a) = Q^\star(\cdot, a')$ pointwise on $\mathcal{H}$, and let $[A]$ denote the induced quotient. Assume a measurable map chooses one representative from each class, as is automatic for finite action spaces. Then $\mathcal{R}_\ell(\varepsilon) = \inf_{\kappa([a]|h) \text{ with } \mathbb{E}[\ell] \le \varepsilon} I(H, [A])$, and in particular $\mathcal{R}_\ell(\varepsilon) \le \log_2 K_\sim$ when there are $K_\sim$ classes.*

Proof. $\ell(h, a)$ depends on a only through $[a]$, so any kernel $\kappa(\cdot \mid h)$ achieves the same expected regret as its push-forward $\bar{\kappa}([a] \mid h)$. The DPI gives $I(H, [A]) \le I(H, A)$, so the quotient kernel is feasible and weakly improves the rate. Conversely, choose the fixed representative of each quotient action. This measurable lift preserves loss and MI because the class is also recoverable from its representative. Taking infima in both directions gives equality. For the cardinality bound, fix any feasible kernel $\bar{\kappa}$ achieving $\mathbb{E}[\ell(H, A)] \le \varepsilon$ (e.g. the regret-minimising one in the quotient) and bound $I(H, [A]) \le H([A]) \le \log_2 K_\sim$, which gives $\mathcal{R}_\ell(\varepsilon) \le \log_2 K_\sim$ whenever the constraint is feasible. □

Multi-stage scaffolds.

Corollary 12 (Multi-stage form, repeated)**.** *Let the context be constructed by a pipeline $H \to C_1 \to C_2 \to \cdots \to C_k \to A$ with the Markov property at each stage, and let $C_{\mathrm{ctx}}^{(i)} \triangleq \sup I(H, C_i)$. If $\mathbb{E}[\ell(H, A)] \le \varepsilon$, then $\min_i C_{\mathrm{ctx}}^{(i)} \ge \mathcal{R}_\ell(\varepsilon)$.*

symbol	meaning	domain / units	definition / reference
Random variables, channels, and decision rules			
H, h	history random variable / realisation	$\mathcal{H}$ (measurable)	§3
C, c	context (prompt + retrieval + memory)	$\mathcal{C}$ (measurable)	§3
A, a	agent's emitted action / tool call	$\mathcal{A}$ (measurable)	§3
C_i	per-stage context in pipeline $H \to C_1 \to \cdots \to C_k \to A$		Cor. 6
X	gold-tool index $X = f(H)$ for the decision task	$\{1, \ldots, M\}$	§3
ψ	stochastic compression map $H \to C$		§3
π	decoding policy $\pi(a \mid c)$		§3
κ	decision kernel $\kappa(a \mid h)$ in rate-distortion optimisation		Def. 2
$\bar{\kappa}$	push-forward kernel under quotient $a \mapsto [a]$		Lem. 11
P_H	marginal law of the history H		§3
$[A]$	action quotient under $Q^\star$-equivalence	quotient set	Lem. 11
$K_\sim$	number of behavioural action classes	$\mathbb{N}$	Lem. 11
f	decision-label map $X = f(H)$	$\mathcal{H} \to \{1, \ldots, M\}$	§3
$(\Sigma_\mathcal{H}, \Sigma_\mathcal{C}, \Sigma_\mathcal{A})$	σ-algebras of standard Borel spaces		Thm. 10
Value, regret, and rate-distortion functions			
$V^\star(h)$	Bayes-optimal value at history h	$\mathbb{R}$	Def. 1
$Q^\star(h, a)$	Bayes Q-value of action a at history h	$\mathbb{R}$	Def. 1
$\ell(h, a)$	regret distortion $V^\star(h) - Q^\star(h, a)$	$\mathbb{R}_{\geq 0}$	Def. 1
ε	regret-tolerance budget for the converse	$\mathbb{R}_{\geq 0}$	Def. 2
ε°	best-constant-action expected regret	$\mathbb{R}_{\geq 0}$	Prop. 3
$\mathcal{R}_\ell(\varepsilon)$	regret rate-distortion function	bits	Def. 2, Eq. (1)
C_{ctx}	context-channel capacity $\sup_\psi I(H, C)$	bits	§3
$C_{\mathrm{ctx}}^{(i)}$	per-stage capacity $\sup I(H, C_i)$ in pipeline	bits	Cor. 12
$h_2(\varepsilon)$	binary entropy $-\varepsilon \log_2 \varepsilon - (1-\varepsilon) \log_2 (1-\varepsilon)$	bits	Eq. (3)
μ	per-cell margin $\widehat{I}_{\mathrm{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon})$	bits	§5, Experiment 1
q_a	off-diagonal mass $\mathbb{P}(X \neq a \mid A = a)$ in the Hamming proof	$[0, 1]$	Thm. 8, proof
c_a	per-action cost in the heterogeneous-cost extension	$\mathbb{R}_{\geq 0}$	App. 3, "Two other distortions"
$p^\star$	optimal action distribution in rational inattention	$[0, 1]$	Prop. 15
$d(\cdot, \cdot)$	metric in classical reconstruction rate distortion	$\mathbb{R}_{\geq 0}$	Rem. 7
β (rate weight)	coefficient of MI in the rate-aware loss	$\mathbb{R}_{\geq 0}$	§6
Information-theoretic quantities			
$I(X, Y)$	mutual information of X and Y	bits ($\log_2$)	§3
$H(X)$	Shannon entropy of X	bits ($\log_2$)	§4
$H(X \mid A)$	conditional entropy of X given A	bits	Thm. 8, proof
$\widehat{H}_{\mathrm{plug}}$	plug-in entropy estimator $-\sum \widehat{p} \log_2 \widehat{p}$	bits	App. 4
$\widehat{H}_{\mathrm{MM}}$	Miller-Madow corrected entropy $\widehat{H}_{\mathrm{plug}} + (k-1)/(2n \ln 2)$	bits	App. 4
$\widehat{I}_{\mathrm{MM}}(X, A)$	MM-corrected MI $\widehat{H}_{\mathrm{MM}}(X) + \widehat{H}_{\mathrm{MM}}(A) - \widehat{H}_{\mathrm{MM}}(X, A)$	bits	App. 4
$\widehat{p}(x, a)$	empirical joint probability $N(x, a)/n$	$[0, 1]$	App. 4
$N(x, a)$	joint count of trials with gold x and prediction a	$\mathbb{Z}_{\geq 0}$	App. 4
k	number of non-empty bins in the entropy being corrected	$\mathbb{N}$	App. 4
n	trials per cell (n=320 throughout)	$\mathbb{N}$	§5, "Experiment Setup"
Empirical observables (per cell)			
$\widehat{\varepsilon}$	empirical fraction of actions with $A \neq X$	$[0, 1]$	§5, "Experiment Setup"
B	number of bootstrap resamples (B=400)	$\mathbb{N}$	App. 4
$\lvert\Delta\widehat{I}_{\mathrm{MM}}\rvert$, $\lvert\Delta\widehat{\varepsilon}\rvert$	absolute pair gaps in rate / regret (Experiment 2)	bits, $[0, 1]$	§5, Experiment 2
α, β, R^2	intercept, slope (per doubling of N), and goodness-of-fit of $\widehat{\varepsilon} = \alpha + \beta \log_2 N$		App. 13
Experimental design and grid			
M	cardinality of the core tool pool (M=16)	$\mathbb{N}$	Eq. (3)
N	number of catalogue entries, including the gold tool	$\{16, 64, 128, 512\}$	§5, "Experiment Setup"
p	insertion position of the gold tool	{BEG,MID,END}	§5, "Experiment Setup"
τ	description-compression intervention	{T0, T1, T2, C1}	§5, "Experiment Setup"
T0, T1, T2, C1	verbatim; first 50% / 25% of description characters; supplied paraphrase, falling back to T1	intervention codes	App. 4
⟨think⟩	emitted reasoning block under CoT	character stream	App. 8

Table 3. Key symbols and their definitions or references.

Proof. DPI applied stage by stage gives $I(H, A) \leq I(H, C_k) \leq \cdots \leq I(H, C_1) \leq C_{\mathrm{ctx}}^{(1)}$, and similarly $I(H, A) \leq C_{\mathrm{ctx}}^{(i)}$ for every i, so $I(H, A) \leq \min_i C_{\mathrm{ctx}}^{(i)}$. Combining with $\mathcal{R}_\ell(\varepsilon) \leq I(H, A)$ gives the claim. □

The tightest stage of any composed scaffold dominates the bound, so a high-recall retriever feeding a lossy summariser is rate-limited by the summariser's C_{ctx}, not the retriever's. Multi-read agents that re-query H during decoding are recovered by setting C to be the concatenation of all reads. The same DPI applies provided the action depends only on this expanded context.

Connections to information bottleneck, PAC-Bayes, and reconstruction rate-distortion. The information bottleneck objective trades representation rate against predictive information [47]. RRD instead constrains a specified decision loss. They are related information-constrained objectives, but need not coincide for general regret. PAC-Bayes and MI generalisation bounds [52] concern the relation between training data and a learned hypothesis. They complement the deployed-channel question. Their training-data MI is not a second numerical upper bound on the history-to-action MI without further assumptions. Choosing $\ell(h, a) = d(h, a)^q$ for a metric d and exponent $q > 0$ recovers a classical reconstruction loss. Choosing zero-one decision loss recovers the Hamming setting. The context adequacy problem, rather than these classical links, is the focus here.

3 Derivation of the Hamming Rate-Distortion Formula

The complete derivation of Equation (3), stated as Theorem 8, appears in the main text. We collect here the figure and table that the main text references, and extend the closed form to two situations that go beyond the uniform zero-one setup, namely heterogeneous action costs and non-uniform X.

Proposition 13 (Exact context budget for uniform decisions)**.** *Let $X = f(H)$ be uniform on $M \geq 2$ tools, with zero-one loss and random context maps allowed. For $0 \leq \varepsilon \leq 1 - 1/M$, there is a context and a copy decoder with expected regret ε and*

$$I(H, C) = I(X, C) = \mathcal{R}_\ell(\varepsilon).$$

For $\varepsilon \geq 1 - 1/M$, a constant context achieves regret at most ε at zero rate. Thus the budget is both necessary and sufficient within this decoder class.

Proof. Given $X = x$, let $C = x$ with probability $1 - \varepsilon$ and choose each other tool with probability $\varepsilon/(M-1)$. Set $A = C$. This gives $H \to X \to C \to A$ and error probability ε. Uniform X gives uniform C, so

$$I(X, C) = \log_2 M - h_2(\varepsilon) - \varepsilon \log_2(M - 1).$$

Since C depends on H only through X, the chain rule gives $I(H, C) = I(X, C)$. For the zero-rate branch, copy a constant tool, whose error is $1 - 1/M$. The converse in Theorem 5 proves necessity. This uses the classical symmetric channel [10]. It is an existence result for a context that transmits a recommendation. It is not a guarantee for an LLM reading a natural-language prompt. □

Proposition 14 (Finite kernels and one-shot context coding)**.** *Let $\mathcal{H}$ and $\mathcal{A}$ be finite, let ℓ be finite, and let the regret constraint be feasible. An optimal kernel $\kappa^\star$ attains $r = \mathcal{R}_\ell(\varepsilon)$. A random context $C \sim \kappa^\star(\cdot \mid H)$ and copy decoder attain the same regret and mutual information r. If encoder and decoder share randomness U independent of H, the same kernel has a discrete context representation with*

$$H(C \mid U) \leq r + \log_2(r + 1) + 4 \quad \textit{bits}.$$

Here the bound concerns conditional entropy given shared randomness, rather than the unconditional length of a prompt.

Proof. Kernels form a finite product of probability simplexes, which is compact. The feasible set is closed because expected loss is continuous. MI is continuous on these finite alphabets, including boundary probabilities with $0 \log 0 = 0$. It therefore attains a minimum. Taking $C = A^\star$ from this kernel and copying it gives the first context construction.

The strong functional representation lemma [22] gives U independent of H and $A^\star = g(H, U)$ with

$$H(A^\star \mid U) \leq I(H, A^\star) + \log_2(I(H, A^\star) + 1) + 4.$$

Set $C = g(H, U)$ and copy C. Its joint law with H is the optimal one. A conditional prefix code for C given U needs expected length less than $H(C \mid U) + 1$ bits. This coding cost excludes the shared seed. If the seed is available to the decoder, the full decoder input is (C, U) and the Markov chain is $H \to (C, U) \to A$. Its information is $I(H, (C, U)) = H(C \mid U)$ because C is determined by (H, U). This full-input rate need not equal r. The result does not claim an arbitrary fixed-length prompt or an actual LLM attains the bound. □

Proposition 15 (Rational inattention and the inverse budget)**.** *For a finite problem with finite values, set $\bar{V}^\star = \mathbb{E}[V^\star(H)]$ and define*

$$V(k) = \max_{\kappa \text{ with } I(H,A) \leq k} \mathbb{E}[Q^\star(H, A)], \qquad e(k) = \bar{V}^\star - V(k).$$

Then V is concave and non-decreasing, and e is convex and non-increasing. The budget functions are generalised inverses,

$$\mathcal{R}_\ell(\varepsilon) = \inf\{k \geq 0 \mid V(k) \geq \bar{V}^\star - \varepsilon\}, \qquad e(k) = \inf\{\varepsilon \geq 0 \mid \mathcal{R}_\ell(\varepsilon) \leq k\}.$$

In particular, $\mathcal{R}_\ell$ is convex and non-increasing. At a constrained optimum supported by a finite positive multiplier, the policy has the Matějka and McKay form

$$p^\star(a \mid h) = \frac{p_0(a) \exp(Q^\star(h, a)/\lambda)}{\sum_b p_0(b) \exp(Q^\star(h, b)/\lambda)}, \qquad p_0(a) = \sum_h P_H(h) p^\star(a \mid h).$$

The prior p_0 is the endogenous action marginal, not an imposed uniform prior. If the rate multiplier is η in value per bit, then $\lambda = \eta/\ln 2$. The formula applies on the support of p_0. Zero-rate and fully informed endpoints are obtained separately or by limits.

Proof. Compactness gives maximisers. A larger k enlarges the feasible set, proving monotonicity. Mix kernels optimal at k_1 and k_2 with weight t. Expected value mixes linearly and MI is at most $tk_1 + (1-t)k_2$ by convexity in the kernel. This proves concavity of V. Subtracting from $\bar{V}^\star$ gives the properties of e. Since $\mathbb{E}[\ell] = \bar{V}^\star - \mathbb{E}[Q^\star]$, the two feasible-set descriptions give the inverse relations. The same mixture argument with a regret constraint proves convexity of $\mathcal{R}_\ell$.

For a positive supporting multiplier, maximise $\mathbb{E}[Q^\star] - \eta I(H, A)$. Differentiation on a positive support, with row-sum constraints and natural logarithms, gives

$$Q^\star(h,a) - \frac{\eta}{\ln 2} \ln \frac{p(a \mid h)}{p_0(a)} = b(h).$$

Normalising each row yields the displayed policy and the marginal equation. Conversely, the objective is concave in the kernel, so its optimality conditions, including the conditions for excluded actions, are sufficient. Supporting multipliers exist in the interior of the attainable tradeoff. Flat segments can use mixtures of optimal policies and endpoints can use limits. This is the classical rational-inattention characterisation [27, 45], expressed as a regret budget at the context cut. □

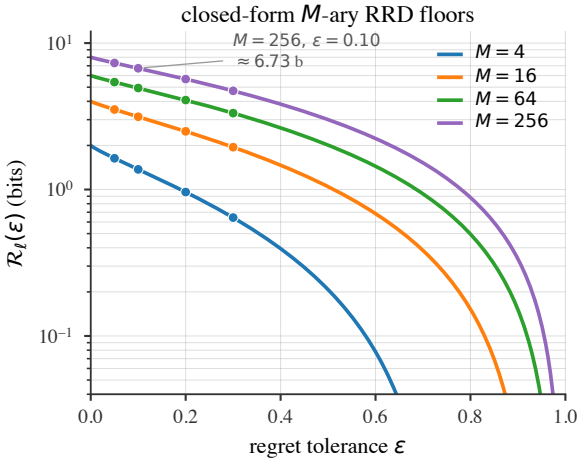

Figure 5. Closed-form Hamming floors $\mathcal{R}_\ell(\varepsilon)$ as functions of regret tolerance for M-ary zero-one decisions (Equation 3). Each curve is the information-theoretic minimum any decision rule must transmit through the prompt to keep expected regret at most ε, so the curves are *forbidden floors*. Below them the converse rules out low-regret behaviour for every model, scaffold, and decoder. Markers are pinned at the practitioner-relevant tolerances $\varepsilon \in \{0.05, 0.1, 0.2, 0.3\}$ and align with the corresponding columns of Table 4. All four curves share the same shape. $\mathcal{R}_\ell(\varepsilon) \to \log_2 M$ as $\varepsilon \to 0$ (perfect action requires the full task entropy) and $\mathcal{R}_\ell(\varepsilon) \to 0$ as $\varepsilon \to (M-1)/M$ (uniform-random guessing), with the slope steepest in the small-ε regime that deployment cares about.

	ε=0.01	0.05	0.10	0.20	0.30	0.50
M=4	1.90	1.63	1.37	0.96	0.64	0.21
M=16	3.88	3.52	3.14	2.50	1.95	1.05
M=64	5.86	5.41	4.93	4.08	3.33	2.01
M=256	7.84	7.31	6.73	5.68	4.72	3.00

Table 4. Numerical values of the closed-form Hamming floors at practitioner-relevant operating points, accurate to two decimal places via Equation 3. At fixed ε the required information grows $\approx \log_2 M$, that is, *linearly in task entropy*, so a 256-way decision needs roughly twice the rate of a 16-way decision for the same regret tolerance. Doubling the inventory adds about $1 - \varepsilon$ bits at fixed regret in the large-M regime. The columns at $\varepsilon \in \{0.05, 0.1, 0.2, 0.3\}$ correspond to the markers in Figure 5.

Two other distortions. Non-uniform priors and heterogeneous costs remain finite decision problems covered by Proposition 14. For costs c_a and $\ell(x,a) = c_a \mathbf{1}[a \neq x]$, solve the regret-constrained kernel problem numerically. There is no general symmetric-channel formula for these costs. For

non-uniform X and $0 \le \varepsilon \le 1 - 1/M$, Fano gives

$$\mathcal{R}_\ell(\varepsilon) \ge \max\{0, H(X) - h_2(\varepsilon) - \varepsilon \log_2(M-1)\}.$$

The exact value follows from the finite-kernel optimisation or its rational-inattention form. Our balanced experiment uses the uniform special case.

4 Estimation Protocol

	Benchmark		
Property	API-BANK	τ-bench	TOOLBENCH
Tool pool $\lvert\mathcal{P}\rvert$	53	26	47,001
Core tools M	16	16	16
Available distractors	37	10	46,985
Benchmark queries	389	156	500

Table 5. Benchmark composition. Nominal catalogue size $N \in \{16, 64, 128, 512\}$, insertion position $p \in \{\text{BEG}, \text{MID}, \text{END}\}$, description compression $\tau \in \{\text{T0}, \text{T1}, \text{T2}, \text{C1}\}$. Distractors are sampled with replacement when fewer than $N-1$ are available, so N counts catalogue entries rather than distinct tools. Each setting uses 320 trials, stratified as 20 per gold tool. Each (model, benchmark) slice has 48 no-CoT settings plus one CoT setting at (N=128, MID, T0), giving $3 \times 3 \times 49 = 441$ main-set cells. A separate 12-cell CoT supplement covers Qwen3 / API-BANK / T0 across all (N, p) settings.

	Benchmark		
Pool size	API-BANK	τ-bench	TOOLBENCH
$N = 16$	817	1,561	2,583
$N = 64$	2,041	4,775	6,705
$N = 128$	3,718	8,961	12,137
$N = 512$	14,041	32,537	32,768

Table 6. Nominal input tokens at (T0, MID, N). Token counts are measured at verbatim (τ=T0) prompts with insertion position p=MID, across the four catalogue sizes N. Entries are the arithmetic mean of the per-cell mean input-token counts across the three models, rounded to the nearest token.

Plug-in MM estimator. For empirical joint counts $N(x, a)$ on the 16×18 table (16 gold tools × {16 core tools, OTHER_DISTRACTOR, INVALID}), the plug-in entropy is $\widehat{H}_{\text{plug}} = -\sum_{x,a} \widehat{p}(x,a) \log_2 \widehat{p}(x,a)$ with $\widehat{p}(x,a) = N(x,a)/n$. The MM correction adds $(k-1)/(2n \ln 2)$ where k is the number of non-empty bins. We apply it to $\widehat{H}(X)$, $\widehat{H}(A)$, and $\widehat{H}(X, A)$ separately and combine $\widehat{I}_{\text{MM}}(X, A) = \widehat{H}_{\text{MM}}(X) + \widehat{H}_{\text{MM}}(A) - \widehat{H}_{\text{MM}}(X, A)$. Combining three corrected entropies is not an exact debiasing of plug-in MI. The resulting estimator behaviour is characterised in Appendix 4.

Bootstrap protocols. The paper analysis draws $B = 400$ resamples within each gold-tool block, keeping 20 trials per class. It recomputes regret and MM MI and reports the 2.5 and 97.5 percentiles. It uses one random stream with seed 12345. The collapse analysis also stratifies by gold and shares resampling indices across all three action variants. Its per-cell seed is $42 + N + \text{ord}(p_1)$, where p_1 is the first letter of the position. Paired means that the variants use the same indices. It is distinct from stratification.

Ordinary resampling draws n indices from the full trial list and need not reproduce the paper's CIs. These protocols can give different intervals at corner cells. Neither CI protocol makes the in-sample Experiment 1 check independent evidence for the theorem. The regret standard-error upper bound is $1/(2\sqrt{320}) \approx 0.028$. Typical MI bootstrap spreads are 0.05 to 0.10 bits.

Estimator-bias diagnostic. The MM correction adds a term of order $(k-1)/(2n \ln 2)$ to each plug-in *entropy* estimate, where k is the number of non-empty bins [29, 32]. Because $\widehat{I}_{\text{MM}} = \widehat{H}_{\text{MM}}(X) + \widehat{H}_{\text{MM}}(A) - \widehat{H}_{\text{MM}}(X, A)$ combines three corrected entropies, the MM correction is *not* an exact debiasing of plug-in MI. The change to plug-in MI is $(k_X + k_A - k_{XA} - 1)/(2n \ln 2)$, which can have either sign. Residual MI bias is also not strictly signed. The plug-in sign check is structural under the balanced empirical law. MM departures require estimator checks before any interpretation.

In our 16×18 table at $n{=}320$ the per-entropy MM term adds up to $(18{\cdot}16{-}1)/(2{\cdot}320{\cdot}\ln 2) \approx 0.65$ b in the worst case (every bin populated), and much less in practice since most cells have k between 20 and 80. As a consequence of combining three independently corrected entropies, a small number of master-table cells report $\widehat{I}_{\mathrm{MM}}$ marginally above $H(X){=}4$ b (e.g. $\widehat{I}_{\mathrm{MM}}{=}4.034$ b on Gemma3 cells where A nearly equals X). We leave the values unclipped to expose the estimator behaviour rather than hide it. The conclusion that $\widehat{I}_{\mathrm{MM}} \geq \mathcal{R}_\ell(\widehat{\varepsilon})$ holds for all 441 cells is robust both at the point estimate and after clipping $\widehat{I}_{\mathrm{MM}} \leq H(X){=}4$ b. Recomputing the clipped margins gives a minimum of 0.000 b among cells with $\widehat{\varepsilon} < 15/16$, compared with $+0.034$ b before clipping. The bootstrap CIs reported alongside each cell make the residual estimator uncertainty visible at the cell level.

configuration entry	value
input budget	32,768 tokens (hard cap)
output budget (no CoT)	128 new tokens
output budget (CoT, Qwen3)	1,024 new tokens
output budget (CoT, Gemma3 / GLM-4)	512 new tokens
temperature	0 (greedy)
top-p	disabled
trial sample	320 queries / cell, stratified by gold tool
batch size	32 trials per forward pass
random seed	20,260,428 (set in run manifest)
hardware	8×NVIDIA A100 80 GB SXM4 (single node, tensor-parallel)
inference engine	vLLM with paged key-value (KV) caching
total compute	$\sim$ 4,000 A100-GPU-hours ($\sim$ 500 wall-clock h on 8-GPU node)

Table 7. Decoding and inference configuration shared across the three models.

benchmark	τ	P5	P25	median	P75	P95	mean
API-BANK	T0	3650	3682	3718	3752	3962	3735
	T1	2849	2876	2890	2904	3015	2904
	T2	2475	2494	2519	2530	2572	2521
	C1	2846	2870	2890	2916	2972	2901
τ-bench	T0	8562	8708	8817	8967	9144	8848
	T1	4942	5011	5077	5129	5225	5075
	T2	3514	3548	3566	3590	3660	3578
	C1	4967	5024	5076	5120	5217	5075
ToolBench	T0	10305	10894	11743	12625	14209	11897
	T1	6825	7164	7462	7889	9422	7686
	T2	4800	4868	5018	5239	5872	5179
	C1	6798	7151	7538	8018	8853	7670

Table 8. Input-prompt token-count distributions by benchmark and intervention for Qwen3 at ($N{=}128$, MID, no-CoT). All entries are in tokens; P5, P25, P75, and P95 denote percentiles. In the prompt builder, T1 and T2 retain the first 50% and 25% of description characters. C1 uses a supplied paraphrase when available and otherwise falls back to the T1 rule. The inference runner supplies no paraphrase lookup. Tool names are retained in all variants.

5 Per-Cell Summary

We summarise all 441 main-grid cells graphically in Figure 6.

Context retention audit and its scope. Table 9 adds presence, identifiability, and recoverability columns for all 441 cells. This is a structural audit of the prompt builder and recorded input lengths, rather than a semantic evaluation of descriptions. The builder places the gold description in the catalogue, applies the chosen text compression, and preserves every tool name. Distractors exclude the entire fixed core pool. Hence the only core identifier in each assembled catalogue is the gold identifier. A reference matcher scans catalogue name fields and returns the sole identifier in that fixed pool. It uses C and the fixed benchmark pool, without the trial's gold label or the model output. This is a property of the controlled task and can make recovery easier than in a real catalogue.

Presence means that the gold entry, including its description after the chosen intervention, is retained. It does not mean that the full T0 description survives compression. Identifiability means that its core identifier is distinct from all distractor identifiers. Recoverability means that the reference

matcher above returns the target. The source code proves all three properties before the input cap. An execution check applies every intervention and position to each recorded core identifier and confirms unique-core recovery. We check this invariant against all logged gold labels and each benchmark's core pool. For each trial with a recorded padded input length strictly below 32,768, the complete prompt fits, so these three properties are certified. Entries report certified counts out of 320. A suffix u marks remaining trials whose length reaches the cap or whose input length is missing. They are unverified rather than failed. Input token IDs after truncation are not stored, so this audit cannot resolve those trials or test a semantic matcher. The same three counts can coincide because the structural matcher needs only the retained name.

Across 141,120 logged trials, 136,199 have a full-input certificate and 4,921 remain unverified. All 320 trials are certified in 407 of 441 cells.

When all 320 trials in a cell are certified, X is recoverable from C and $I(X, C) = 4$ bits by construction. For other cells we make no such equality claim. The per-cell table also gives all three action-side rates. No model inference or image generation is involved.

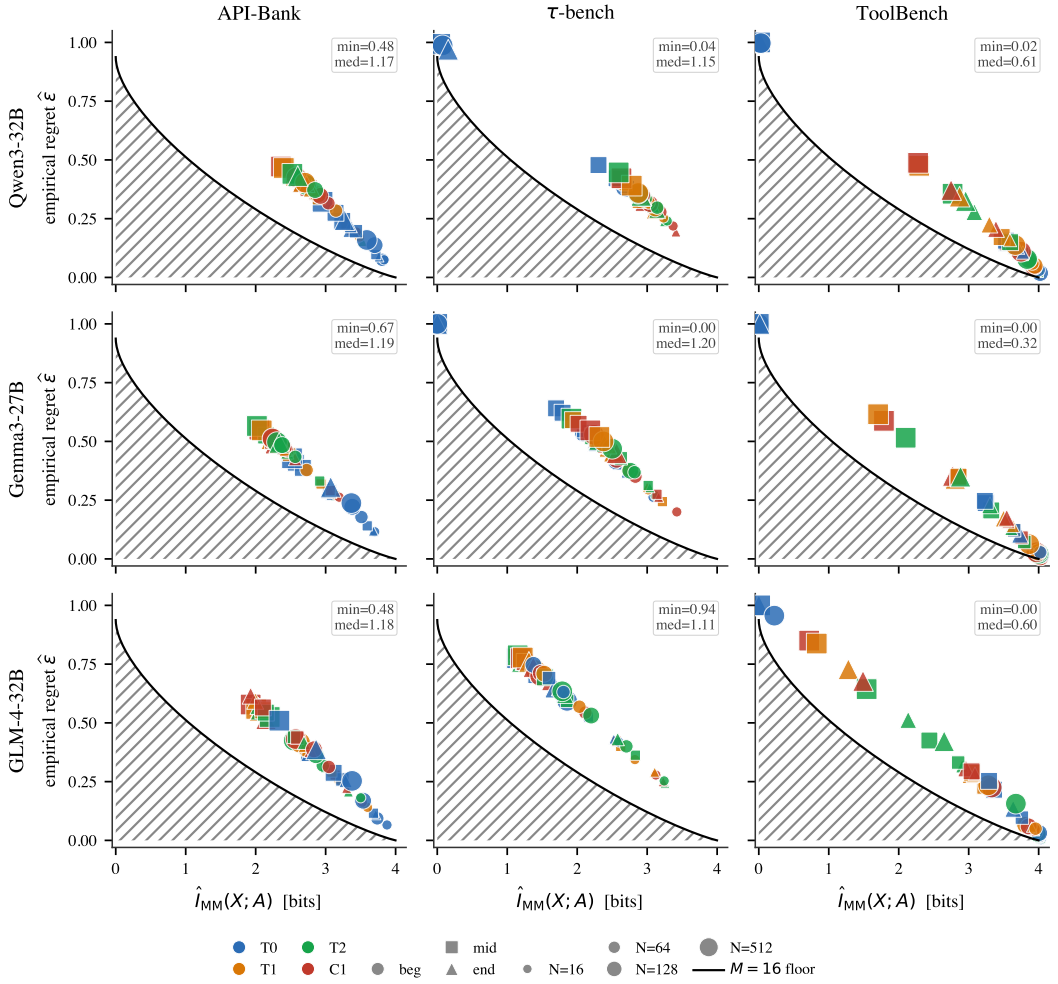

Figure 6. Per-cell summary of the 441-cell main grid. A 3×3 panel grid (rows are *Qwen3-32B / Gemma3-27B / GLM-4-32B*, columns are *API-Bank / τ-bench / ToolBench*) plots every one of the 49 cells of the corresponding (model, bench) slice in $(\widehat{I}_{\mathrm{MM}}, \widehat{\varepsilon})$ space. Marker colour encodes the compression intervention $\tau \in \{\mathrm{T0}, \mathrm{T1}, \mathrm{T2}, \mathrm{C1}\}$, marker shape encodes insertion position $p \in \{\mathrm{BEG}, \mathrm{MID}, \mathrm{END}\}$, and marker size grows with catalogue size $N \in \{16, 64, 128, 512\}$. The black curve is the analytical M=16 Hamming floor (Equation 3). The shaded region beneath it is forbidden. Per-panel boxes report the (min, median) margin $\mu = \widehat{I}_{\mathrm{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon})$ in bits. *No marker in any panel intrudes into the forbidden region.*

Reading the grid. Each of the nine panels collects 49 cells: $4 \times 3 \times 4$ no-CoT settings from the (N, p, τ) sweep and one main-set CoT setting at $(N{=}128, \mathrm{MID}, \mathrm{T0})$. Read down a column to compare models on the same benchmark, and read across a row to track a single model across the

difficulty gradient. The forbidden-region shading falls below the analytical floor and is the visual statement of Experiment 1, in that every coloured marker is on or above the curve.

API-Bank, easiest, no saturation. Across all three models, API-Bank cells cluster in the $\widehat{I}_{\mathrm{MM}} \in [1.9, 3.9]$ b, $\widehat{\varepsilon} \in [0.07, 0.62]$ region, with margin minima comfortably above the floor (Qwen3 $+0.485$ b, Gemma3 $+0.669$ b, GLM-4 $+0.482$ b) and *zero* saturation cells. The verbatim T0 prompts at every N stay under the 32,768-token cap, so the four interventions and three positions fan out smoothly in $\widehat{I}_{\mathrm{MM}}$ rather than collapsing. This panel column shows what the converse looks like *outside* the saturation regime, where compression and position act on $\widehat{I}_{\mathrm{MM}}$, regret tracks the floor, and no operating point comes close to violating it. Median per-panel regrets are 0.34 (Qwen3), 0.43 (Gemma3), and 0.38 (GLM-4), close across models. The converse does not require models to have similar regrets.

τ-bench, a partial-saturation regime, with a GLM-4 outlier. The middle column is where the three models begin to diverge. Qwen3 and Gemma3 still cover the same $\widehat{I}_{\mathrm{MM}}$ range as on API-Bank but pick up 1 and 3 saturation cells respectively at $N{=}512$, where input tokens exceed the cap and $\widehat{I}_{\mathrm{MM}}$ drops below 0.05 b. GLM-4 is the visible outlier, with its cells compressed into the upper-left of the panel ($\widehat{I}_{\mathrm{MM}} \in [1.13, 3.25]$ b, $\widehat{\varepsilon} \in [0.24, 0.79]$, median $\widehat{\varepsilon}{=}0.65$), well off the floor. The larger margin describes a greater distance from the attainable rate frontier at the measured regret. It does not establish that this model's failures are caused by the decoder. That conclusion needs a context audit at a specified target regret.

ToolBench, heavy saturation tail. The right column is dominated by the $N{=}512$ saturation cluster, where markers pile up at the upper-left of the panel ($\widehat{I}_{\mathrm{MM}} \to 0$, $\widehat{\varepsilon} \to 1$) on the floor itself and both sides of the inequality have collapsed simultaneously. Across all 49 cells per model, ToolBench has the highest median $\widehat{I}_{\mathrm{MM}}$ ($3.75/3.79/3.36$ b across the three models) and the lowest median regret. Median regrets are 0.11 (Qwen3), 0.07 (Gemma3), and 0.22 (GLM-4). Saturation counts are $3, 3, 2$ across the three models, all at $N{=}512$ on T0. Compressed prompts can improve on capped T0 cells, the inversion that motivates the "compression-as-recovery" regime of Appendix 11.

Aggregate audit, computed cell by cell. Across the 441 cells the margin distribution is $\mu_{\min}{=}0.000$ b, $\mu_{\mathrm{med}}{=}1.073$ b, $\mu_{\max}{=}1.343$ b, matching the headline statistics of Table 10 and Figure 2 (both recomputed from the same cell-level results). The minimum is touched only at the 5 Gemma3 saturation cells where $\mathcal{R}_\ell(\widehat{\varepsilon}){=}0$ and $\widehat{I}_{\mathrm{MM}}{=}0$ to four decimal places. The $+0.016$ b Qwen3 / ToolBench / $N = 512$ / T0 / end cell lies on the $\widehat{\varepsilon} \geq 15/16$ zero-rate branch. Restricting to $\widehat{\varepsilon} < 15/16$ gives the $+0.034$ b minimum used in the main text. Per-model margin medians are 1.05 (Qwen3), 1.09 (Gemma3), and 1.06 (GLM-4), within 0.04 b of one another, as a descriptive feature of this grid.

Relation to main claims. Figure 6 describes the margin geometry used in the Experiment 1 audit. API-Bank lies well above the floor, ToolBench has tight high-information cells and a saturation tail, and τ-bench is intermediate. A large margin with high regret is allowed by the converse. It does not imply that retrieval changes cannot help. The context columns below separate what can be certified about the prompt from what is measured at the action.

Table 9. Per-cell context audit and action-label collapse results. Model codes Q, G, and L denote Qwen3, Gemma3, and GLM-4. Benchmark codes A, T, and B denote API-Bank, τ-bench, and ToolBench. Setting is (N, p, τ), with b, m, and e for position. A star marks the logged CoT cell included in the 441-cell grid. Rates are in bits. P, I, and R give certified presence, identifiability, and recoverability counts out of 320. A suffix u gives the unverified remainder. These are source-based retention certificates, not semantic matcher scores on stored prompts.

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
G	A	16,b,C1	0.262	3.194	3.047	3.186	320	320	320
G	A	16,b,T0	0.116	3.701	3.610	3.701	320	320	320
G	A	16,b,T1	0.272	3.143	2.989	3.131	320	320	320
G	A	16,b,T2	0.272	3.117	2.953	3.109	320	320	320
G	A	16,e,C1	0.291	3.062	2.916	3.062	320	320	320

Continued on next page

Continued

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
G	A	16,e,T0	0.119	3.682	3.567	3.682	320	320	320
G	A	16,e,T1	0.284	3.087	2.969	3.087	320	320	320
G	A	16,e,T2	0.269	3.136	2.973	3.128	320	320	320
G	A	16,m,C1	0.291	3.066	2.958	3.066	320	320	320
G	A	16,m,T0	0.141	3.594	3.481	3.594	320	320	320
G	A	16,m,T1	0.319	2.934	2.765	2.917	320	320	320
G	A	16,m,T2	0.331	2.916	2.761	2.888	320	320	320
G	A	64,b,C1	0.459	2.450	2.259	2.450	320	320	320
G	A	64,b,T0	0.178	3.512	3.370	3.512	320	320	320
G	A	64,b,T1	0.378	2.728	2.589	2.728	320	320	320
G	A	64,b,T2	0.434	2.566	2.419	2.547	320	320	320
G	A	64,e,C1	0.425	2.564	2.462	2.564	320	320	320
G	A	64,e,T0	0.278	3.097	2.992	3.097	320	320	320
G	A	64,e,T1	0.466	2.428	2.273	2.419	320	320	320
G	A	64,e,T2	0.453	2.451	2.271	2.451	320	320	320
G	A	64,m,C1	0.491	2.265	2.086	2.265	320	320	320
G	A	64,m,T0	0.287	3.056	2.937	3.056	320	320	320
G	A	64,m,T1	0.444	2.478	2.295	2.468	320	320	320
G	A	64,m,T2	0.472	2.394	2.247	2.394	320	320	320
G	A	128,b,C1	0.472	2.399	2.305	2.399	320	320	320
G	A	128,b,T0	0.222	3.382	3.331	3.382	320	320	320
G	A	128,b,T1	0.500	2.257	2.148	2.257	320	320	320
G	A	128,b,T2	0.484	2.378	2.260	2.367	320	320	320
G	A	128,e,C1	0.503	2.243	2.129	2.243	320	320	320
G	A	128,e,T0	0.294	3.041	2.944	3.041	320	320	320
G	A	128,e,T1	0.484	2.314	2.219	2.314	320	320	320
G	A	128,e,T2	0.506	2.234	2.112	2.234	320	320	320
G	A	128,m,C1	0.503	2.248	2.163	2.248	320	320	320
G	A	128,m,T0	0.387	2.674	2.564	2.674	320	320	320
G	A	128,m,T0*	0.409	2.573	2.447	2.573	320	320	320
G	A	128,m,T1	0.553	2.038	1.921	2.028	320	320	320
G	A	128,m,T2	0.525	2.149	2.025	2.149	320	320	320
G	A	512,b,C1	0.512	2.240	2.074	2.240	320	320	320
G	A	512,b,T0	0.238	3.368	3.281	3.368	320	320	320
G	A	512,b,T1	0.509	2.245	2.135	2.245	320	320	320
G	A	512,b,T2	0.497	2.305	2.170	2.305	320	320	320
G	A	512,e,C1	0.491	2.277	2.194	2.277	320	320	320
G	A	512,e,T0	0.306	3.072	2.995	3.072	320	320	320
G	A	512,e,T1	0.509	2.223	2.126	2.223	320	320	320
G	A	512,e,T2	0.494	2.336	2.275	2.336	320	320	320
G	A	512,m,C1	0.550	2.053	1.926	2.053	320	320	320
G	A	512,m,T0	0.428	2.519	2.462	2.519	320	320	320
G	A	512,m,T1	0.547	2.087	2.011	2.087	320	320	320
G	A	512,m,T2	0.566	2.018	1.941	2.018	320	320	320
G	T	16,b,C1	0.200	3.423	3.313	3.423	320	320	320
G	T	16,b,T0	0.262	3.084	3.012	3.084	320	320	320
G	T	16,b,T1	0.294	3.014	2.906	3.014	320	320	320
G	T	16,b,T2	0.269	3.116	3.053	3.116	320	320	320
G	T	16,e,C1	0.262	3.164	3.025	3.164	320	320	320
G	T	16,e,T0	0.366	2.695	2.642	2.695	320	320	320
G	T	16,e,T1	0.306	2.999	2.850	2.999	320	320	320
G	T	16,e,T2	0.303	3.038	2.859	3.038	320	320	320
G	T	16,m,C1	0.275	3.140	2.984	3.140	320	320	320
G	T	16,m,T0	0.400	2.573	2.546	2.573	320	320	320
G	T	16,m,T1	0.244	3.218	3.055	3.218	320	320	320
G	T	16,m,T2	0.312	3.014	2.850	3.014	320	320	320
G	T	64,b,C1	0.350	2.832	2.770	2.832	320	320	320
G	T	64,b,T0	0.409	2.537	2.481	2.537	320	320	320
G	T	64,b,T1	0.378	2.729	2.695	2.729	320	320	320
G	T	64,b,T2	0.369	2.820	2.776	2.820	320	320	320
G	T	64,e,C1	0.488	2.370	2.298	2.370	320	320	320
G	T	64,e,T0	0.500	2.243	2.243	2.243	320	320	320
G	T	64,e,T1	0.475	2.470	2.451	2.470	320	320	320
G	T	64,e,T2	0.475	2.434	2.352	2.434	320	320	320

Continued on next page

Continued

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
G	T	64,m,C1	0.506	2.306	2.286	2.306	320	320	320
G	T	64,m,T0	0.544	2.085	2.085	2.085	320	320	320
G	T	64,m,T1	0.472	2.456	2.396	2.456	320	320	320
G	T	64,m,T2	0.428	2.622	2.585	2.622	320	320	320
G	T	128,b,C1	0.422	2.562	2.545	2.562	320	320	320
G	T	128,b,T0	0.534	2.083	2.063	2.083	320	320	320
G	T	128,b,T1	0.438	2.516	2.507	2.516	320	320	320
G	T	128,b,T2	0.375	2.760	2.706	2.760	320	320	320
G	T	128,e,C1	0.484	2.405	2.376	2.405	320	320	320
G	T	128,e,T0	0.494	2.272	2.252	2.272	320	320	320
G	T	128,e,T1	0.469	2.430	2.402	2.430	320	320	320
G	T	128,e,T2	0.491	2.390	2.330	2.390	320	320	320
G	T	128,m,C1	0.575	2.020	2.009	2.020	320	320	320
G	T	128,m,T0	0.622	1.791	1.791	1.791	320	320	320
G	T	128,m,T0*	0.641	1.698	1.698	1.698	320	320	320
G	T	128,m,T1	0.591	1.940	1.940	1.940	320	320	320
G	T	128,m,T2	0.522	2.228	2.209	2.228	320	320	320
G	T	512,b,C1	0.528	2.238	2.238	2.238	320	320	320
G	T	512,b,T0	1.000	0.000	0.000	0.000	0+320u	0+320u	0+320u
G	T	512,b,T1	0.500	2.376	2.376	2.376	320	320	320
G	T	512,b,T2	0.469	2.500	2.481	2.500	320	320	320
G	T	512,e,C1	0.450	2.563	2.483	2.563	320	320	320
G	T	512,e,T0	1.000	0.000	0.000	0.000	0+320u	0+320u	0+320u
G	T	512,e,T1	0.456	2.523	2.464	2.523	320	320	320
G	T	512,e,T2	0.506	2.309	2.248	2.309	320	320	320
G	T	512,m,C1	0.547	2.188	2.188	2.188	320	320	320
G	T	512,m,T0	1.000	0.000	0.000	0.000	0+320u	0+320u	0+320u
G	T	512,m,T1	0.519	2.317	2.317	2.317	320	320	320
G	T	512,m,T2	0.597	1.918	1.908	1.918	320	320	320
G	B	16,b,C1	0.019	3.991	3.926	3.991	320	320	320
G	B	16,b,T0	0.016	4.002	3.930	3.989	320	320	320
G	B	16,b,T1	0.003	4.034	4.013	4.034	320	320	320
G	B	16,b,T2	0.009	4.023	3.981	4.023	320	320	320
G	B	16,e,C1	0.012	4.023	3.960	4.011	320	320	320
G	B	16,e,T0	0.000	4.034	4.034	4.034	320	320	320
G	B	16,e,T1	0.028	3.944	3.882	3.944	320	320	320
G	B	16,e,T2	0.016	3.997	3.938	3.997	320	320	320
G	B	16,m,C1	0.016	3.997	3.938	3.997	320	320	320
G	B	16,m,T0	0.009	4.023	3.981	4.023	320	320	320
G	B	16,m,T1	0.019	3.991	3.926	3.991	320	320	320
G	B	16,m,T2	0.022	3.978	3.906	3.978	320	320	320
G	B	64,b,C1	0.016	4.016	3.958	4.016	320	320	320
G	B	64,b,T0	0.028	4.023	3.912	3.995	320	320	320
G	B	64,b,T1	0.000	4.034	4.034	4.034	320	320	320
G	B	64,b,T2	0.022	3.997	3.904	3.976	320	320	320
G	B	64,e,C1	0.159	3.567	3.340	3.506	320	320	320
G	B	64,e,T0	0.066	3.840	3.688	3.818	320	320	320
G	B	64,e,T1	0.141	3.632	3.451	3.609	320	320	320
G	B	64,e,T2	0.128	3.609	3.458	3.609	320	320	320
G	B	64,m,C1	0.091	3.756	3.596	3.747	320	320	320
G	B	64,m,T0	0.122	3.650	3.470	3.618	320	320	320
G	B	64,m,T1	0.122	3.642	3.495	3.642	320	320	320
G	B	64,m,T2	0.072	3.794	3.682	3.794	320	320	320
G	B	128,b,C1	0.006	4.025	3.993	4.025	320	320	320
G	B	128,b,T0	0.016	4.034	3.975	4.034	320	320	320
G	B	128,b,T1	0.028	3.986	3.925	3.986	320	320	320
G	B	128,b,T2	0.022	4.011	3.915	3.987	320	320	320
G	B	128,e,C1	0.175	3.544	3.294	3.494	320	320	320
G	B	128,e,T0	0.103	3.733	3.580	3.716	320	320	320
G	B	128,e,T1	0.172	3.500	3.301	3.441	320	320	320
G	B	128,e,T2	0.228	3.305	3.019	3.240	320	320	320
G	B	128,m,C1	0.347	2.849	2.567	2.821	320	320	320
G	B	128,m,T0	0.247	3.231	2.981	3.181	320	320	320
G	B	128,m,T0*	0.238	3.243	2.988	3.184	320	320	320

Continued on next page

Continued

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
G	B	128,m,T1	0.350	2.855	2.621	2.810	320	320	320
G	B	128,m,T2	0.206	3.323	3.190	3.314	320	320	320
G	B	512,b,C1	0.028	3.967	3.860	3.944	312+8u	312+8u	312+8u
G	B	512,b,T0	1.000	0.000	0.000	0.000	0+320u	0+320u	0+320u
G	B	512,b,T1	0.062	3.866	3.712	3.839	309+11u	309+11u	309+11u
G	B	512,b,T2	0.022	4.000	3.912	3.984	320	320	320
G	B	512,e,C1	0.353	2.771	2.575	2.719	312+8u	312+8u	312+8u
G	B	512,e,T0	1.000	0.013	0.000	0.000	0+320u	0+320u	0+320u
G	B	512,e,T1	0.341	2.808	2.628	2.784	314+6u	314+6u	314+6u
G	B	512,e,T2	0.350	2.884	2.684	2.827	320	320	320
G	B	512,m,C1	0.588	1.789	1.583	1.774	312+8u	312+8u	312+8u
G	B	512,m,T0	1.000	0.000	0.000	0.000	0+320u	0+320u	0+320u
G	B	512,m,T1	0.616	1.705	1.483	1.677	318+2u	318+2u	318+2u
G	B	512,m,T2	0.516	2.099	1.928	2.079	320	320	320
L	A	16,b,C1	0.162	3.524	3.391	3.501	320	320	320
L	A	16,b,T0	0.066	3.877	3.777	3.856	320	320	320
L	A	16,b,T1	0.141	3.598	3.475	3.589	320	320	320
L	A	16,b,T2	0.181	3.501	3.327	3.456	320	320	320
L	A	16,e,C1	0.222	3.300	3.150	3.279	320	320	320
L	A	16,e,T0	0.109	3.716	3.572	3.670	320	320	320
L	A	16,e,T1	0.250	3.252	3.103	3.227	320	320	320
L	A	16,e,T2	0.206	3.326	3.158	3.295	320	320	320
L	A	16,m,C1	0.278	3.136	3.021	3.127	320	320	320
L	A	16,m,T0	0.116	3.683	3.533	3.646	320	320	320
L	A	16,m,T1	0.291	3.089	2.915	3.061	320	320	320
L	A	16,m,T2	0.309	3.000	2.865	2.964	320	320	320
L	A	64,b,C1	0.312	3.046	2.866	3.031	320	320	320
L	A	64,b,T0	0.094	3.738	3.623	3.728	320	320	320
L	A	64,b,T1	0.262	3.201	3.043	3.182	320	320	320
L	A	64,b,T2	0.319	2.959	2.857	2.949	320	320	320
L	A	64,e,C1	0.381	2.785	2.582	2.704	320	320	320
L	A	64,e,T0	0.253	3.286	3.098	3.188	320	320	320
L	A	64,e,T1	0.403	2.703	2.500	2.651	320	320	320
L	A	64,e,T2	0.416	2.691	2.451	2.635	320	320	320
L	A	64,m,C1	0.438	2.595	2.331	2.502	320	320	320
L	A	64,m,T0	0.259	3.239	3.066	3.179	320	320	320
L	A	64,m,T1	0.384	2.755	2.572	2.721	320	320	320
L	A	64,m,T2	0.441	2.549	2.355	2.497	320	320	320
L	A	128,b,C1	0.387	2.833	2.624	2.710	320	320	320
L	A	128,b,T0	0.169	3.536	3.479	3.521	320	320	320
L	A	128,b,T1	0.394	2.710	2.554	2.644	320	320	320
L	A	128,b,T2	0.363	2.859	2.711	2.818	320	320	320
L	A	128,e,C1	0.616	1.928	1.554	1.678	320	320	320
L	A	128,e,T0	0.372	2.765	2.563	2.641	320	320	320
L	A	128,e,T1	0.594	1.993	1.723	1.813	320	320	320
L	A	128,e,T2	0.572	2.039	1.810	1.925	320	320	320
L	A	128,m,C1	0.566	2.103	1.857	1.933	320	320	320
L	A	128,m,T0	0.372	2.763	2.574	2.670	320	320	320
L	A	128,m,T0*	0.287	3.114	2.980	3.099	320	320	320
L	A	128,m,T1	0.553	2.109	1.881	2.007	320	320	320
L	A	128,m,T2	0.544	2.145	1.938	2.050	320	320	320
L	A	512,b,C1	0.431	2.572	2.339	2.525	320	320	320
L	A	512,b,T0	0.253	3.378	3.149	3.248	320	320	320
L	A	512,b,T1	0.416	2.635	2.469	2.594	320	320	320
L	A	512,b,T2	0.425	2.543	2.392	2.533	320	320	320
L	A	512,e,C1	0.516	2.147	1.994	2.136	320	320	320
L	A	512,e,T0	0.387	2.864	2.601	2.699	320	320	320
L	A	512,e,T1	0.550	2.052	1.915	2.042	320	320	320
L	A	512,e,T2	0.556	2.070	1.968	2.056	320	320	320
L	A	512,m,C1	0.578	1.928	1.760	1.917	320	320	320
L	A	512,m,T0	0.509	2.339	2.026	2.136	320	320	320
L	A	512,m,T1	0.559	2.007	1.890	1.997	320	320	320
L	A	512,m,T2	0.525	2.201	2.020	2.173	320	320	320
L	T	16,b,C1	0.278	3.124	3.041	3.105	320	320	320

Continued on next page

Continued

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
L	T	16,b,T0	0.419	2.593	2.556	2.593	320	320	320
L	T	16,b,T1	0.344	2.825	2.737	2.798	320	320	320
L	T	16,b,T2	0.253	3.243	3.152	3.222	320	320	320
L	T	16,e,C1	0.250	3.230	3.164	3.185	320	320	320
L	T	16,e,T0	0.431	2.524	2.484	2.511	320	320	320
L	T	16,e,T1	0.291	3.107	3.052	3.082	320	320	320
L	T	16,e,T2	0.244	3.254	3.186	3.245	320	320	320
L	T	16,m,C1	0.422	2.561	2.528	2.528	320	320	320
L	T	16,m,T0	0.525	2.160	2.111	2.131	320	320	320
L	T	16,m,T1	0.400	2.622	2.592	2.610	320	320	320
L	T	16,m,T2	0.363	2.832	2.761	2.795	320	320	320
L	T	64,b,C1	0.544	2.116	2.027	2.027	320	320	320
L	T	64,b,T0	0.631	1.804	1.713	1.713	320	320	320
L	T	64,b,T1	0.569	2.032	1.930	1.930	320	320	320
L	T	64,b,T2	0.400	2.703	2.643	2.669	320	320	320
L	T	64,e,C1	0.666	1.584	1.557	1.557	320	320	320
L	T	64,e,T0	0.644	1.728	1.655	1.655	320	320	320
L	T	64,e,T1	0.644	1.683	1.642	1.642	320	320	320
L	T	64,e,T2	0.431	2.579	2.510	2.528	320	320	320
L	T	64,m,C1	0.706	1.428	1.407	1.407	320	320	320
L	T	64,m,T0	0.691	1.598	1.468	1.468	320	320	320
L	T	64,m,T1	0.725	1.396	1.338	1.338	320	320	320
L	T	64,m,T2	0.594	1.856	1.852	1.852	320	320	320
L	T	128,b,C1	0.713	1.480	1.383	1.383	320	320	320
L	T	128,b,T0	0.747	1.375	1.254	1.254	320	320	320
L	T	128,b,T1	0.709	1.529	1.394	1.394	320	320	320
L	T	128,b,T2	0.531	2.200	2.095	2.105	320	320	320
L	T	128,e,C1	0.772	1.172	1.153	1.153	320	320	320
L	T	128,e,T0	0.791	1.131	1.082	1.082	320	320	320
L	T	128,e,T1	0.759	1.234	1.207	1.207	320	320	320
L	T	128,e,T2	0.619	1.826	1.785	1.785	320	320	320
L	T	128,m,C1	0.759	1.250	1.206	1.206	320	320	320
L	T	128,m,T0	0.778	1.176	1.129	1.129	320	320	320
L	T	128,m,T0*	0.703	1.416	1.416	1.416	320	320	320
L	T	128,m,T1	0.781	1.137	1.118	1.118	320	320	320
L	T	128,m,T2	0.694	1.538	1.455	1.455	320	320	320
L	T	512,b,C1	0.700	1.471	1.450	1.450	320	320	320
L	T	512,b,T0	0.594	1.855	1.803	1.803	287+33u	287+33u	287+33u
L	T	512,b,T1	0.725	1.402	1.335	1.335	320	320	320
L	T	512,b,T2	0.634	1.786	1.724	1.724	320	320	320
L	T	512,e,C1	0.738	1.373	1.301	1.301	320	320	320
L	T	512,e,T0	0.650	1.682	1.622	1.622	308+12u	308+12u	308+12u
L	T	512,e,T1	0.759	1.310	1.205	1.205	320	320	320
L	T	512,e,T2	0.759	1.207	1.203	1.203	320	320	320
L	T	512,m,C1	0.775	1.194	1.129	1.129	320	320	320
L	T	512,m,T0	0.775	1.141	1.105	1.105	295+25u	295+25u	295+25u
L	T	512,m,T1	0.778	1.222	1.137	1.137	320	320	320
L	T	512,m,T2	0.787	1.150	1.097	1.097	320	320	320
L	B	16,b,C1	0.012	4.021	3.970	4.021	320	320	320
L	B	16,b,T0	0.003	4.034	4.013	4.034	320	320	320
L	B	16,b,T1	0.012	4.021	3.970	4.021	320	320	320
L	B	16,b,T2	0.009	4.023	3.981	4.023	320	320	320
L	B	16,e,C1	0.044	3.932	3.791	3.898	320	320	320
L	B	16,e,T0	0.012	4.002	3.972	4.002	320	320	320
L	B	16,e,T1	0.056	3.899	3.770	3.868	320	320	320
L	B	16,e,T2	0.009	4.025	3.972	4.014	320	320	320
L	B	16,m,C1	0.041	3.942	3.820	3.922	320	320	320
L	B	16,m,T0	0.022	3.999	3.915	3.987	320	320	320
L	B	16,m,T1	0.037	3.928	3.830	3.928	320	320	320
L	B	16,m,T2	0.022	3.991	3.912	3.984	320	320	320
L	B	64,b,C1	0.047	3.929	3.802	3.912	320	320	320
L	B	64,b,T0	0.019	4.013	3.928	3.993	320	320	320
L	B	64,b,T1	0.050	3.956	3.821	3.935	320	320	320
L	B	64,b,T2	0.034	3.964	3.864	3.957	320	320	320

Continued on next page

Continued

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
L	B	64,e,C1	0.284	3.090	2.774	3.013	320	320	320
L	B	64,e,T0	0.056	3.882	3.741	3.862	320	320	320
L	B	64,e,T1	0.275	3.059	2.824	3.032	320	320	320
L	B	64,e,T2	0.316	2.907	2.665	2.864	320	320	320
L	B	64,m,C1	0.219	3.301	3.011	3.228	320	320	320
L	B	64,m,T0	0.097	3.760	3.613	3.704	320	320	320
L	B	64,m,T1	0.225	3.212	2.976	3.168	320	320	320
L	B	64,m,T2	0.331	2.850	2.587	2.809	320	320	320
L	B	128,b,C1	0.059	3.857	3.725	3.849	320	320	320
L	B	128,b,T0	0.031	4.007	3.895	3.983	320	320	320
L	B	128,b,T1	0.066	3.807	3.677	3.807	320	320	320
L	B	128,b,T2	0.028	3.985	3.884	3.967	320	320	320
L	B	128,e,C1	0.306	2.961	2.638	2.883	320	320	320
L	B	128,e,T0	0.134	3.636	3.420	3.599	320	320	320
L	B	128,e,T1	0.278	3.026	2.783	2.992	320	320	320
L	B	128,e,T2	0.512	2.136	1.807	2.043	320	320	320
L	B	128,m,C1	0.294	3.042	2.696	2.937	320	320	320
L	B	128,m,T0	0.253	3.290	3.060	3.139	320	320	320
L	B	128,m,T0*	0.216	3.363	3.200	3.273	320	320	320
L	B	128,m,T1	0.287	3.049	2.703	2.942	320	320	320
L	B	128,m,T2	0.425	2.438	2.157	2.378	320	320	320
L	B	512,b,C1	0.225	3.328	3.060	3.279	317+3u	317+3u	317+3u
L	B	512,b,T0	0.956	0.222	0.164	0.177	0+320u	0+320u	0+320u
L	B	512,b,T1	0.234	3.279	3.023	3.219	317+3u	317+3u	317+3u
L	B	512,b,T2	0.156	3.674	3.459	3.624	320	320	320
L	B	512,e,C1	0.678	1.489	1.293	1.388	319+1u	319+1u	319+1u
L	B	512,e,T0	1.000	0.001	0.000	0.000	0+320u	0+320u	0+320u
L	B	512,e,T1	0.728	1.281	1.043	1.169	317+3u	317+3u	317+3u
L	B	512,e,T2	0.422	2.650	2.383	2.537	320	320	320
L	B	512,m,C1	0.850	0.720	0.533	0.609	319+1u	319+1u	319+1u
L	B	512,m,T0	1.000	0.020	0.000	0.000	0+320u	0+320u	0+320u
L	B	512,m,T1	0.838	0.828	0.493	0.672	320	320	320
L	B	512,m,T2	0.644	1.541	1.260	1.513	320	320	320
Q	A	16,b,C1	0.194	3.438	3.305	3.438	320	320	320
Q	A	16,b,T0	0.075	3.835	3.735	3.835	320	320	320
Q	A	16,b,T1	0.172	3.484	3.358	3.484	320	320	320
Q	A	16,b,T2	0.191	3.453	3.291	3.453	320	320	320
Q	A	16,e,C1	0.200	3.365	3.244	3.354	320	320	320
Q	A	16,e,T0	0.088	3.760	3.664	3.760	320	320	320
Q	A	16,e,T1	0.219	3.323	3.208	3.323	320	320	320
Q	A	16,e,T2	0.153	3.527	3.409	3.527	320	320	320
Q	A	16,m,C1	0.225	3.291	3.185	3.291	320	320	320
Q	A	16,m,T0	0.097	3.718	3.602	3.718	320	320	320
Q	A	16,m,T1	0.228	3.265	3.135	3.265	320	320	320
Q	A	16,m,T2	0.225	3.284	3.179	3.284	320	320	320
Q	A	64,b,C1	0.316	3.040	2.859	3.040	320	320	320
Q	A	64,b,T0	0.075	3.807	3.717	3.807	320	320	320
Q	A	64,b,T1	0.284	3.148	2.991	3.139	320	320	320
Q	A	64,b,T2	0.344	2.909	2.781	2.900	320	320	320
Q	A	64,e,C1	0.356	2.842	2.724	2.842	320	320	320
Q	A	64,e,T0	0.219	3.355	3.262	3.355	320	320	320
Q	A	64,e,T1	0.359	2.852	2.690	2.834	320	320	320
Q	A	64,e,T2	0.384	2.744	2.559	2.744	320	320	320
Q	A	64,m,C1	0.369	2.795	2.644	2.795	320	320	320
Q	A	64,m,T0	0.197	3.432	3.322	3.432	320	320	320
Q	A	64,m,T1	0.338	2.904	2.740	2.895	320	320	320
Q	A	64,m,T2	0.394	2.677	2.527	2.677	320	320	320
Q	A	128,b,C1	0.347	2.924	2.828	2.924	320	320	320
Q	A	128,b,T0	0.137	3.700	3.638	3.700	320	320	320
Q	A	128,b,T1	0.353	2.890	2.785	2.890	320	320	320
Q	A	128,b,T2	0.372	2.850	2.694	2.841	320	320	320
Q	A	128,e,C1	0.397	2.729	2.564	2.729	320	320	320
Q	A	128,e,T0	0.206	3.365	3.273	3.365	320	320	320
Q	A	128,e,T1	0.378	2.799	2.641	2.789	320	320	320

Continued on next page

Continued

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
Q	A	128,e,T2	0.403	2.678	2.553	2.678	320	320	320
Q	A	128,m,C1	0.381	2.754	2.675	2.754	320	320	320
Q	A	128,m,T0	0.241	3.237	3.139	3.237	320	320	320
Q	A	128,m,T0*	0.275	3.141	3.013	3.129	320	320	320
Q	A	128,m,T1	0.425	2.546	2.451	2.546	320	320	320
Q	A	128,m,T2	0.419	2.585	2.471	2.585	320	320	320
Q	A	512,b,C1	0.425	2.583	2.398	2.583	320	320	320
Q	A	512,b,T0	0.159	3.588	3.491	3.588	320	320	320
Q	A	512,b,T1	0.403	2.706	2.524	2.706	320	320	320
Q	A	512,b,T2	0.394	2.722	2.595	2.722	320	320	320
Q	A	512,e,C1	0.406	2.672	2.534	2.672	320	320	320
Q	A	512,e,T0	0.244	3.269	3.191	3.269	320	320	320
Q	A	512,e,T1	0.406	2.674	2.561	2.674	320	320	320
Q	A	512,e,T2	0.431	2.602	2.499	2.602	320	320	320
Q	A	512,m,C1	0.472	2.362	2.251	2.362	320	320	320
Q	A	512,m,T0	0.322	2.956	2.877	2.956	320	320	320
Q	A	512,m,T1	0.466	2.402	2.320	2.402	320	320	320
Q	A	512,m,T2	0.441	2.525	2.392	2.525	320	320	320
Q	T	16,b,C1	0.219	3.370	3.255	3.370	320	320	320
Q	T	16,b,T0	0.284	3.093	3.038	3.093	320	320	320
Q	T	16,b,T1	0.253	3.234	3.133	3.234	320	320	320
Q	T	16,b,T2	0.241	3.286	3.177	3.286	320	320	320
Q	T	16,e,C1	0.194	3.411	3.360	3.411	320	320	320
Q	T	16,e,T0	0.297	3.051	3.036	3.051	320	320	320
Q	T	16,e,T1	0.294	3.066	3.010	3.066	320	320	320
Q	T	16,e,T2	0.238	3.248	3.171	3.248	320	320	320
Q	T	16,m,C1	0.284	3.080	3.016	3.080	320	320	320
Q	T	16,m,T0	0.312	2.912	2.873	2.912	320	320	320
Q	T	16,m,T1	0.278	3.087	3.023	3.087	320	320	320
Q	T	16,m,T2	0.284	3.104	3.019	3.104	320	320	320
Q	T	64,b,C1	0.278	3.188	3.094	3.188	320	320	320
Q	T	64,b,T0	0.369	2.707	2.623	2.707	320	320	320
Q	T	64,b,T1	0.319	3.021	2.952	3.021	320	320	320
Q	T	64,b,T2	0.297	3.143	3.046	3.143	320	320	320
Q	T	64,e,C1	0.328	2.952	2.861	2.952	320	320	320
Q	T	64,e,T0	0.331	2.968	2.968	2.968	320	320	320
Q	T	64,e,T1	0.334	2.971	2.963	2.971	320	320	320
Q	T	64,e,T2	0.303	3.085	3.018	3.085	320	320	320
Q	T	64,m,C1	0.403	2.661	2.595	2.661	320	320	320
Q	T	64,m,T0	0.384	2.723	2.723	2.723	320	320	320
Q	T	64,m,T1	0.397	2.694	2.649	2.694	320	320	320
Q	T	64,m,T2	0.422	2.615	2.558	2.615	320	320	320
Q	T	128,b,C1	0.312	3.040	2.973	3.040	320	320	320
Q	T	128,b,T0	0.381	2.669	2.594	2.669	320	320	320
Q	T	128,b,T1	0.294	3.115	3.019	3.115	320	320	320
Q	T	128,b,T2	0.334	2.953	2.861	2.953	320	320	320
Q	T	128,e,C1	0.344	2.927	2.875	2.927	320	320	320
Q	T	128,e,T0	0.341	2.889	2.889	2.889	320	320	320
Q	T	128,e,T1	0.284	3.107	3.061	3.107	320	320	320
Q	T	128,e,T2	0.287	3.141	3.086	3.141	320	320	320
Q	T	128,m,C1	0.425	2.602	2.565	2.602	320	320	320
Q	T	128,m,T0	0.422	2.563	2.563	2.563	320	320	320
Q	T	128,m,T0*	0.478	2.304	2.304	2.304	320	320	320
Q	T	128,m,T1	0.416	2.615	2.559	2.615	320	320	320
Q	T	128,m,T2	0.422	2.620	2.530	2.620	320	320	320
Q	T	512,b,C1	0.350	2.898	2.781	2.898	320	320	320
Q	T	512,b,T0	0.988	0.077	0.051	0.051	6+314u	6+314u	6+314u
Q	T	512,b,T1	0.359	2.875	2.757	2.875	320	320	320
Q	T	512,b,T2	0.356	2.872	2.789	2.872	320	320	320
Q	T	512,e,C1	0.325	2.970	2.891	2.970	320	320	320
Q	T	512,e,T0	0.969	0.156	0.127	0.127	15+305u	15+305u	15+305u
Q	T	512,e,T1	0.338	2.925	2.822	2.925	320	320	320
Q	T	512,e,T2	0.347	2.915	2.872	2.915	320	320	320
Q	T	512,m,C1	0.422	2.633	2.506	2.633	320	320	320

Continued on next page

Continued

M	Bench	Setting	$\widehat{\varepsilon}$	$\widehat{I}_{\mathrm{MM},18}$	$\widehat{I}_{\mathrm{MM},16}$	$\widehat{I}_{\mathrm{MM},17}$	P	I	R
Q	T	512,m,T0	0.994	0.043	0.025	0.025	6+314u	6+314u	6+314u
Q	T	512,m,T1	0.391	2.780	2.681	2.780	320	320	320
Q	T	512,m,T2	0.447	2.592	2.523	2.592	320	320	320
Q	B	16,b,C1	0.009	4.023	3.981	4.023	320	320	320
Q	B	16,b,T0	0.025	3.975	3.897	3.975	320	320	320
Q	B	16,b,T1	0.019	4.019	3.954	4.019	320	320	320
Q	B	16,b,T2	0.016	4.020	3.962	4.020	320	320	320
Q	B	16,e,C1	0.050	3.909	3.795	3.909	320	320	320
Q	B	16,e,T0	0.022	3.991	3.919	3.991	320	320	320
Q	B	16,e,T1	0.041	3.914	3.812	3.914	320	320	320
Q	B	16,e,T2	0.053	3.870	3.752	3.870	320	320	320
Q	B	16,m,C1	0.037	3.937	3.839	3.937	320	320	320
Q	B	16,m,T0	0.019	4.020	3.954	4.019	320	320	320
Q	B	16,m,T1	0.031	3.964	3.875	3.964	320	320	320
Q	B	16,m,T2	0.022	3.998	3.926	3.998	320	320	320
Q	B	64,b,C1	0.041	3.956	3.854	3.956	320	320	320
Q	B	64,b,T0	0.031	4.000	3.912	4.000	320	320	320
Q	B	64,b,T1	0.059	3.910	3.786	3.910	320	320	320
Q	B	64,b,T2	0.028	3.980	3.897	3.980	320	320	320
Q	B	64,e,C1	0.159	3.514	3.321	3.514	320	320	320
Q	B	64,e,T0	0.106	3.775	3.621	3.717	320	320	320
Q	B	64,e,T1	0.159	3.589	3.422	3.589	320	320	320
Q	B	64,e,T2	0.113	3.723	3.581	3.723	320	320	320
Q	B	64,m,C1	0.109	3.720	3.548	3.713	320	320	320
Q	B	64,m,T0	0.106	3.723	3.566	3.704	320	320	320
Q	B	64,m,T1	0.100	3.746	3.577	3.736	320	320	320
Q	B	64,m,T2	0.100	3.758	3.595	3.753	320	320	320
Q	B	128,b,C1	0.044	3.924	3.817	3.924	320	320	320
Q	B	128,b,T0	0.019	4.020	3.940	4.006	320	320	320
Q	B	128,b,T1	0.050	3.943	3.829	3.943	320	320	320
Q	B	128,b,T2	0.075	3.856	3.717	3.856	320	320	320
Q	B	128,e,C1	0.206	3.389	3.170	3.383	320	320	320
Q	B	128,e,T0	0.134	3.678	3.534	3.644	320	320	320
Q	B	128,e,T1	0.225	3.296	3.114	3.288	320	320	320
Q	B	128,e,T2	0.278	3.080	2.836	3.073	320	320	320
Q	B	128,m,C1	0.150	3.556	3.363	3.551	320	320	320
Q	B	128,m,T0	0.156	3.534	3.313	3.504	320	320	320
Q	B	128,m,T0*	0.147	3.551	3.371	3.531	320	320	320
Q	B	128,m,T1	0.172	3.471	3.292	3.464	320	320	320
Q	B	128,m,T2	0.150	3.592	3.404	3.592	320	320	320
Q	B	512,b,C1	0.109	3.748	3.546	3.711	315+5u	315+5u	315+5u
Q	B	512,b,T0	0.997	0.029	0.013	0.013	0+320u	0+320u	0+320u
Q	B	512,b,T1	0.137	3.667	3.452	3.634	313+7u	313+7u	313+7u
Q	B	512,b,T2	0.078	3.835	3.685	3.826	320	320	320
Q	B	512,e,C1	0.372	2.748	2.605	2.725	318+2u	318+2u	318+2u
Q	B	512,e,T0	1.000	0.016	0.000	0.000	0+320u	0+320u	0+320u
Q	B	512,e,T1	0.344	2.870	2.677	2.819	315+5u	315+5u	315+5u
Q	B	512,e,T2	0.325	2.957	2.775	2.928	320	320	320
Q	B	512,m,C1	0.488	2.277	2.017	2.151	315+5u	315+5u	315+5u
Q	B	512,m,T0	1.000	0.019	0.000	0.000	0+320u	0+320u	0+320u
Q	B	512,m,T1	0.475	2.293	2.109	2.228	320	320	320
Q	B	512,m,T2	0.356	2.770	2.492	2.748	320	320	320

6 Experiment 1: Measurement audit and margin structure

Table 10 reports the extremal statistics of the margin distribution (the minimum, the median, and the maximum) broken out per (model, benchmark) and across-model. The remainder of this appendix examines the entire distribution and the cells that approach the floor most closely.

Distributional shape. Figure 7 shows the per-benchmark histogram and *empirical cumulative distribution function* (ECDF) of the 441 margins. On API-BANK, the lower-margin cells are mostly T0, while the compressed variants have median margins of approximately 1.20 to 1.22 b. τ-bench

	API-BANK			τ-bench			TOOLBENCH		
Model	*min*	*med*	*max*	*min*	*med*	*max*	*min*	*med*	*max*
Qwen3-32B	0.485	1.172	1.273	0.043	1.154	1.330	0.016	0.605	1.181
Gemma3-27B	0.669	1.190	1.269	0.000	1.201	1.343	0.000	0.318	1.185
GLM-4-32B	0.482	1.182	1.341	0.938	1.109	1.273	0.001	0.598	1.280
All 3 models	**0.482**	**1.180**	**1.341**	**0.000**	**1.154**	**1.343**	**0.000**	**0.467**	**1.280**

Table 10. Experiment 1 measurement and margin audit on the 441-cell main grid. Entries are min, median, and max of the margin $\mu \triangleq \widehat{I}_{\mathrm{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon})$ in bits (49 cells per (model, benchmark) slice; 147 cells per benchmark pooled across models). All 441 cells satisfy $\mu \geq 0$, with grand min / median / max $= 0.000$ / 1.073 / 1.343 bits.

has a single mode near 1.1 b, and ToolBench is heavy-tailed to the left, dominated by the saturation cells at $N{=}512$ where both $\widehat{I}_{\mathrm{MM}}$ and $\mathcal{R}_\ell$ collapse to zero. No benchmark places any mass in the negative-margin region.

Smallest-margin cells. Figure 8 dissects the 40 tightest cells from four complementary angles, and Table 12 enumerates the bottom 30. Panel (a) shows that these cells split cleanly into two qualitatively distinct clusters on the rate-distortion plane, namely a saturation cluster ($\widehat{\varepsilon}\to 1$, $\widehat{I}_{\mathrm{MM}}\to 0$, both sides of the bound collapsed to ≈ 0) and a near-zero-regret cluster ($\widehat{\varepsilon}\to 0$, $\widehat{I}_{\mathrm{MM}}\to \log_2 M$) sitting on the high-information corner of the floor. Panel (b) sorts the cells by their reported CI corner statistic $\widehat{I}^{\mathrm{lo}}_{\mathrm{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon}^{\mathrm{hi}})$, showing that this quantity is non-negative for every cell, with the saturation cluster clamped at zero and the near-zero-regret cluster reaching up to ~ 0.33 b. Panel (c) plots the point margin and the two reported corner statistics from the marginal CIs. These endpoints are not a paired bootstrap CI on the margin. Panel (d) reveals where these tight cells live in the design grid. 34 of 40 are on ToolBench (concentrated at $N{=}16$ and $N{=}512$), 6 are on τ-bench at $N{=}512$, and API-BANK contributes none.

Relation to main claims. The distribution and tight-cell plots locate operating points relative to the exact frontier. This is more informative than the sign check, which is structural for balanced plug-in counts. The tight cells split between the zero-rate branch and the high-information corner. MM correction and confidence intervals audit estimator behaviour. They do not turn Experiment 1 into an independent test of Fano.

Absorbing-label robustness. We use the existing 441-cell post-processing results. The original action has 18 values. The 16-valued version maps both absorbing labels to the alphabetically first core tool. The 17-valued version merges them into one absorbing label. Both maps are deterministic functions of the original action. Plug-in DPI holds cell by cell. The median decrease in MM MI from 18 to 16 values is 0.105 bits and the maximum is 0.374 bits. All per-cell rates appear in Table 9.

Action values	Cells	Negative	Min	Median	Max	Min below $15/16$
18 original	441	0	0.000	1.073	1.343	0.034
16 collapsed	441	0	0.000	0.949	1.343	0.034
17 merged	441	0	0.000	1.044	1.343	0.034

Table 11. Margins in bits after action-label post-processing. Every variant uses the original regret to evaluate the floor. The final column restricts original regret to $\widehat{\varepsilon} < 15/16$, rather than restricting token counts. Negative counts use point estimates.

The floor uses original regret $\widehat{\varepsilon}$, where both absorbing labels are errors. We change only the MI labels. A mapped absorbing output can become correct when its gold tool is the chosen target. This occurs in 356 cells and would lower rescored regret by at most 0.0625. This rescored regret is not used for the reported floor. Thus this is not a standard 16-way table with both MI and regret rescored. The weaker original check retains information within error labels. The collapsed check removes some of it. Neither check is independent evidence for the theorem.

The paired gold-stratified bootstrap uses $B = 400$ as in Appendix 4. The reported corner statistic is $\widehat{I}^{\mathrm{lo}}_{\mathrm{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon}^{\mathrm{hi}})$, which is non-negative in all 441 collapsed cells. Since the floor decreases with regret, this is not the minimum margin over the CI rectangle. That minimum would use $\widehat{\varepsilon}^{\mathrm{lo}}$. The existing figures retain their original corner display. Its non-negative value must not be read as showing that every point of each rectangle is above the floor.

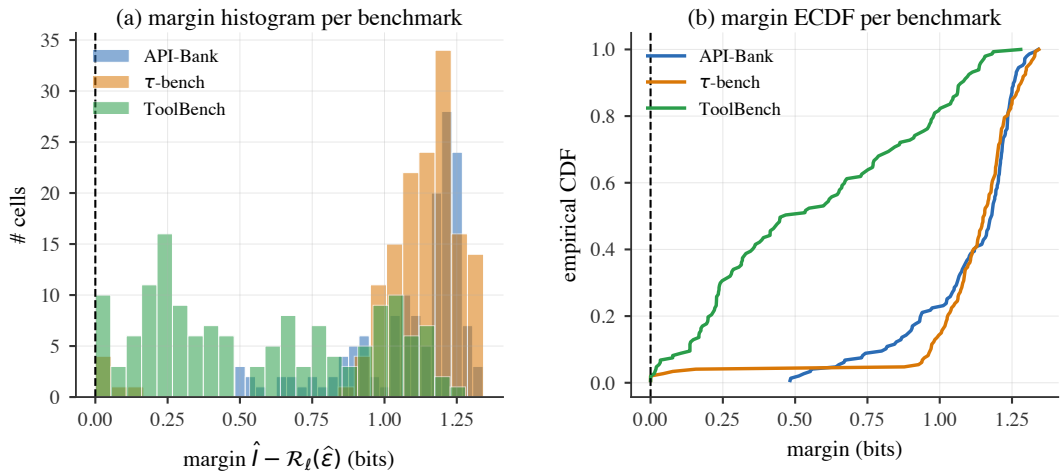

Figure 7. Margin distribution and ECDF over the 441-cell main grid. *(Left)* histogram of per-cell margins $\widehat{I}_{\mathrm{MM}} - \mathcal{R}_\ell(\widehat{\varepsilon})$, split by benchmark. API-BANK has the largest median margin (mass concentrated near $+1.2$ b). τ-bench is unimodal near $+1.1$ b, and ToolBench is heavy-tailed to the left, dominated by saturation cells at $N{=}512$ where both $\widehat{I}_{\mathrm{MM}}$ and $\mathcal{R}_\ell$ collapse to zero. *(Right)* the ECDF of the same margins, and no benchmark places any mass below the floor.

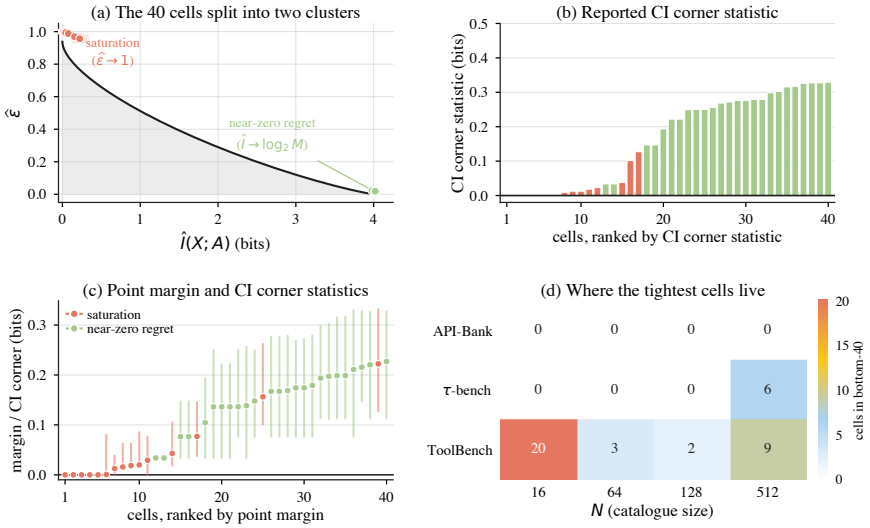

Figure 8. The 40 smallest-margin cells of the 441-cell grid. *(a)* Two clusters on the rate-distortion plane, namely saturation ($\widehat{\varepsilon}{\to}1$, both sides $\approx$0) and near-zero regret ($\widehat{I}_{\mathrm{MM}}{\to}\log_2 M$). *(b)* Reported CI corner statistic sorted, using the upper regret endpoint as described above. *(c)* Point margin and marginal-CI corner statistics per cell. *(d)* The tightest cells concentrate in ToolBench at $N{\in}\{16, 512\}$.

7 Experiment 2: Matched-rate regret clustering

Experiment 2 records an empirical regularity expected when cells lie close to the achievability frontier, namely that pairs of cells with similar action-side rates $\widehat{I}_{\mathrm{MM}}$ tend to achieve similar regrets. (The converse, being one-sided, does not directly imply this. We report the regularity as a near-frontier diagnostic, not a theorem prediction.) The main table (Table 1) reports aggregate overlap rates. Here we break the comparison down geometrically and per (model, bench).

Geometry. Figure 9 plots $|\Delta\widehat{\varepsilon}|$ against $|\Delta\widehat{I}_{\mathrm{MM}}|$ for the 1,658 unordered pairs with $|\Delta\widehat{I}_{\mathrm{MM}}| \le 0.50$ bits within fixed (model, bench, N) slices of the 432 no-CoT main-set cells. The 629 pairs to the left of the 0.10-bit cutoff form the loose matched-rate subset. Points cluster tightly near $\Delta\widehat{\varepsilon}{=}0$ as $|\Delta\widehat{I}_{\mathrm{MM}}| \to 0$, with the slope steepening continuously and the large-regret-gap region at small $|\Delta\widehat{I}_{\mathrm{MM}}|$ empty. That region is feasible under classification mechanics. Pairs whose CIs do not overlap (red) concentrate at the high-$|\Delta\widehat{\varepsilon}|$ end of the scatter rather than scattering at random. The

model	bench	N	p	τ	$\widehat{\varepsilon}$	$\widehat{I}$	$\mathcal{R}_\ell$	margin (b)
Gemma3-27B	ToolBench	512	BEG	T0	1.000	0.000	0.000	+0.000
Gemma3-27B	τ-bench	512	MID	T0	1.000	0.000	0.000	+0.000
Gemma3-27B	τ-bench	512	END	T0	1.000	0.000	0.000	+0.000
Gemma3-27B	τ-bench	512	BEG	T0	1.000	0.000	0.000	+0.000
Gemma3-27B	ToolBench	512	MID	T0	1.000	0.000	0.000	+0.000
GLM-4-32B	ToolBench	512	END	T0	1.000	0.001	0.000	+0.001
Gemma3-27B	ToolBench	512	END	T0	1.000	0.013	0.000	+0.013
Qwen3-32B	ToolBench	512	END	T0	1.000	0.016	0.000	+0.016
Qwen3-32B	ToolBench	512	MID	T0	1.000	0.019	0.000	+0.019
GLM-4-32B	ToolBench	512	MID	T0	1.000	0.020	0.000	+0.020
Qwen3-32B	ToolBench	512	BEG	T0	0.997	0.029	0.000	+0.029
Gemma3-27B	ToolBench	16	END	T0	0.000	4.034	4.000	+0.034
Gemma3-27B	ToolBench	64	BEG	T1	0.000	4.034	4.000	+0.034
Qwen3-32B	τ-bench	512	MID	T0	0.994	0.043	0.000	+0.043
Gemma3-27B	ToolBench	16	BEG	T1	0.003	4.034	3.957	+0.077
GLM-4-32B	ToolBench	16	BEG	T0	0.003	4.034	3.957	+0.077
Qwen3-32B	τ-bench	512	BEG	T0	0.988	0.077	0.000	+0.077
Gemma3-27B	ToolBench	128	BEG	C1	0.006	4.025	3.921	+0.104
Gemma3-27B	ToolBench	16	MID	T0	0.009	4.023	3.887	+0.136
Gemma3-27B	ToolBench	16	BEG	T2	0.009	4.023	3.887	+0.136
GLM-4-32B	ToolBench	16	BEG	T2	0.009	4.023	3.887	+0.136
Qwen3-32B	ToolBench	16	BEG	C1	0.009	4.023	3.887	+0.136
GLM-4-32B	ToolBench	16	END	T2	0.009	4.025	3.887	+0.139
GLM-4-32B	ToolBench	16	END	T0	0.012	4.002	3.854	+0.148
Qwen3-32B	τ-bench	512	END	T0	0.969	0.156	0.000	+0.156
GLM-4-32B	ToolBench	16	BEG	T1	0.012	4.021	3.854	+0.167
GLM-4-32B	ToolBench	16	BEG	C1	0.012	4.021	3.854	+0.167
Gemma3-27B	ToolBench	16	END	C1	0.012	4.023	3.854	+0.169
Gemma3-27B	ToolBench	16	END	T2	0.016	3.997	3.823	+0.174
Gemma3-27B	ToolBench	16	MID	C1	0.016	3.997	3.823	+0.174

Table 12. Smallest-margin 30 cells. Saturation cells reach the floor only because both sides of the inequality have collapsed to $\widehat{I}_{\mathrm{MM}} \approx 0, \mathcal{R}_\ell \approx 0$.

small absolute gaps provide evidence beyond CI overlap. Overlap alone cannot establish equality of underlying regret.

Breakdown by (model, bench). Table 14 reports overlap rates per (model, bench) slice. Of 9 slices, 7 have 100% overlap at the loose threshold ($|\Delta\widehat{I}_{\mathrm{MM}}| \leq 0.10$ b) and $9/9$ exceed 87%. The lowest overlap (87.4%, Gemma3 / ToolBench) coincides with the slice that contains the heaviest saturation tail and the widest spread of regret values.

Confusion-matrix null comparison. Uniform 16-way classification allows a wide regret range at fixed rate. The symmetric channel in Theorem 8 has $I(X, A) = 2$ bits at regret about 0.29. For the other endpoint, let $A = X + J$ modulo 16, where J is uniform on four nonzero shifts. Then $H(A) = 4$ and $H(A \mid X) = 2$, so the rate is 2 bits and regret is 1. For any error e between these endpoints, mix the two noise laws to get error e and rate at most 2, by convexity of MI. A second channel that outputs X with probability $1 - e$ and $X + 1$ modulo 16 otherwise has the same error and rate $4 - h_2(e) \geq 3$. Mix these two channels. Error remains e, and continuity of MI gives a mixture with rate exactly 2. Thus the feasible gap is about 0.7 at exactly matched rate.

Comparison	Regret range	Regret gap
Uniform mechanics at 2 bits	$[0.29, 1]$	about 0.71 allowed
Observed 629 pairs with $\lvert\Delta\widehat{I}_{\mathrm{MM}}\rvert \leq 0.10$		0.0125 median
Observed pairs with $\lvert\Delta\widehat{I}_{\mathrm{MM}}\rvert \leq 0.05$		0.0063 median

Table 13. Text and table null comparison for the unchanged Experiment 2 scatter. The theoretical range is at exactly 2 bits, while the observed rows pool matched-rate pairs across rates. This is a feasible-range comparison, not a sampled null distribution or a significance test.

Relation to main claims. The observed small gaps are not forced by balance or the confusion-matrix constraint. They support Experiment 2 as a companion empirical regularity. CI overlap is a consistency statistic rather than proof of equal underlying regret. The lowest overlap slice is Gemma3 / ToolBench at 87.4%, rising to 88.2% at the tighter threshold. Saturation and estimation noise can affect matching there. The data do not isolate their relative effects. Large regret differences at the same rate would be allowed by the converse.

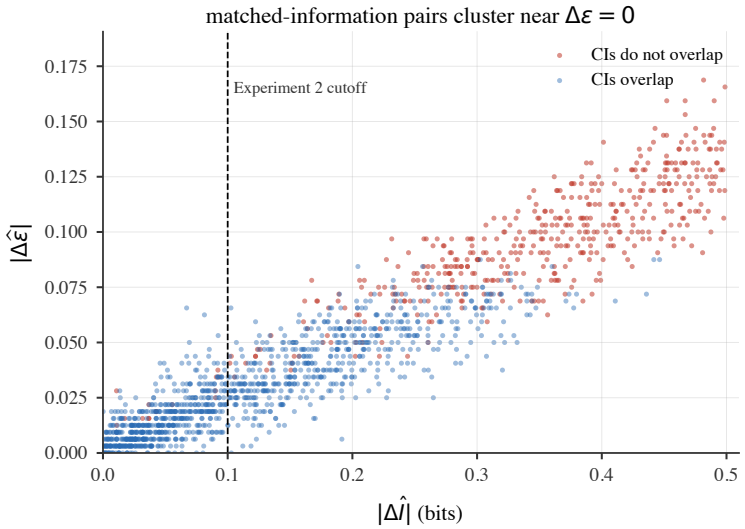

Figure 9. Experiment 2 pair scatter ($|\Delta\widehat{I}_{\mathrm{MM}}|, |\Delta\widehat{\varepsilon}|$) over 1,658 no-CoT pairs with $|\Delta\widehat{I}_{\mathrm{MM}}| \leq 0.50$ bits. Each point is one unordered cell pair within a fixed (model, bench, N) slice. The dashed 0.10-bit cutoff selects the 629 loose matched-rate pairs. Points cluster near $\Delta\widehat{\varepsilon}{=}0$ as $|\Delta\widehat{I}_{\mathrm{MM}}| \to 0$, the behaviour expected near the achievability frontier. The large-regret-gap region at small rate gap is empty in this dataset, but is feasible under the null comparison in Table 13. CI overlap alone does not prove equal underlying regret.

		loose ($\leq$ 0.10 b)			strict ($\leq$ 0.05 b)		
model	bench	n	$\cap$	overlap	n	$\cap$	overlap
Qwen3-32B	API-BANK	59	59	100.0%	27	27	100.0%
Qwen3-32B	τ-bench	76	76	100.0%	49	49	100.0%
Qwen3-32B	ToolBench	94	94	100.0%	56	56	100.0%
Gemma3-27B	API-BANK	68	68	100.0%	35	35	100.0%
Gemma3-27B	τ-bench	51	51	100.0%	26	26	100.0%
Gemma3-27B	ToolBench	103	90	87.4%	85	75	88.2%
GLM-4-32B	API-BANK	47	47	100.0%	19	19	100.0%
GLM-4-32B	τ-bench	51	51	100.0%	26	26	100.0%
GLM-4-32B	ToolBench	80	77	96.2%	48	48	100.0%

Table 14. Experiment 2 overlap broken down by (model, bench). The minimum overlap rate is 87.4% at the 0.10-bit threshold. CI widths and saturation affect this statistic.

8 Experiment 3: CoT shifts

The main paper's CoT figure (Figure 3) summarises Experiment 3 across four panels: arrows on the rate-distortion plane (a), per-cell regret and information shifts (b, c), and distance-to-floor margins (d) over all 12 matched cells. Table 15 reports the corresponding numerical results. These 12 supplement runs are independent of the nine main-set CoT runs in Appendix 5; the repeated (Qwen3, API-BANK, $N{=}128$, MID, T0) setting therefore has separate MI estimates in the two sets. Here we report the inference-time cost of the tested CoT setting.

Reasoning length and wall time. Figure 10 reports the mean number of characters inside the ⟨think⟩ block per trial and the mean wall time for each of the 12 matched cells. Reasoning length ranges from about 989 to 1,492 characters per trial. CoT increases wall time in every tested cell, with on/off ratios ranging from about 1.2 to 13.7; these ratios vary across pool sizes and positions.

		Regret $\hat{\varepsilon}$			*Mutual information* $\hat{I}$ (b)			*Derived* (b)	
N	p	off	on	Δ	off	on	Δ	$\mathcal{R}_\ell$	dist†
16	BEG	0.075	0.113	$+0.038$	3.835	3.735	-0.100	3.053	$+0.682$
	MID	0.097	0.106	$+0.009$	3.718	3.723	$+0.005$	3.096	$+0.627$
	END	0.088	0.122	$+0.034$	3.760	3.676	-0.084	2.989	$+0.687$
64	BEG	0.075	0.106	$+0.031$	3.807	3.776	-0.031	3.096	$+0.680$
	MID	0.197	0.247	$+0.050$	3.432	3.297	-0.135	2.229	$+1.068$
	END	0.219	0.247	$+0.028$	3.355	3.290	-0.066	2.229	$+1.060$
128	BEG	0.137	0.169	$+0.031$	3.700	3.565	-0.135	2.686	$+0.879$
	MID	0.241	0.275	$+0.034$	3.237	3.147	-0.090	2.077	$+1.070$
	END	0.206	0.228	$+0.022$	3.365	3.269	-0.096	2.334	$+0.935$
512	BEG	0.159	0.141	-0.019	3.588	3.679	$+0.091$	2.865	$+0.815$
	MID	0.322	0.297	-0.025	2.956	3.080	$+0.124$	1.963	$+1.117$
	END	0.244	0.300	$+0.056$	3.269	3.040	-0.229	1.947	$+1.093$
	mean	**0.172**	**0.196**	$\mathbf{+0.024}$	**3.502**	**3.440**	$\mathbf{-0.062}$	**2.547**	$\mathbf{+0.893}$

Table 15. Empirical regret and MI at each (N, p) cell, CoT off versus on, for the Qwen3 / API-BANK / T0 supplement. Subscripts *off*/*on* denote CoT disabled and enabled. Δ is the on$-$off gap. $\hat{I}$ is the MM estimate of $I(X, A)$ in bits, and $\mathcal{R}_\ell$ is the analytical $M{=}16$ Hamming floor at $\widehat{\varepsilon}_{\text{on}}$. †Distance to the floor at $\widehat{\varepsilon}_{\text{on}}$ in bits per trial. Last row reports the mean over all 12 cells.

The token cost paid for these reasoning chains is non-trivial, yet the mean regret does not fall. Extra reasoning changes information use at the action, rather than proving an information deficit in C.

Token / time / regret trade-off. Table 16 reports the full comparison between CoT off and CoT on along four axes: input tokens, output character counts, reasoning length, and wall time. The ratio of CoT-on to CoT-off output characters is approximately 38 to 76$\times$, with most of the increase concentrated in the $\langle\text{think}\rangle$ block. The mean regret does not improve on this slice. These costs do not locate a bottleneck between H and C.

Detailed per-cell reading of Table 15. Twelve (N, p) cells let us observe the measured directions of change. *(i)* Of the 12 cells, 10 have $\Delta\widehat{\varepsilon}{>}0$ and 2 have $\Delta\widehat{\varepsilon}{<}0$, so CoT slightly *worsens* regret on average ($+0.024$ pooled), an effect that is small in absolute terms but the opposite of the popular "CoT always helps" folk theory. *(ii)* The $\Delta\widehat{I}_{\text{MM}}$ column has the opposite pooled sign (-0.062 b on average): on average the rate decreases while regret increases. Every CoT-on cell still lands above the $\mathcal{R}_\ell$ floor at its own $\widehat{\varepsilon}_{\text{on}}$, with mean distance $+0.893$ b. The converse itself does not predict this co-movement (it is one-sided), but near the frontier rate and regret should track, which is what we observe. *(iii)* The two cells with $\Delta\widehat{\varepsilon}{<}0$ ($N{=}512$, BEG and $N{=}512$, MID) both have increased $\widehat{I}_{\text{MM}}$. A third cell, $N{=}16$, MID, has $\Delta\widehat{I}_{\text{MM}} = +0.005$ b together with $\Delta\widehat{\varepsilon} = +0.009$, so the rate–regret directions do not align in every cell. The absence of below-floor points is part of the same Experiment 1 audit, not separate support for Experiment 3.

Relation to main claims. The supplement measures the cost of the small mean shifts in Experiment 3. It uses up to 1,492 reasoning characters per trial, with wall-time ratios of about 1.2 to 13.7 across cells. Input tokens decrease by about 0.6% to 15.6% per cell, while output characters grow approximately 38 to 76 times. This shows that the tested reasoning setting is costly. It does not rule out gains from a different CoT policy, more compute, or another benchmark. We leave those extensions to future work.

9 Lost-in-the-Middle Heatmap

Position changes can alter how a decoder uses an otherwise complete prompt [24]. The heatmap reports regret, and the rate panels report action-side MI. When the full prompt is retained, moving an entry need not change syntactic $I(X, C)$ at all. The measurements therefore do not establish a position-induced drop in C_{ctx}.

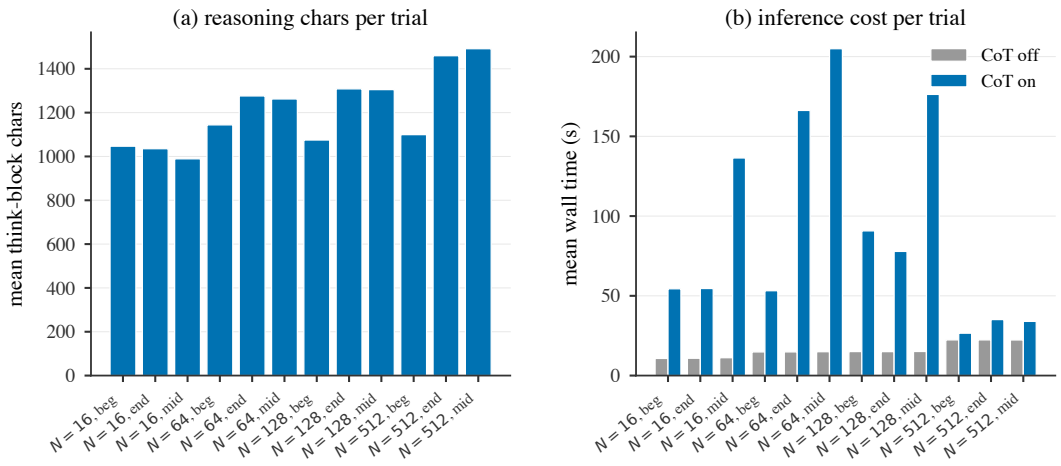

Figure 10. Inference cost of CoT across the 12 matched cells at (Qwen3, API-BANK, T0). *(Left)* mean characters in the ⟨think⟩ block per trial as a function of (N, p), ranging from about 989 to 1,492 characters. *(Right)* mean wall time per trial. CoT increases wall time by factors of about 1.2 to 13.7 across cells, with no mean improvement in regret (Figure 3).

		input toks		output chars			wall time (s)	
N	p	off	on	off	on	think chars	off	on
16	BEG	792	693	31	1217	1047	10.7	54.4
16	END	796	696	31	1202	1035	10.9	54.6
16	MID	823	695	31	1752	989	11.2	136.4
64	BEG	2029	1966	31	1193	1144	14.8	53.2
64	END	2033	1970	31	2065	1276	14.8	166.3
64	MID	2047	1971	31	2167	1262	15.0	205.0
128	BEG	3726	3666	31	1481	1075	15.1	90.7
128	END	3720	3665	31	1479	1308	15.0	77.9
128	MID	3735	3674	31	2330	1305	15.1	176.3
512	BEG	14089	14003	31	1508	1100	22.4	26.6
512	END	14111	14026	31	2109	1460	22.4	35.1
512	MID	14098	14016	31	2032	1492	22.3	34.0

Table 16. Extended CoT comparison covering input tokens, output character counts, think-block lengths, and wall times for off versus on. CoT costs 1.2–13.7× the wall time across cells and produces ∼1,000 chars of hidden reasoning per trial, with no mean regret gain.

Reading the heatmap. Position effects vary across models and benchmarks. At N=16, for example, τ-bench regret at BEG/MID/END is $0.262/0.400/0.366$ for Gemma3 and $0.419/0.525/0.431$ for GLM-4. At N=512, the MID cell is worst or tied for worst on all three benchmarks. API-Bank middle-position regret ranges from 0.32 to 0.51, while the corresponding values on the harder benchmarks range from 0.775 to 1.00.

Connection to the converse. The heatmap is consistent with the converse *interpretation*, where we read position p as an intervention on the realized history-to-action channel, and observe that $\widehat{\varepsilon}$ tracks $\widehat{I}_{\mathrm{MM}}$ in a way consistent with cells sitting near the achievability frontier. We do not claim the position effect is causally mediated by $I(H, C)$ alone. The converse only forbids below-floor cells, which we never observe.

Relation to main claims. The position effect is an action-side pattern relevant to U. At large pool sizes it can involve both decoder use and truncation. The context audit distinguishes certified full inputs from unresolved capped trials. The heatmap does not by itself separate these causes.

10 Tokens-Versus-Bits Gap, Full Panels

The main paper's Figure 4 shows the input-tokens-per-bit ratio at (T0, MID) as a three-panel bar summary (one panel per benchmark, four bars per panel for $N \in \{16, 64, 128, 512\}$, with bar

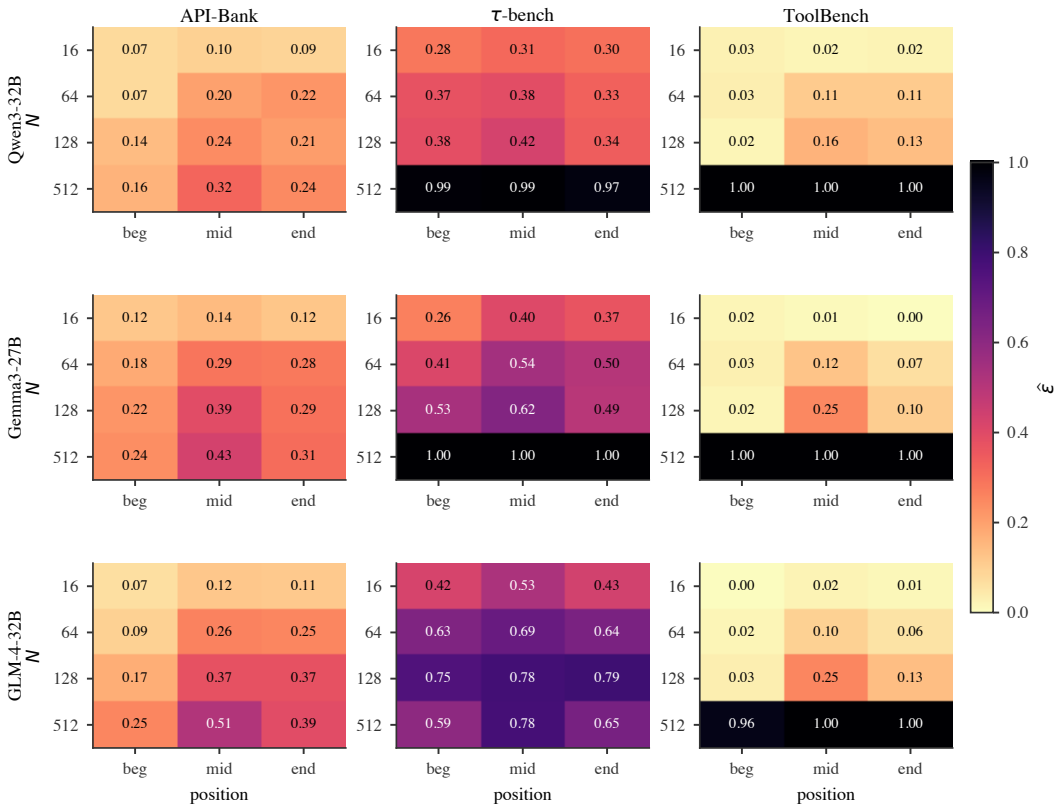

Figure 11. Position effects on empirical regret. Heatmap of empirical regret $\hat{\varepsilon}$ across (position, N) for each of the 9 (model, bench) cells under τ=T0. Some slices already show position differences at N=16. At N=512, the middle position is worst or tied for worst on the harder benchmarks. We interpret the position knob as an intervention on the realized history-to-action channel. The empirical $\hat{\varepsilon}$ tracks the action-side rate $\hat{I}_{\mathrm{MM}}$, consistent with cells sitting near the achievability frontier of the converse.

heights pooled as cross-model medians). Here we disaggregate across all four interventions and three positions.

$\hat{I}_{\mathrm{MM}}$ vs N across all interventions and positions. Figure 12 presents a 3×3 panel grid (one panel per (model, bench) cell) with 4 interventions and 3 positions plotted on each panel. The panels show three patterns. (i) At N=16, rates vary across models, benchmarks, interventions, and positions. ToolBench rates cluster near $H(X)$=4 b, with some unclipped MM estimates slightly above 4 b. (ii) As N grows the curves often fan out, with T0 (verbatim) generally higher and T2 (quarter-length text) often lower in the unsaturated regime. On ToolBench at N=512, T0 hits the input cap and *collapses below* the compressed interventions. (iii) Position effects vary across panels and can already appear at N=16, consistent with the heatmap in Appendix 9.

Token cost vs N. Figure 13 shows mean input tokens at (T0, MID) as a function of N. The 32,768-token input cap is hit by all three models on ToolBench at N=512. On τ-bench at N=512, all Gemma3 trials, most Qwen3 trials, and some GLM-4 trials hit the cap; their mean token counts need not equal the cap. The cap acts as a hard ceiling, since once it is reached, further content may be truncated. The logs alone do not establish which target entries survive. Low $\hat{I}_{\mathrm{MM}}$ measures action-side collapse.

Relation to main claims. Figure 12 and Figure 13 *support* T (the tokens-versus-bits gap) and *refine* the three-panel summary of Figure 4 along two axes the main figure suppresses. The first refinement is *intervention*. T0 is highest in the unsaturated regime but *collapses below* all three compressed interventions on ToolBench at N=512, exactly when the input cap forces a silent

truncation. The second is *position*. Position differences vary across panels and can already appear at $N{=}16$, consistent with the lost-in-the-middle picture of Appendix 9. Neither refinement contradicts the headline statement that tokens are not bits. Both make the case stronger by showing where in the operational map the gap opens. The saturation markers also have high regret. Context audits are needed before attributing these markers to missing prompt information.

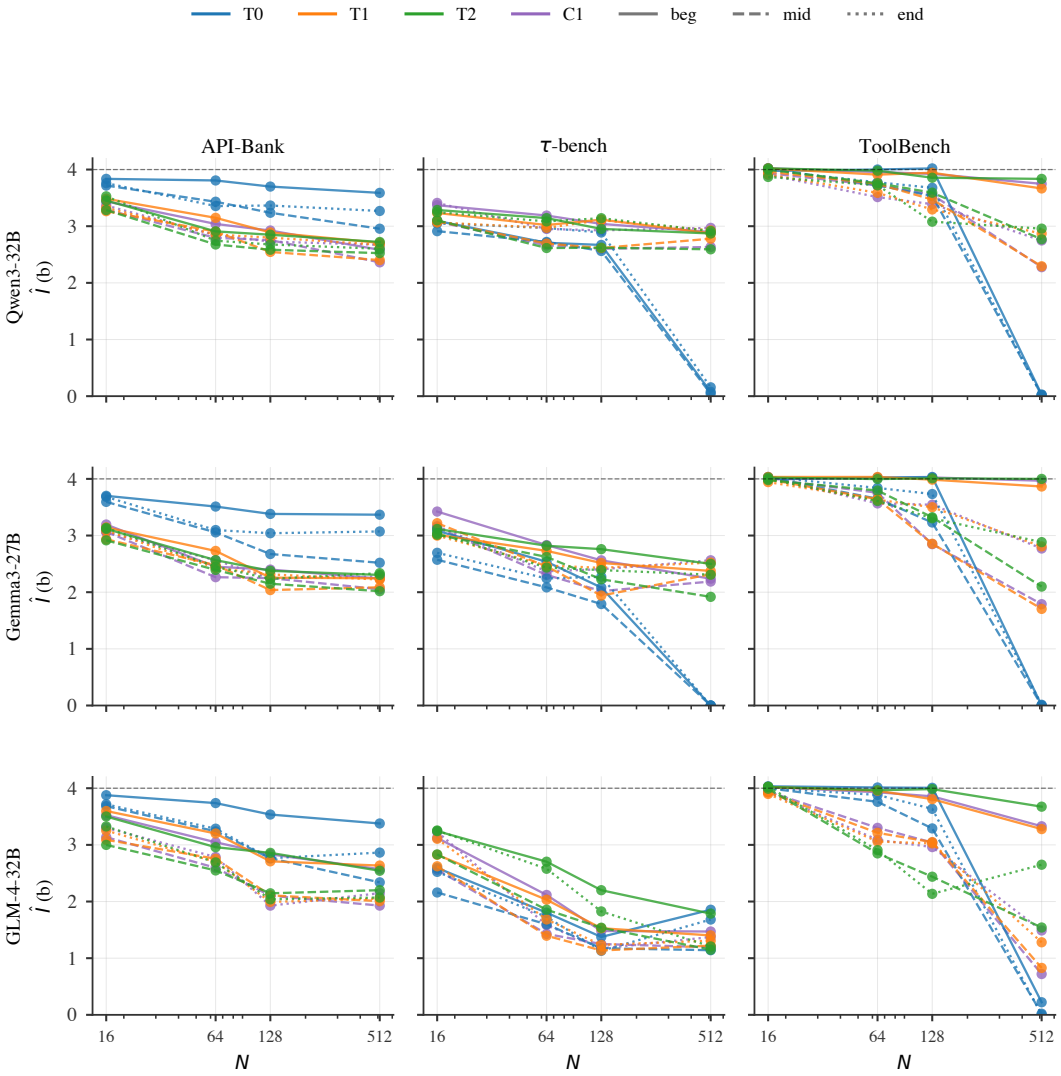

Figure 12. Mutual information $\widehat{I}_{\mathrm{MM}}$ versus pool size N across the full grid. A 3×3 panel grid (one panel per (model, bench) pair) plots $\widehat{I}_{\mathrm{MM}}$ as a function of N under all four compression interventions and three positions. Information *collapses* toward zero at $N{=}512$ on τ-bench and ToolBench, where the verbatim prompt can exceed the input cap. At $N{=}16$, rates vary across panels; ToolBench rates cluster near $H(X){=}4$ b, with some unclipped MM estimates slightly above 4 b. Position differences can already appear at this pool size. The curves often fan out as N grows, with T0 generally higher in the unsaturated regime and collapsing below the compressed interventions once truncation kicks in.

11 Compression Dose Response

The headline table reports compression $\tau \in \{\mathrm{T1}, \mathrm{T2}, \mathrm{C1}\}$ at the single point $(N{=}128, \textsc{mid})$. The full dose response across N is more revealing.

Two regimes. Figure 14 plots $\widehat{\varepsilon}$ by compression intervention (T0, T1, T2, C1) at MID, separately for each (model, bench) cell and each N. These categorical codes are not a strict severity ordering. Two regimes appear. **(i)** In uncapped API-Bank cells, compression generally raises regret and lowers action-side rate. The changes are not monotone across codes, and other benchmark slices include improvements even at $N \leq 128$. **(ii)** In the *saturation* regime (capped T0 cells on ToolBench and

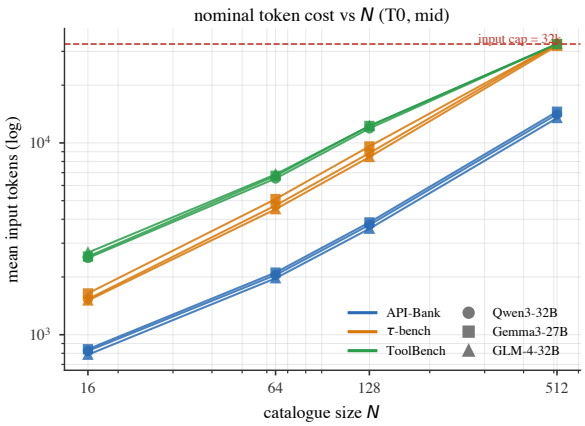

Figure 13. Token cost as a function of pool size at (T0, MID). Mean input tokens versus N for each of the 9 (model, bench) cells. The 32,768-token cap is hit by all three models on ToolBench at $N{=}512$. On τ-bench at $N{=}512$, all Gemma3 trials, most Qwen3 trials, and some GLM-4 trials hit the cap; this differs from a mean token count reaching the cap. Once the cap is reached, further content may be truncated. These token counts do not show which target entries survive, and $\widehat{I}_{\mathrm{MM}}$ describes the action side.

τ-bench at $N{=}512$), compression can lower regret and raise action-side rate by allowing more catalogue content to fit. The context audit is needed to assess retention in these cells.

Why compression sometimes helps. The two pipeline views are $H \to$ assembled catalogue $\to$ compressed catalogue $\to A$ and $H \to$ assembled catalogue $\to$ truncated prompt $\to A$. The query is included in each context state. Compression can keep the prompt within the window, while truncation can remove content. When both prompts fit, compression may instead make retained information harder for the decoder to use. The measured action-side rates do not isolate these mechanisms. Corollary 6 supplies a necessary budget at each cut, rather than a monotone prediction for regret under compression.

Relation to main claims. The two observed directions refine the headline compression result. On uncapped API-Bank cells, compression reduces action-side rate and raises regret. In capped T0 cells, compression can improve both. This motivates context retention checks and decoder checks together. It is not a dedicated retrieve-then-select experiment.

12 Failure Modes and Confusion Matrices

The label space contains 18 outcomes, namely 16 core tools, the OTHER_DISTRACTOR absorbing label (the model called a non-core tool), and the INVALID absorbing label (the output was not parseable as a tool call). Failure modes are qualitatively different and trace back to different mechanisms.

Four representative cells. Figure 15 shows the row-normalised confusion matrix for four cells. **(a)** A high-accuracy cell (Qwen3 / ToolBench / $N{=}16$ / BEG / T0), where the diagonal dominates with negligible mass in the OTHER / INVALID columns. **(b)** A cell with frequent distractor errors (Gemma3 / API-Bank / $N{=}128$ / MID / T2), where mass spreads onto OTHER because the compressed prompt loses the cues that distinguish similar tools, but INVALID remains low, that is, the model emits a *plausible* but wrong tool call. **(c)** CoT on (Qwen3 / API-Bank / $N{=}128$ / MID / T0), where the diagonal weakens slightly relative to T0 without CoT, with some leakage to OTHER. **(d)** A high-INVALID cell (GLM-4 / τ-bench / $N{=}128$ / MID / T0), where mass concentrates in the INVALID column, indicating that the parser frequently fails to extract a tool call from the model output entirely.

Mode mix interpretation. The absorbing labels can retain correlations with X even when every such output is scored as wrong. Their collapse in Appendix 6 removes this information. A concentration on INVALID may reflect output format, parsing, decoder use, or lost input. The confusion matrix alone does not choose among these causes.

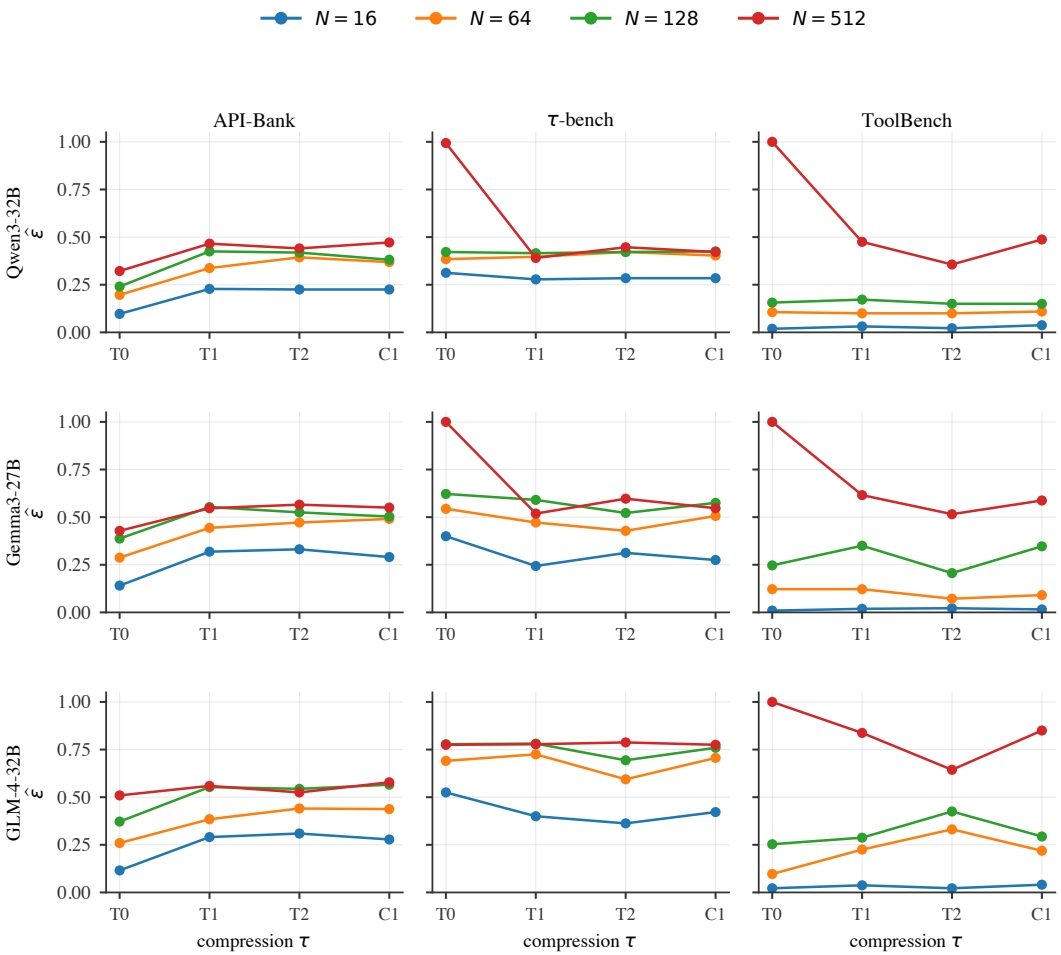

Figure 14. Compression intervention comparison. Regret $\widehat{\varepsilon}$ by intervention code $\tau \in \{\mathrm{T0}, \mathrm{T1}, \mathrm{T2}, \mathrm{C1}\}$ at p=MID, for each of the 9 (model, bench) cells and each pool size N. Compression generally raises regret in uncapped API-Bank cells, with non-monotone differences across codes and exceptions on other benchmarks. In capped T0 cells on ToolBench and τ-bench at N=512, compression can lower regret by allowing more catalogue content to fit.

Relation to main claims. The panels describe output failures for the Experiment 1 audit and motivate U. They do not establish that a cell needs only decoder changes or only context changes. The retained-context columns and the action-side labels answer different questions.

13 Per-Knob Slope Visualisation

For each (model, bench, intervention, position) cell we fit a least-squares line $\widehat{\varepsilon} = \alpha + \beta \log_2 N$ over the four pool sizes $N \in \{16, 64, 128, 512\}$. The slope β measures how fast regret grows per doubling of catalogue size at a fixed compression and position. The coefficient R^2 measures how well the linear fit captures the N-dependence. Across the 108 fits, the slope is positive in every cell. This describes an empirical relation between pool size and regret. It is not a direct measurement of C_{ctx}. The fit summarises the measured range. We do not validate prediction at new pool sizes or outside this range.

What the figure shows at a glance. On API-BANK, the panel that hugs the leftmost band, slopes lie in $[+0.019, +0.080]$ across all three models, so the benchmark is easy enough that even the verbatim T0 sweep does not produce dramatic N-dependence. On τ-bench, slopes spread over $[+0.006, +0.147]$, with the largest bars reserved for T0 where the prompt is most token-heavy and approaches the input cap. ToolBench hosts the most extreme bars (β up to $+0.195$, corresponding to an increase of nearly 0.2 in $\widehat{\varepsilon}$ per doubling of N) and also the only near-zero slopes, concentrated at compressions C1, T2 and T1 with BEG insertion (slopes $+0.001$, $+0.002$ and $+0.013$ respectively for Gemma3-27B), where the prompt is short enough that no saturation regime is reached. The grid describes how regret

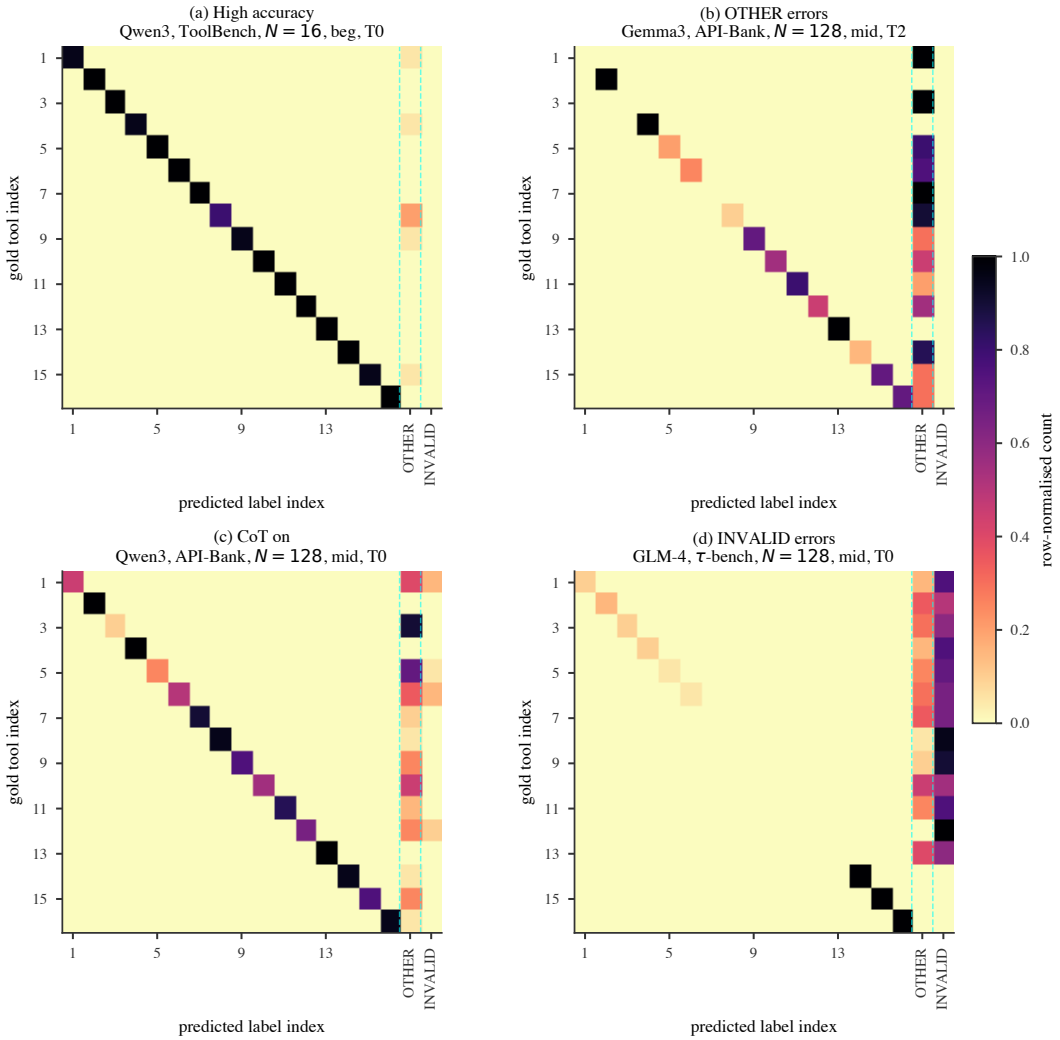

Figure 15. Row-normalised 16×18 confusion matrices for four representative cells. The right two columns are the absorbing labels OTHER_DISTRACTOR and INVALID. *(a) High-accuracy cell* (Qwen3 / ToolBench / N=16 / BEG / T0). The diagonal dominates with negligible mass on the absorbing labels. *(b) Frequent distractor errors* (Gemma3 / API-Bank / N=128 / MID / T2). Mass spreads onto OTHER as the compressed prompt loses cues that distinguish similar tools. *(c) CoT on* (Qwen3 / API-Bank / N=128 / MID / T0). The diagonal weakens slightly relative to the same cell without CoT. *(d) High-*INVALID *cell* (GLM-4 / τ-bench / N=128 / MID / T0). Mass concentrates in the INVALID column, indicating that the parser frequently fails to extract a tool call from the model output entirely.

changes with pool size under each position and compression setting. The figure makes cross-cell comparisons visual.

Relation to main claims. The slope grid gives a descriptive summary of pool-size effects. It complements the action-side rate panels, but positive slopes do not prove a decrease in context capacity. Most R^2 values exceed 0.7. Weak fits also occur without saturation. The per-knob values expand the summary in Table 2.

14 Limitations

1. *Bayes regret requires an oracle.* Theorem 5 uses the Bayes regret distortion. In practice, $V^\star$ may not be known. In our M=16 instantiation the gold tool supplies the optimum under zero-one utility, as defined by the benchmark annotations [23, 34, 55]. For open-ended generation, an auditable surrogate value function and its own optimum must define a non-negative surrogate regret; the resulting budget is conditional on that surrogate. A strong baseline alone need not supply the Bayes optimum. Lemma 11 preserves the rate-distortion optimum when utility-

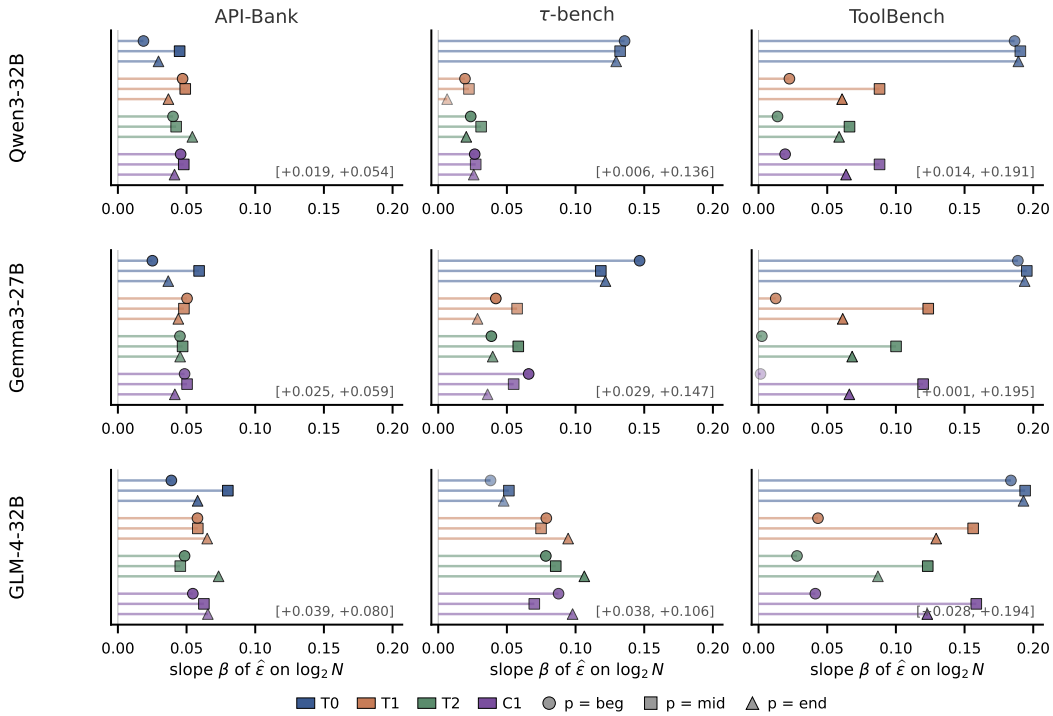

Figure 16. Per-knob slope grid. 108 least-squares fits of $\widehat{\varepsilon} = \alpha + \beta \log_2 N$, one per (model, bench, intervention, position) cell. Rows are models (Qwen3-32B / Gemma3-27B / GLM-4-32B), columns are benchmarks (API-BANK / τ-bench / ToolBench). Within each panel, the four rows correspond to the four compression levels $\tau \in \{\text{T0}, \text{T1}, \text{T2}, \text{C1}\}$ (intervention colour) and three sub-rows per compression encode the insertion position $p \in \{\text{BEG}, \text{MID}, \text{END}\}$ (marker shape). Marker opacity grows with R^2 so the most reliable fits read darkest. Per-panel slope ranges are annotated in the bottom right corner. Two patterns are immediate. **(i)** every bar points to the right, so all fitted slopes are positive; this does not imply strictly increasing regret between every adjacent pair of pool sizes. **(ii)** the largest slopes occur on the harder benchmarks (τ-bench, ToolBench) under verbatim T0 prompts, where the bar reaches $\beta \approx +0.20$ per doubling, the regime in which the input cap forces the steepest rate collapse.

equivalent actions are quotiented, but does not resolve unknown utility. The class count gives an upper bound $\log_2 K_\sim$ on the quotient rate, rather than a general closed-form distortion curve.

2. *Plug-in MI bias is not strictly signed.* We apply the MM correction to each plug-in *entropy* [29], but combining three corrected entropies is not an exact debiasing of plug-in MI. The residual MI bias depends on the joint distribution and is not strictly upward (see Appendix 4). In the $|\mathcal{A}| = 18$ setup the plug-in estimator is tractable, and our Experiment 1 conclusion holds at the point estimate *and* after clipping $\widehat{I}_{\text{MM}} \leq H(X) = 4$ b. For high-dimensional or rare-event actions, variational lower bounds [6, 33, 48] should replace plug-in.

3. *Markov assumption.* The bound requires $I(H, A \mid C) = 0$. Multi-read agents that re-fetch H during decoding need a new application of the converse at each decision, with C enlarged to the full read history. The assumption is restrictive only for agents whose decoder bypasses the prompt entirely (e.g. parameter-efficient adapters that read H via auxiliary projections), in which case the converse should be re-derived on the joint $(C, \text{auxiliary})$ pipeline.

4. *Model scope.* Our experiments use three open-source LLMs ($\sim$30B parameters) [14, 35, 46]. Closed-source frontier models [1, 13, 31] may sit at different operating points (in particular, with longer effective contexts than the 32,768-token cap we use), but the converse, being model-agnostic, makes the same prediction at every cell of the larger grid.

5. *Single decision step.* Our experimental task is a single-step tool selection, while real agents emit decision trajectories. The multi-stage form (Corollary 6) covers closed pipelines that add no new observations. Fresh-read loops need decision-by-decision application. Instrumenting per-step $I(H, C_t)$ along a long-horizon harness is left to future work (Section 6).

6. *Benchmark coverage.* The three benchmarks we use cover short, medium, and long tool-description regimes, but real-world agent deployments also include retrieval-augmented question

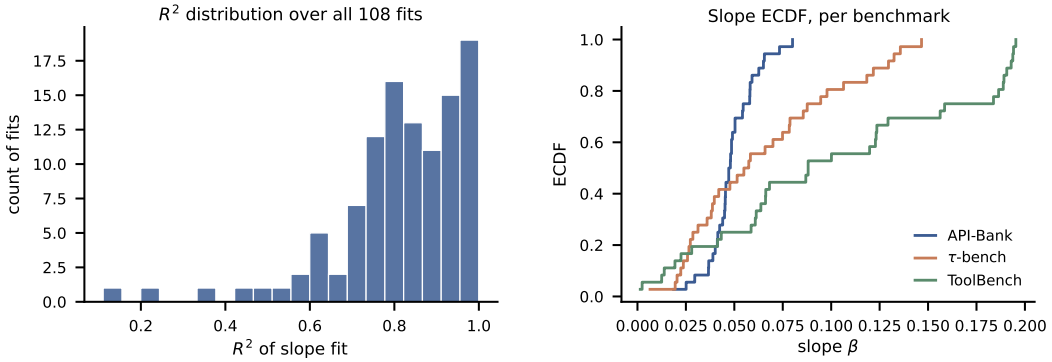

Figure 17. Slope-fit reliability and per-benchmark distribution. *(Left)* Histogram of R^2 across all 108 fits, where the bulk of the mass sits above 0.7. Fits below 0.7 occur mostly on τ-bench, with additional ToolBench and API-Bank cases. *(Right)* ECDF of the slope β split by benchmark. API-BANK hugs the leftmost band ($\beta \in [+0.019, +0.080]$), τ-bench is more dispersed ($[+0.006, +0.147]$), and ToolBench has the heaviest right tail ($[+0.001, +0.195]$), mirroring the absolute regret swing each benchmark exposes.

answering, code-patching environments, and multi-step web navigation [19, 53, 60]. The same DPI argument applies when the context-cut assumptions hold, but estimating $\mathcal{R}_\ell$ for each task requires specifying its prior, loss, and relevant action space. Action cardinality M alone does not determine the curve outside the uniform zero-one setting.

7. *Action quotient and surface-form metrics.* Lemma 11 collapses behaviourally-equivalent actions to a quotient that is well-defined for utility-defined regret. It does not extend to surface-form metrics such as bilingual evaluation understudy (BLEU) or Recall-Oriented Understudy for Gisting Evaluation (ROUGE), where two strings with identical reward can incur different distortion. The converse is therefore not directly operational on generation tasks scored by lexical overlap unless the metric is first replaced by an auditable utility surrogate.

8. *Achievability depends on the decoder and coding model.* For uniform M-way zero-one decisions, converse and achievability meet with a random recommendation and a copy decoder. General finite problems also admit optimal kernels. With shared randomness, one-shot context coding has the conditional-entropy overhead stated in Proposition 14. These results do not guarantee that an actual LLM attains the same budget. The 1.07-bit median margin describes its rate distance from an attainable frontier at the achieved regret, rather than proving which component failed.

9. *Action-side information is not context capacity.* We estimate $I(X, A)$, rather than directly measuring general $I(H, C)$ or C_{ctx}. These experiments alone cannot distinguish missing prompt information from unused available information. The retained-context audit certifies a structural matcher on complete prompts. It does not assess semantic recovery after truncation. Capped trials remain unverified. Our reference matcher also uses the special core-pool construction, so its success should not be assumed for general tool catalogues.

15 Reproducibility

The numerical values in Table 4 and the curves of Figure 5 can be reproduced exactly from Equation (3). We list exact decoding configurations and benchmark composition in Appendix 4. All randomness is fixed. **Compute.** The full $441{+}12$-cell sweep was executed on a single node of $8\times$NVIDIA A100 80 GB SXM4 GPUs, with each open-source LLM served via tensor-parallel batched inference. Counting the main grid together with all bootstrap, pair, slope, CoT, and million-token analyses, as well as preliminary tuning runs that did not enter the reported numbers, the project consumed approximately 4,000 A100-GPU-hours in total, corresponding to roughly 500 wall-clock hours on the 8-GPU node. Peak memory reached about 60 GB per GPU during 32,768-token ToolBench prompts.

NeurIPS Paper Checklist

1. **Claims**

   Question. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer. [Yes]

   Justification. The core contribution is the formal context adequacy problem. Section 5 separates Experiment 1 (Measurement audit), Experiment 2 (Matched-rate regret clustering), and Experiment 3 (CoT shifts). The experiments cover balanced, single-step, 16-way zero-one tool selection. The theory states its own conditions and includes context achievability in Appendix 3.

   Guidelines.

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question. Does the paper discuss the limitations of the work performed by the authors?

   Answer. [Yes]

   Justification. Appendix 14 enumerates nine limitations covering the Bayes regret oracle assumption, the residual finite-sample bias of plug-in mutual information, the Markov assumption between history and action, the open-source $\sim$30B-parameter model scope, the single-decision-step setting, benchmark coverage, the action-quotient restriction to utility-defined regret, and decoder-dependent achievability, and the limits of action-side diagnosis. Each limitation states how violations would affect the claims and where the paper hedges accordingly.

   Guidelines.

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question. For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer. [Yes]

   Justification. Theorem 5 (the regret rate distortion converse), Theorem 8 (the closed form for the M-ary Hamming case), and the supporting Lemma 4 and Corollary 6 state their assumptions. The main text provides complete proofs of the converse, the supporting lemma, and the Hamming formula. Appendix 2 gives the general standard-Borel formulation, the quotient-action proof, and the multi-stage proof. Appendix 3 adds proofs for context achievability, finite kernels, shared-randomness coding, and rational inattention.

   Guidelines.

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question. Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer. [Yes]

   Justification. Appendix 4 documents the 441-cell main grid plus the 12-cell CoT supplement, the benchmark composition (Table 5), the input-token counts (Table 6), the decoding configuration shared across the three models (Table 7), and Qwen3 input-prompt token-count distributions by intervention at (N=128, MID) (Table 8). The Miller-Madow estimator and the stratified bootstrap are specified in the same appendix. Appendix 15 summarises compute resources.

   Guidelines.

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers. Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question. Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer. [No]

   Justification. No supplementary code or data files are provided. The theoretical results and their proofs are included in the paper. The three benchmarks (API-Bank, τ-bench, ToolBench) are publicly available, and Appendix 4 states the benchmark composition and the (N, p, τ) knob settings used in our sweep.

   Guidelines.

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question. Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer. [Yes]

   Justification. We do not train any model. The empirical evaluation runs inference only on three open-source LLMs at fixed weights. Decoding parameters (temperature, sampling, batch size, seed, 32,768-token input cap) are listed in Table 7, and the full (N, p, τ) grid is documented in Table 5 together with per-cell trial counts (n=320). The bootstrap protocol (B=400 resamples) is documented in Appendix 4.

Guidelines.

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question. Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer. [Yes]

   Justification. Each cell reports a 95% stratified bootstrap confidence interval for both empirical regret $\widehat{\varepsilon}$ and Miller-Madow mutual information $\widehat{I}_{\mathrm{MM}}$, with n=320 trials per cell and B=400 resamples stratified by gold tool. The bootstrap procedure is described in Appendix 4, including the observed MI spread of about 0.05 to 0.10 b. Appendix 6, "Absorbing-label robustness", explains the reported CI-corner statistic in Figure 8 panel (b); it is not the minimum margin over the CI rectangle and does not certify that the entire rectangle lies above the floor.

   Guidelines.

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question. For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer. [Yes]

   Justification. All experiments ran on a single node with 8×NVIDIA A100 80 GB SXM4 GPUs, with each open-source LLM served via tensor-parallel batched inference. Counting the main grid together with all bootstrap, pair, slope, CoT, and million-token analyses, as well as preliminary tuning runs that did not enter the reported numbers, the project consumed approximately 4,000 A100-GPU-hours in total, corresponding to roughly 500 wall-clock hours on the 8-GPU node. Peak memory reached about 60 GB per GPU during 32,768-token ToolBench prompts. See Appendix 15.

   Guidelines.

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question. Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer. [Yes]

   Justification. We reviewed the NeurIPS Code of Ethics. The work uses publicly released benchmarks and open-source LLM weights under their respective licenses, runs only inference, and involves no human subjects, no scraping of new data, and no release of new model weights. The contribution is diagnostic in nature.

   Guidelines.

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question. Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer. [Yes]

    Justification. Section 6 discusses broader impact and compute implications. The framework gives engineers a decision budget to use together with context audits and decoder comparisons. The study introduces no new model training. The discussion identifies risks from estimation error, incomplete context logs, and treating a necessary budget as a capability guarantee, and recommends reporting the target loss, measured channel, uncertainty, and audit limits. It also distinguishes the observed token-per-bit gap from evidence that removing tokens saves compute at fixed regret.

    Guidelines.

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question. Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer. [N/A]

    Justification. We release no new model weights and no scraped or generated dataset.

    Guidelines.

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question. Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer. [Yes]

    Justification. All three model families are cited at first use (Qwen3 [35], Gemma3 [14], GLM-4 [46]), and so are the three benchmarks (API-Bank [23], τ-bench [55], ToolBench [34]). We use each asset under its public release terms (Apache-2.0 for Qwen3 weights, the MIT licence for GLM-4 weights, the Gemma usage licence for Gemma3, and the original benchmark licences accompanying the respective releases), without modification of weights and without redistribution.

    Guidelines.

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question. Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer. [N/A]

Justification. We do not release new datasets, model weights, or software assets.

Guidelines.

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question. For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer. [N/A]

Justification. The work involves no crowdsourcing and no human-subject research. All gold labels come from the existing public benchmark annotations.

Guidelines.

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question. Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer. [N/A]

Justification. The work involves no human subjects, so no IRB review applies.

Guidelines.

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question. Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer. [Yes]

Justification. The three open-source LLMs (Qwen3-32B [35], Gemma3-27B [14], GLM-4-32B [46]) are the empirical evaluation subjects of the converse, with decoding configurations and prompt construction rules documented in Appendix 4.

Guidelines.

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.