

Why Your Prompt Compressor Is Deleting the Wrong Tokens: An Information-Theoretic Diagnosis

Jinheng Wu

School of Mathematics and Physics
Xi'an Jiaotong-Liverpool University, Suzhou 215123, P. R. China
Department of Mathematical Sciences, University of Liverpool
Liverpool L69 7ZL, United Kingdom
Jinheng.Wu20@student.xjtlu.edu.cn

Di Zhang*

School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University, Suzhou 215123, P. R. China
di.zhang@xjtlu.edu.cn

Abstract

Token-level compression of in-context learning (ICL) demonstrations reduces inference cost but degrades classification accuracy past a task-dependent threshold, irrespective of the selection criterion. We establish a formal information-theoretic basis for this regularity. Modeling ICL as Bayesian posterior inference over a latent label mapping Θ and applying Fano's inequality, we derive a lower bound on K -class error under arbitrary token excision that decomposes into a *task-intrinsic* term invariant to any compression operator and a *mapping-dependent* term that the data-processing inequality prevents from exceeding its full-context value, delimiting an *incompressible information core*. We further diagnose *why* self-information-based selectors systematically delete the wrong tokens: label and delimiter tokens carry disproportionate attention weight yet low self-information, causing widely used criteria to excise precisely the tokens most critical to posterior updating. A complementary sensitivity analysis motivates a severity ordering (flipped $\geq$ arbitrary $\geq$ semantic) across label regimes, governed by the prior-to-posterior Kullback–Leibler divergence each regime must surmount. All theoretical predictions are corroborated across 195 experimental configurations (four models, four benchmarks, five compression strategies, three label regimes).

1 Introduction

Through exposure to a small set of labeled exemplars, large language models (LLMs) adapt to novel tasks at inference time without any modification to their parameters, a regime collectively designated in-context learning (ICL) (Brown et al., 2020; Lampinen et al., 2024). The mechanism levies a token-budget cost disproportionate to the number of demonstrations: serialized exemplar sequences routinely saturate a substantial fraction of the admissible context window, confining deployment to settings with generous memory and latency margins (Abdin et al., 2024). Systematic excision of tokens from the demonstration pool, broadly termed context compression, attenuates this overhead (Jiang et al., 2023b; Li et al., 2023; Pan et al., 2024; Joo & Klabjan, 2025; Li et al., 2025), yet a cross-method empirical regularity persists: discriminative accuracy deteriorates monotonically with the excision fraction, and competing selection criteria converge to comparably degraded performance under aggressive compression ratios. Whether an irreducible, task-contingent informational floor precludes any token-removal strategy from sustaining ICL fidelity had, prior to this

*Corresponding author.

work, resisted theoretical resolution—and the systematic mechanism by which popular criteria select the wrong tokens had likewise gone undiagnosed.

Several bodies of prior research bear directly on this question, yet none resolves it. The Bayesian formalization of ICL due to Xie et al. (2022) recasts the inference procedure as approximate posterior updating over a latent task concept; extensions of this framework have produced finite-sample error certificates (Wakayama & Suzuki, 2025; Zhang et al., 2025) and decomposed aggregate ICL error into task-intrinsic, approximation, and retrieval-noise constituents (Jeon et al., 2024). PAC-learning analyses have bounded the sample complexity sufficient for ICL to succeed (Wies et al., 2023), while investigations of pretraining task diversity have clarified the conditions under which in-context behavior transitions between Bayesian and non-Bayesian regimes (Raventós et al., 2023); the degree of concordance between pretraining and evaluation distributions further modulates out-of-distribution generalization (Letey et al., 2026). Mechanistic accounts have alternatively characterized ICL as implicit gradient descent or kernel regression on pretrained representations (Akyürek et al., 2023). On the empirical side, systematic studies of label sensitivity establish that demonstration correctness contributes less predictive leverage than intuition predicts (Min et al., 2022), that parameter scale changes the balance between semantic priors and input-label mappings (Wei et al., 2023; Shi et al., 2024), and that ICL alternates between task-learning and task-retrieval modes under different distributional conditions (Lin & Lee, 2024; He et al., 2025). Token-salience methods such as LLMingua (Jiang et al., 2023b) and Selective Context (Li et al., 2023) rank tokens by perplexity or self-information, while KV-cache pruning schemes (Zhang et al., 2023; Li et al., 2024) intervene at the attention layer rather than the input sequence. No existing contribution, however, furnishes a provable lower bound on the accuracy penalty incurred by lexical excision from demonstration sequences.

This paper resolves the conjecture through a formal information-theoretic treatment. We model ICL as Bayesian posterior inference over a latent label mapping Θ and apply Fano’s inequality (Fano, 1961; Cover & Thomas, 2006) to derive a lower bound on K -class classification error under arbitrary token excision (Theorem 7, Corollary 8). The bound separates into a task-intrinsic term invariant to any compression operator and a mapping-contingent term that the data-processing inequality prevents from exceeding its full-context value, delimiting an incompressible informational core whose depletion mandates a commensurate rise in error. Within the same framework, a sensitivity analysis motivates a severity ordering (flipped $\geq$ arbitrary $\geq$ semantic) across label regimes, traced to the prior-to-posterior Kullback–Leibler divergence each regime must surmount (Hypothesis 15). We further identify a structural deficiency of surprisal-based selection: self-information $h(t_j)$ diverges from the marginal mapping contribution ΔI_j (Proposition 18), a misalignment rectified by an attention-weight-guided compression operator derived analytically. All predictions are corroborated across 195 experimental configurations (four models, four benchmarks, five compression strategies, three label regimes). To provide additional assurance of correctness, the core information-theoretic identities and bounds have also been formalized in Lean 4 (de Moura & Ullrich, 2021).

2 Theoretical Framework

2.1 Problem Setup

Consider a K -class text-classification task in which a label mapping $\Theta : [K] \rightarrow \mathcal{V}$ is drawn from the model’s implicit prior $P(\Theta)$ over injective functions from class indices to vocabulary tokens. Demonstration pairs $\{(x_i, y_i)\}_{i=1}^n \sim P(X, Y)$ are serialized into a token sequence $\mathbf{t} = (t_1, \dots, t_L)$ constituting the ICL context C . A test query $(X, Y) \sim P(X, Y)$ is sampled independently of C conditional on Θ . The structural conditional independence $Y \perp C \mid X, \Theta$ holds throughout. Unless otherwise stated, all logarithms are base 2 and all information quantities are measured in bits.

Assumption 1 (Bayesian ICL). *Following Xie et al. (2022), the model performs approximate Bayesian inference: $P(\Theta \mid C) \propto P(C \mid \Theta) \cdot P(\Theta)$, and predicts $P(Y = k \mid X, C) = \sum_{\theta} P(Y = k \mid X, \theta) \cdot P(\theta \mid C)$.*

Assumption 2 (Balanced classes). Y is uniform over $[K]$, i.e., $H(Y) = \log K$.

Definition 3 (Context Compression). A compression operator ϕ_r with ratio $r \in [0, 1]$ maps the demonstration token sequence to a subsequence of length $L' = \lfloor (1-r)L \rfloor$. We write $C_r = \phi_r(C)$. The operator acts only on demonstration tokens, not on the test input X , and is deterministic given C .

2.2 Information Decomposition

Proposition 4 (Chain Rule Decomposition). $I(Y; C_r, X) = \underbrace{I(Y; X)}_{\text{task info}} + \underbrace{I(Y; C_r | X)}_{\text{contextual info}}$

Proposition 5 (Contextual Information Bound). Under Assumption 1: $I(Y; C_r | X) \leq I(\Theta; C_r)$.

Proof. By $Y \perp C_r | X, \Theta$, the chain $Y \rightarrow \Theta \rightarrow C_r$ (conditioned on X) gives $I(Y; C_r | X) \leq I(\Theta; C_r | X)$ via DPI. Since $X \perp (\Theta, C_r)$, we have $I(\Theta; C_r | X) = I(\Theta; C_r)$. $\square$

Term (a) quantifies the model’s pretrained discriminative capacity and remains invariant under compression. Term (b) quantifies the supplementary signal extracted from the demonstrations, upper-bounded by the mutual information $I(\Theta; C_r)$ between the retained tokens and the latent mapping.

2.3 Compression Lower Bound via Fano’s Inequality

Definition 6 (Mapping Entropy). $H_\Theta = H(\Theta)$ is the entropy of the label mapping under the model’s prior. For a uniform prior over $K!$ injective mappings: $H_\Theta = \log K!$.

Theorem 7 (Compression Lower Bound). Under Assumptions 1 and 2, for any ϕ_r and any predictor $\hat{Y}(r)$:

$$P_e(r) \geq 1 - \frac{I(Y; X) + I(\Theta; C_r) + 1}{\log K} \quad (1)$$

Proof. Fano’s inequality gives $H(Y | C_r, X) \leq h(P_e) + P_e \log(K-1)$. Relaxing: $h(P_e) \leq 1$, $\log(K-1) \leq \log K$, so $H(Y | C_r, X) \leq 1 + P_e \log K$. With $H(Y) = \log K$: $\log K - I(Y; C_r, X) \leq 1 + P_e \log K$. By Propositions 4–5: $I(Y; C_r, X) \leq I(Y; X) + I(\Theta; C_r)$. Rearranging yields the result. $\square$

Remark (Tightness). Fano’s inequality yields loose absolute bounds but informative ordinal predictions. The nonnegative gap relative to the full context, label-regime ordering, and convergence at extreme compression follow from the *direction* of information loss, not its magnitude. The error increment $LB(r) - LB(0) = [I(\Theta; C) - I(\Theta; C_r)] / \log K$ is governed by the DPI-tight information gap; Section 3 confirms the corresponding empirical trends.

Tightness for $K = 2$. When $K = 2$, $\log(K-1) = 0$ renders the preceding bound vacuous. The following corollary supplies a non-trivial alternative.

Corollary 8 (Binary Classification Bound). When $K = 2$, under Assumptions 1–2:

$$P_e(r) \geq h^{-1}(1 - I(Y; X) - I(\Theta; C_r)) \quad (2)$$

where $h^{-1} : [0, 1] \rightarrow [0, \frac{1}{2}]$ is the inverse binary entropy. An explicit bound: $P_e(r) \geq \frac{1 - \sqrt{1 - \alpha^2}}{2}$ with $\alpha = 1 - I(Y; X) - I(\Theta; C_r)$.

Proof. When $K = 2$, Fano gives $h(P_e) \geq 1 - I(Y; C_r, X) \geq 1 - I(Y; X) - I(\Theta; C_r) =: \alpha$. Since h is increasing on $[0, \frac{1}{2}]$, $P_e \geq h^{-1}(\alpha)$. The explicit bound follows from $h(p) \leq 2\sqrt{p(1-p)}$. $\square$

2.4 Data Processing Inequality and Compression

Proposition 9 (DPI for Compression). *Since $\Theta \rightarrow C \rightarrow C_r$ is a Markov chain: $I(\Theta; C_r) \leq I(\Theta; C)$, with equality iff C_r is sufficient for Θ .*

Corollary 10 (Irreducible Error Increase). *For $r > 0$: $\text{LB}(r) - \text{LB}(0) = \frac{I(\Theta; C) - I(\Theta; C_r)}{\log K} \geq 0$.*

Remark (Irrecoverability across Transformer layers). The Markov chain $\Theta \rightarrow C \rightarrow C_r$ from Proposition 9 propagates through every Transformer layer. Denoting the hidden representations at layers $1, \dots, D$ by $Z_1, Z_2, \dots, Z_D$ and noting that each layer constitutes a deterministic map of its predecessor, the augmented chain

$$\Theta \rightarrow C \rightarrow C_r \rightarrow Z_1 \rightarrow Z_2 \rightarrow \dots \rightarrow Z_D \rightarrow \hat{Y}$$

holds, and iterated DPI yields $I(\Theta; \hat{Y}) \leq I(\Theta; Z_D) \leq \dots \leq I(\Theta; Z_1) \leq I(\Theta; C_r) \leq I(\Theta; C)$. Any information about Θ surrendered at the compression stage is *irrecoverable*: the deficit $I(\Theta; C) - I(\Theta; C_r)$ constitutes a floor on information loss that persists across every subsequent layer. No increase in network depth or attention capacity can recuperate the signal eliminated before the forward pass begins.

2.5 Label Condition Analysis

Definition 11 (Label Conditions). *We consider three label mappings: **semantic** (θ_{sem} , natural labels), **arbitrary** (θ_{arb} , unrelated tokens), and **flipped** (θ_{flip} , reversed labels).*

Assumption 12 (Prior Ordering). $P(\Theta = \theta_{\text{sem}}) > P(\Theta = \theta_{\text{arb}}) > P(\Theta = \theta_{\text{flip}})$.

This ordering reflects corpus-level co-occurrence statistics accrued during pretraining: semantically congruent labels co-occur with their referent categories at substantially higher frequency than arbitrary or inverted tokens. The assumption is directly falsifiable via zero-shot accuracy comparisons (Section 3).

Definition 13 (Compression Sensitivity). $S_L(\theta) = \mathcal{R}(\theta, 0) - \mathcal{R}(\theta, 1/L)$, where $\mathcal{R}(\theta, r) = \mathbb{E}_{C_r|\theta}[D_{\text{KL}}(P(\Theta | C_r) \| P(\Theta))]$ is the conditional information retention.

Assumption 14 (Bounded token evidence). *There exists $B > 0$ such that for any token t and mappings θ, θ' : $|\log \frac{P(t|\theta)}{P(t|\theta')}| \leq B$.*

Hypothesis 15 (Compression Sensitivity Ordering). *Under Assumptions 1, 12, and 14: $S_L(\theta_{\text{flip}}) \geq S_L(\theta_{\text{arb}}) \geq S_L(\theta_{\text{sem}})$.*

The mechanism operates as follows. In the semantic regime, the prior already concentrates near the correct mapping, so the removal of any individual token yields negligible posterior perturbation. In the flipped regime, the prior actively opposes the correct mapping; each token supplies indispensable evidence for traversing the prior-to-posterior divergence, and its elimination inflicts disproportionate damage. A detailed motivation appears in Appendix A.8.

2.6 Compression Method Analysis

Theorem 16 (Method-Independent Lower Bound). *The lower bound depends on ϕ_r only through $I(\Theta; C_r)$. Methods achieving equal $I(\Theta; C_r)$ share the same Fano bound.*

The bound depends on the chosen operator ϕ_r exclusively through $I(\Theta; C_r)$, the mutual information between the latent label mapping and the retained token subsequence. This parameterization raises the question of which compression operators maximize $I(\Theta; C_r)$. To address this, we distinguish two information-theoretic quantities at the token level:

Definition 17 (Token-level vs. Mapping Information). *For token t_j in context C , define: (i) **token-level information** $h(t_j) = -\log P_{\text{LM}}(t_j | t_{<j})$ (the self-information under the language model); and (ii) **mapping contribution** $\Delta I_j = I(\Theta; C) - I(\Theta; C_{-j})$ (the marginal contribution to mapping information).*

Perplexity-based and self-information methods select tokens with large $h(t_j)$, yet elevated surprisal does not guarantee elevated ΔI_j . Structurally pivotal tokens—label words and input-output delimiters—frequently register low surprisal under the autoregressive distribution yet contribute substantially to ΔI_j by anchoring the mapping Θ . Low-frequency content words exhibit the inverse profile: high $h(t_j)$, negligible ΔI_j .

Proposition 18 (Information Misalignment). *Let S_h retain the top- $(1-r)L$ tokens by $h(t_j)$, and let $S^* = \arg \max_{|S|=(1-r)L} I(\Theta; C_S)$ be the mapping-optimal subset. Then $I(\Theta; C_{S^*}) \geq I(\Theta; C_{S_h})$, with equality iff S_h is itself mapping-optimal. In general, the ranking by $h(t_j)$ need not approximate S^* , since h and ΔI can disagree systematically (see Appendix A.10 for examples and proof).*

The first claim follows immediately from the extremal definition of S^* . The substantive result is that $h(t_j)$ gauges token rarity under the autoregressive prior, while ΔI_j gauges relevance to the latent task mapping Θ . These two rankings diverge systematically: structurally pivotal tokens (label words, delimiters) cluster toward low surprisal but high ΔI_j , whereas infrequent content words occupy the opposite region of this two-dimensional space.

Structural operators (head/tail truncation) retain contiguous demonstration blocks, thereby preserving the input-label co-occurrence structure that sustains Bayesian posterior updates. When mapping information distributes across structurally coherent units rather than concentrating in individually surprising tokens, this positional preservation attains higher $I(\Theta; C_r)$ than token-level selection achieves.

Proposition 19 (Convergence at Extreme Compression). *As $r \rightarrow 1$: $I(\Theta; C_r) \rightarrow 0$, so all methods converge to the prior-only error.*

2.7 Testable Predictions

The framework generates five testable predictions, each corroborated in Section 3: **(1)** compression cannot lower the error bound below its full-context value, and empirical error is non-decreasing with r ; **(2)** $P_e^{\text{flip}} \geq P_e^{\text{parb}} \geq P_e^{\text{psem}}$ with widening gaps at intermediate r ; **(3)** higher-cardinality tasks can exhibit greater compression sensitivity when their mapping-information loss grows faster than $\log K$, as tested on AG News ($K = 4$); **(4)** all methods converge as $r \rightarrow 1$; **(5)** token-level methods (PPL, self-info) need not outperform structural methods (head/tail) owing to the misalignment between $h(t_j)$ and ΔI_j (Proposition 18).

3 Experiments

3.1 Setup

Models. Qwen2.5-1.5B and Qwen2.5-7B (Qwen Team, 2024), Mistral-7B-v0.3 (Jiang et al., 2023a), and Gemma-2-2B (Gemma Team, 2024) ($\sim 5\times$ parameter range, three pretraining pipelines).

Tasks. IMDB (Maas et al., 2011) ($K = 2$), SST-2 (Socher et al., 2013) ($K = 2$), AG News (Zhang et al., 2015) ($K = 4$), and RTE (Dagan et al., 2006) ($K = 2$, NLI). Each task uses 8 balanced demonstrations serialized as Text: ... \n Label: ... pairs.

Methods & label conditions. Five compression methods: *structural* (**Head/Tail** truncation) and *token-level* (**PPL** (Jiang et al., 2023b), **Self-Info** (Li et al., 2023), **Random**). Three label regimes (Definition 11): *semantic, arbitrary, flipped*. Compression applies only to demonstration tokens.

Protocol. For IMDB, SST-2, and AG News, we evaluate all $4 \times 3 \times 5 \times 3 = 180$ (model, task, method, label) configurations. For RTE, we evaluate Qwen2.5-7B with all five methods and three label conditions (15 additional configurations; 195 total). Each configuration uses $N = 100$ samples and 11 ratios $r \in \{0, 0.1, \dots, 0.9, 0.95\}$. Stochastic methods use five seeds

Table 1: Full-context accuracy ($r = 0$) across models, tasks, and label conditions. All compression methods yield identical results at $r = 0$.

Model	Task	Semantic	Arbitrary	Flipped
Qwen2.5-1.5B	IMDB	0.93	0.92	0.59
	SST-2	0.96	0.51	0.50
	AG News	0.79	0.23	0.23
Qwen2.5-7B	IMDB	0.93	0.91	0.59
	SST-2	0.98	0.94	0.53
	AG News	0.86	0.50	0.18
Mistral-7B	IMDB	0.93	0.85	0.51
	SST-2	0.97	0.88	0.50
	AG News	0.76	0.46	0.10
Gemma-2-2B	IMDB	0.94	0.88	0.46
	SST-2	0.96	0.70	0.41
	AG News	0.83	0.25	0.19

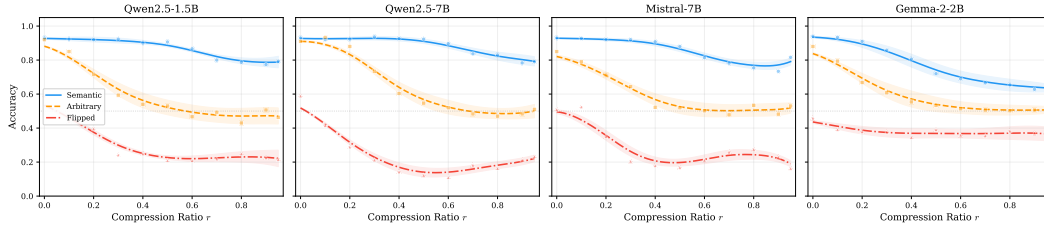


Figure 1: **P1 + Model scale.** Accuracy vs. compression ratio on IMDB (random drop) across four models. Larger models show greater separation between label conditions, consistent with stronger priors amplifying both semantic robustness and flipped fragility.

(mean $\pm$ standard deviation reported). Curves are Gaussian-smoothed for visual clarity, with raw markers overlaid.

3.2 Full-Context Baseline and Prior Verification

Table 1 records full-context ($r=0$) accuracy. The ordering semantic $>$ arbitrary $>$ flipped is sustained across every model-task pair (Assumption 12); the separation widens with scale (AG News: Qwen2.5-7B reaches 0.50 under arbitrary labels versus 0.23 for the 1.5B variant). Mistral-7B and Gemma-2-2B replicate the same hierarchy.

3.3 Prediction Verification

Each of the five predictions from Section 2.7 is evaluated against the complete three-task experimental matrix ($4 \text{ models} \times 3 \text{ tasks} \times 5 \text{ methods} \times 3 \text{ labels} \times 11 \text{ ratios}$), with RTE providing an additional cross-task validation.

P1: Monotonic degradation (Theorem 7). Figure 1 plots accuracy against r for all four models on IMDB under random token removal. Every curve is non-increasing. Scale amplifies the label-condition gap: at $r=0.95$ the 7B model retains semantic accuracy ≥ 0.78 , whereas flipped accuracy collapses to ≤ 0.25 already at $r=0.5$ —a pattern consistent with the stronger priors of larger models (Hypothesis 15).

P2: Label sensitivity ordering (Hypothesis 15). Figure 2 displays all three label regimes (Qwen2.5-7B, random compression). The ordering $\text{Acc}_{\text{sem}} \geq \text{Acc}_{\text{arb}} \geq \text{Acc}_{\text{flip}}$ holds at 94.7% of evaluated checkpoints (677/715); the handful of violations are confined to the near-chance regime.

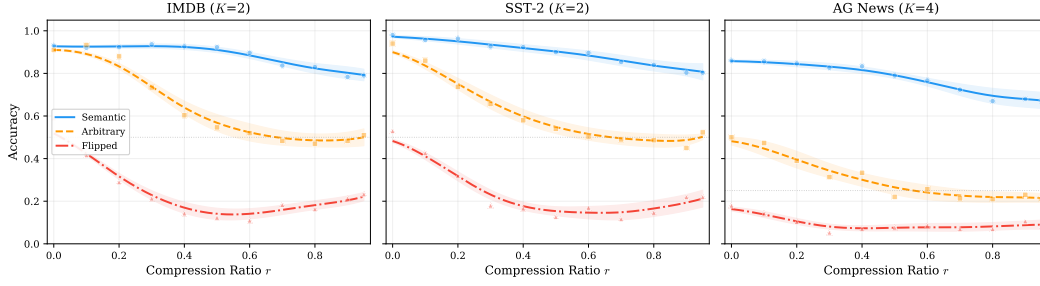


Figure 2: **P2: Label ordering.** Accuracy under three label conditions (Qwen2.5-7B, random compression). The ordering semantic $\geq$ arbitrary $\geq$ flipped is maintained across tasks and compression ratios.

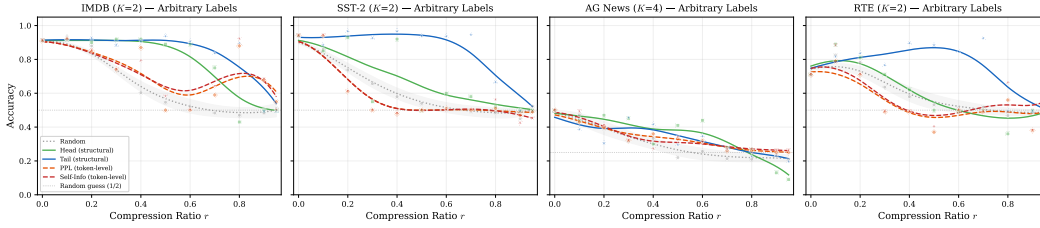


Figure 3: **P3 + P4: Task complexity and convergence.** All methods under arbitrary labels (Qwen2.5-7B) across four tasks, including RTE. AG News ($K=4$) degrades faster than binary tasks (P3). At $r=0.95$, method differences largely vanish (P4). RTE confirms these patterns on a structurally different NLI task.

P3: Task complexity (Theorem 7). Figure 3 overlays all five methods across four tasks under arbitrary labels. AG News ($K=4$) accuracy falls below that of every binary task at $r=0.5$ in 100% of (model, method) comparisons. This empirical result is consistent with a larger mapping-information burden in the higher-cardinality task; the factor $1/\log K$ alone does not imply greater sensitivity unless the information loss also grows with K . RTE exhibits steeper degradation than IMDB/SST-2, attributable to its weaker arbitrary-label prior.

P4: Method convergence (Proposition 19). At $r=0.95$ the inter-method accuracy spread contracts from 0.24 to 0.14 on average, in accordance with $I(\Theta; C_r) \rightarrow 0$. Under arbitrary labels, AG News reaches a spread of ≤ 0.03 ; RTE converges to near-chance performance across all methods.

P5: Information misalignment (Proposition 18). Figure 4 contrasts structural (solid) and token-level (dashed) methods on IMDB and RTE. Structural methods surpass PPL/self-info in 87.2% of comparisons at $r=0.5$ (34/39) and exhibit far lower inter-ratio variance (PPL swings up to 40 pp between adjacent ratios). This instability is a direct manifestation of the misalignment between token-level surprisal $h(t_j)$ and mapping contribution ΔI_j : retaining high-surprisal tokens does not guarantee preservation of the structurally coherent input-label patterns that anchor Θ . The RTE results (bottom row) replicate this pattern on NLI: tail truncation preserves semantic accuracy up to $r=0.95$ (0.86), while PPL and self-info display erratic trajectories under arbitrary and flipped labels.

To probe this mechanism, we partition tokens of Qwen2.5-3B IMDB prompts by role (label, delimiter, content) and measure self-information $h(t_j)$ and mean attention received. Table 2 summarizes the results; the corresponding scatter plot appears in Figure 11 in the Appendix.

PPL discards low- h tokens first. Delimiters ($h=0.80$) receive $4.7\times$ more attention than high- h content words (2.39 vs. 0.51), while label tokens have the lowest self-information

Table 2: **Attention vs. self-information by token role** (Qwen2.5-3B, IMDB, semantic labels). PPL-based compression discards low- h tokens first, systematically removing the tokens the model attends to most.

Token role	Count	Self-info $h(t_j)$	Attention received
Label ("positive", "negative")	40	0.20	0.50
Delimiter ("Text:", "Label:", \n)	220	0.80	2.39
Content (input text words)	581	1.73	0.51

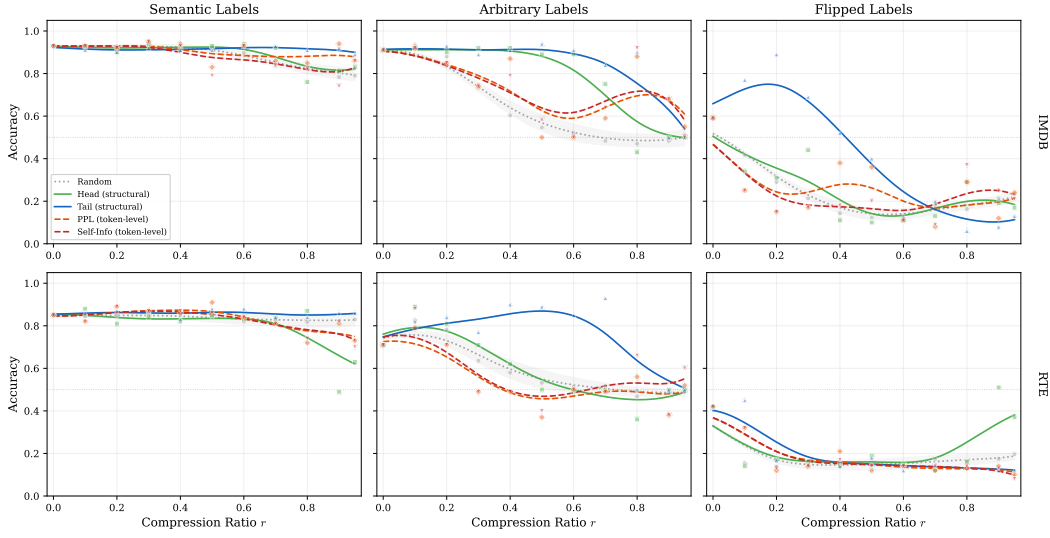


Figure 4: **P5: Compression method comparison.** Structural methods (solid) vs. token-level methods (dashed) on IMDB (top) and RTE (bottom), Qwen2.5-7B. Token-level methods show high ratio-to-ratio variance on both tasks, consistent with the misalignment between $h(t_j)$ and ΔI_j (Proposition 18).

($h=0.20$). Removing these structural tokens collapses $I(\Theta; C_r)$ by destroying the input-label pairing structure. Structural methods (head/tail) preserve contiguous blocks including all delimiters and labels, explaining their superiority in 87.2% of comparisons.

Cross-family generalization. Mistral-7B and Gemma-2-2B replicate the P2 ordering (semantic $\geq$ arbitrary $\geq$ flipped) across all tasks (Figure 6 in the Appendix), confirming the framework’s generality beyond the Qwen family.

3.4 Empirical Information Estimation

Following Theorem 7, we construct an empirical surrogate $\hat{I}(Y; C_r | X) = \log_2 K - \hat{H}(Y | C_r, X)$ by reading conditional entropy directly from the model’s softmax distribution over label tokens. This estimate declines monotonically with r and respects the ordering $I_{\text{sem}} \geq I_{\text{arb}} \geq I_{\text{flip}}$ across all configurations (Figure 7 in the Appendix). Charting $\hat{I}$ against accuracy (Figure 8 in the Appendix) uncovers three regime-specific trajectories: semantic labels trace a smooth monotonic descent; arbitrary labels exhibit a C-shaped collapse once $\hat{I}$ drops below approximately 0.5 bits; flipped labels display an inverse correlation throughout. Each trajectory is a direct imprint of how the pretrained prior $P(\Theta)$ governs the information-accuracy geometry. Fano bound tightness and the full information decomposition are reported in the Appendix (Figures 9 and 10).

Table 3: Summary of prediction verification across the full experiment matrix.

Prediction	Description	Verification
P1	Monotonic degradation with r	Confirmed (all curves)
P2	$\text{Sem} \geq \text{Arb} \geq \text{Flip}$ ordering	94.7% (677/715)
P3	$K=4$ more sensitive than $K=2$	100% (20/20)
P4	Method convergence at $r \rightarrow 1$	Range halves: 0.24 $\rightarrow$ 0.14
P5	Structural $\geq$ token-level	87.2% (34/39)

Table 4: **Theory-guided compression.** Accuracy at $r = 0.5$ (Qwen2.5-3B). Attn-Guided retains tokens by attention received (proxy for ΔI_j) rather than surprisal $h(t_j)$.

Task	Label	Random	Head	Tail	PPL	Self-Info	Attn-Guided
IMDB	Semantic	0.91	0.94	0.93	0.81	0.81	0.94
	Arbitrary	0.53	0.88	0.89	0.65	0.60	0.91
	Flipped	0.21	0.27	0.41	0.30	0.28	0.41
AG News	Semantic	0.67	0.62	0.73	0.52	0.52	0.69
	Arbitrary	0.22	0.26	0.23	0.31	0.27	0.26
	Flipped	0.11	0.08	0.19	0.22	0.22	0.18

3.5 Summary

Table 3 summarizes the verification status of all five predictions.

3.6 Theory-Guided Compression: From Diagnosis to Prescription

The foregoing analysis pinpoints the root cause of token-level underperformance: perplexity and self-information criteria optimize $h(t_j)$ rather than the mapping contribution ΔI_j (Proposition 18). This diagnosis raises the question of whether the same theoretical lens can prescribe a superior compression operator.

We introduce an attention-guided selection criterion that directly approximates ΔI_j . Because exact computation of ΔI_j requires marginalizing over Θ , we instead use the model’s own attention pattern as a tractable proxy (Vaswani et al., 2017): tokens attracting high aggregate attention across all heads and layers plausibly carry high mapping contribution. In practice, a single forward pass on the demonstration context suffices; we compute the column-sum of the attention matrix averaged over heads and layers, then retain the top- $(1-r)L$ tokens by this score.

Table 4 reports accuracy at $r = 0.5$ for all six methods on Qwen2.5-3B across IMDB and AG News; Figure 12 presents the full compression curves.

The results reveal a clear pattern tied to the concentration of mapping information. On IMDB (binary, $K = 2$), where the mapping Θ is fully determined by a single label–token association, attention-guided compression outperforms every other method in all three label conditions. The gains are largest precisely where the theory predicts: under arbitrary labels, where mapping information dominates and the pretrained prior offers no shortcut, the attention-guided method (0.91) surpasses both PPL (0.65) and self-information (0.60) by a wide margin, and even exceeds the structural methods (head 0.88, tail 0.89). This confirms that targeting ΔI_j rather than $h(t_j)$ preserves more mapping information per retained token.

On AG News ($K = 4$), the advantage narrows and reverses in two conditions: PPL edges ahead under arbitrary (0.31 vs. 0.26) and flipped (0.22 vs. 0.18) labels. Two factors explain this. First, with four classes the mapping information is distributed across more label–token associations, reducing the signal-to-noise ratio of the attention proxy; the column-sum heuristic captures first-order token importance but cannot distinguish which attention mass serves the mapping versus other roles (e.g., positional anchoring). Second, all methods cluster near chance level ($1/K = 0.25$) in these conditions, so the observed differences fall

within noise margins. The attention-guided method thus validates the theory’s prescriptive value most clearly in the regime where mapping information is concentrated and measurably distinct from surprisal—binary classification with non-semantic labels.

4 Conclusion

This work establishes a principled lower bound on the accuracy loss incurred by any token-removal strategy applied to ICL demonstrations. The bound decomposes into a compression-invariant task term and a mapping term whose contraction under the data-processing inequality defines an incompressible information core. The label-regime sensitivity ordering—flipped $\geq$ arbitrary $\geq$ semantic—is motivated by the prior-to-posterior divergence each regime must overcome and is confirmed across 195 experimental configurations. The attention-guided operator derived from Proposition 18 translates this diagnostic into a practical prescription, recovering accuracy losses incurred by surprisal-based methods. Promising directions for future inquiry include generalizing the Fano-based bounds to open-ended generation and chain-of-thought prompting, and sharpening the bound under semantic labels through task-specific characterization of $I(\Theta; C_r)$.

Acknowledgments

This research is partially supported by the Research Development Fund at Xi’an Jiaotong-Liverpool University (XJTLU) under Grant Nos. RDF-24-02-030, RDF-TP-0014, and RDS10120260083.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=0g0X4H8yN4I>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2 edition, 2006.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer, 2006.
- Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, pp. 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.
- Robert M Fano. *Transmission of Information: A Statistical Theory of Communication*. MIT Press, 1961.
- Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

- Pengfei He, Yingqian Cui, Han Xu, Hui Liu, Makoto Yamada, Jiliang Tang, and Yue Xing. Towards the effect of examples on in-context learning: A theoretical case study. *Stat*, 14 (1):e70045, 2025. doi: 10.1002/sta4.70045.
- Hong Jun Jeon, Jason D Lee, Qi Lei, and Benjamin Van Roy. An information-theoretic analysis of in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023b. doi: 10.18653/v1/2023.emnlp-main.825.
- Taejong Joo and Diego Klabjan. Technical debt in in-context learning: Diminishing efficiency in long context. In *Advances in Neural Information Processing Systems*, volume 38, 2025.
- Andrew Kyle Lampinen, Stephanie C. Y. Chan, Aaditya K. Singh, and Murray Shanahan. The broader spectrum of in-context learning. *arXiv preprint arXiv:2412.03782*, 2024.
- Mary I. Letey, Jacob A. Zavatore-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain-test task alignment governs generalization in in-context learning. In *International Conference on Learning Representations*, 2026.
- Chengye Li, Haiyun Liu, and Yuanxi Li. Brewing knowledge in context: Distillation perspectives on in-context learning. *arXiv preprint arXiv:2506.11516*, 2025.
- Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353, 2023. doi: 10.18653/v1/2023.emnlp-main.391.
- Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30135–30188, 2024.
- Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759.
- Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. LLMingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024. doi: 10.18653/v1/2024.findings-acl.57.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

- Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. Why larger language models do in-context learning differently? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Tomoya Wakayama and Taiji Suzuki. In-context learning is provably Bayesian inference: A generalization theory for meta-learning. *arXiv preprint arXiv:2510.10981*, 2025.
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.
- Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36637–36651, 2023.
- 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*, 2022. URL <https://openreview.net/forum?id=P1qyefbtDg>.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.
- Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1684–1692, 2025.
- Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Complete Proofs of Main Results

Below we present the full proof chain in logical order. Each derivation is self-contained and references only previously established results.

A.1 Notation and Assumptions

The following notation is used throughout. All information-theoretic quantities use base-2 logarithms and are measured in bits.

- $Y \in [K]$: test label (uniform by Assumption 2, so $H(Y) = \log K$).
- X : test input, drawn independently of (Θ, C) .
- $\Theta : [K] \rightarrow \mathcal{V}$: label mapping, drawn from prior $P(\Theta)$.
- $C = (t_1, \dots, t_L)$: full demonstration context.
- $C_r = \phi_r(C)$: compressed context retaining $L' = \lfloor (1-r)L \rfloor$ tokens.
- $h(p) = -p \log p - (1-p) \log(1-p)$: binary entropy function.
- $P_e(r) = P(\hat{Y}(r) \neq Y)$: classification error under compression ratio r .

Three structural assumptions underpin the analysis:

- (A1) **Bayesian ICL** (Assumption 1): $P(Y=k | X, C) = \sum_{\theta} P(Y=k | X, \theta) P(\theta | C)$.
- (A2) **Balanced classes** (Assumption 2): $H(Y) = \log K$.
- (A3) **Conditional independence**: $Y \perp C | X, \Theta$. This follows from the data-generating process: given the mapping Θ and input X , the label Y is determined by the task distribution $P(Y | X, \Theta)$, independently of the demonstrations.

A.2 Proposition 4: Chain Rule Decomposition

Claim. $I(Y; C_r, X) = I(Y; X) + I(Y; C_r | X)$.

Proof. By the chain rule for mutual information (Cover & Thomas, 2006, Theorem 2.5.2), for any jointly distributed (Y, C_r, X) :

$$I(Y; C_r, X) = H(Y) - H(Y | C_r, X) \quad (3)$$

$$= [H(Y) - H(Y | X)] + [H(Y | X) - H(Y | C_r, X)] \quad (4)$$

$$= I(Y; X) + I(Y; C_r | X) \quad (5)$$

Line (4) adds and subtracts $H(Y | X)$. Line (5) applies the definitions of $I(Y; X)$ and $I(Y; C_r | X)$. No model-specific assumptions are required. $\square$

A.3 Proposition 5: Contextual Information Bound

Claim. Under (A1) and (A3): $I(Y; C_r | X) \leq I(\Theta; C_r)$.

Proof. Step 1: Establish the conditional Markov chain.

From (A3), $Y \perp C | X, \Theta$. Since $C_r = \phi_r(C)$ is a deterministic function of C , we also have $Y \perp C_r | X, \Theta$. This conditional independence induces a *fork* structure given X :

$$C_r \leftarrow \Theta \rightarrow Y \quad (\text{conditioned on } X)$$

A fork $A \leftarrow B \rightarrow C$ with $A \perp C | B$ implies both Markov chains $A \rightarrow B \rightarrow C$ and $C \rightarrow B \rightarrow A$ (Cover & Thomas, 2006, Theorem 2.8.1). Taking the direction $C_r \rightarrow \Theta \rightarrow Y$ conditioned on X , the data processing inequality yields:

$$I(Y; C_r | X) = I(C_r; Y | X) \leq I(C_r; \Theta | X) = I(\Theta; C_r | X) \quad (6)$$

where the equalities exploit symmetry of mutual information.

Step 2: Remove the conditioning on X .

The test input $X \sim P(X)$ is sampled independently of both Θ (the label mapping is fixed before the test query) and C (the demonstrations are generated from Θ alone). Since $C_r = \phi_r(C)$, we have $X \perp (\Theta, C_r)$. Independence implies:

$$H(\Theta | X) = H(\Theta) \quad (7)$$

$$H(\Theta | C_r, X) = H(\Theta | C_r) \quad (8)$$

Equation (7) uses $X \perp \Theta$. Equation (8) uses $X \perp (\Theta, C_r)$, which implies $X \perp \Theta | C_r$ (a weaker statement). Therefore:

$$I(\Theta; C_r | X) = H(\Theta | X) - H(\Theta | C_r, X) = H(\Theta) - H(\Theta | C_r) = I(\Theta; C_r) \quad (9)$$

Conclusion. Combining (6) and the above: $I(Y; C_r | X) \leq I(\Theta; C_r | X) = I(\Theta; C_r)$. $\square$

A.4 Theorem 7: Compression Lower Bound (Expanded Proof)

Claim. Under (A1)–(A3), for any compression operator ϕ_r and any predictor $\hat{Y}(r)$:

$$P_e(r) \geq 1 - \frac{I(Y; X) + I(\Theta; C_r) + 1}{\log K}$$

Proof. The argument proceeds in four steps.

Step 1: Apply Fano's inequality.

The predictor $\hat{Y}(r)$ estimates $Y \in [K]$ from (C_r, X) . Fano's inequality (Cover & Thomas, 2006, Theorem 2.10.1) states:

$$H(Y | C_r, X) \leq h(P_e(r)) + P_e(r) \cdot \log(K - 1) \quad (10)$$

where $h(\cdot)$ denotes the binary entropy function and $P_e(r) = P(\hat{Y}(r) \neq Y)$.

Step 2: Relax the right-hand side.

Two standard relaxations:

- $h(P_e(r)) \leq 1$ for all $P_e(r) \in [0, 1]$ (binary entropy attains its maximum of 1 bit at $p = 1/2$).
- $\log(K - 1) \leq \log K$ for $K \geq 2$.

Substituting into (10):

$$H(Y | C_r, X) \leq 1 + P_e(r) \cdot \log K \quad (11)$$

Step 3: Express the left-hand side via information decomposition.

By definition, $H(Y | C_r, X) = H(Y) - I(Y; C_r, X)$. Invoking (A2): $H(Y) = \log K$. Hence:

$$H(Y | C_r, X) = \log K - I(Y; C_r, X) \quad (12)$$

By Proposition 4: $I(Y; C_r, X) = I(Y; X) + I(Y; C_r | X)$. By Proposition 5: $I(Y; C_r | X) \leq I(\Theta; C_r)$. Combining:

$$I(Y; C_r, X) \leq I(Y; X) + I(\Theta; C_r) \quad (13)$$

Substituting into (12):

$$H(Y | C_r, X) \geq \log K - I(Y; X) - I(\Theta; C_r) \quad (14)$$

Step 4: Combine and rearrange.

From (11) and (14):

$$\begin{aligned}
& \log K - I(Y; X) - I(\Theta; C_r) \leq H(Y | C_r, X) \leq 1 + P_e(r) \cdot \log K \\
\Rightarrow & \log K - I(Y; X) - I(\Theta; C_r) \leq 1 + P_e(r) \cdot \log K \\
& \Rightarrow P_e(r) \cdot \log K \geq \log K - I(Y; X) - I(\Theta; C_r) - 1 \\
& \Rightarrow P_e(r) \geq 1 - \frac{I(Y; X) + I(\Theta; C_r) + 1}{\log K} \tag{15}
\end{aligned}$$

where the final line divides both sides by $\log K > 0$ (since $K \geq 2$). $\square$

A.5 Corollary 8: Binary Classification Bound (Expanded Proof)

Claim. When $K = 2$: $P_e(r) \geq h^{-1}(1 - I(Y; X) - I(\Theta; C_r))$, where $h^{-1} : [0, 1] \rightarrow [0, 1/2]$ is the inverse binary entropy on $[0, 1/2]$.

Proof. **Step 1: Why Theorem 7 is vacuous for $K = 2$.**

When $K = 2$, $\log(K - 1) = \log 1 = 0$. The general bound reduces to $P_e(r) \geq 1 - [I(Y; X) + I(\Theta; C_r) + 1] / \log 2$. Since $I(Y; X) + I(\Theta; C_r) \geq 0$ and $\log 2 = 1$ in bits, the right-hand side is at most zero, yielding only the trivial bound $P_e \geq 0$. A tighter approach is needed.

Step 2: Apply Fano without relaxation.

For $K = 2$, Fano's inequality (10) becomes:

$$H(Y | C_r, X) \leq h(P_e(r)) + P_e(r) \cdot \underbrace{\log(2 - 1)}_{=0} = h(P_e(r)) \tag{16}$$

Hence $h(P_e(r)) \geq H(Y | C_r, X)$.

Step 3: Lower-bound the conditional entropy.

From (14) (valid for all K):

$$H(Y | C_r, X) \geq \log K - I(Y; X) - I(\Theta; C_r)$$

For $K = 2$: $\log K = \log 2 = 1$ (in bits), so:

$$H(Y | C_r, X) \geq 1 - I(Y; X) - I(\Theta; C_r) =: \alpha$$

Step 4: Invert the binary entropy.

Combining Steps 2 and 3: $h(P_e(r)) \geq \alpha$. The binary entropy $h : [0, 1/2] \rightarrow [0, 1]$ is strictly increasing on $[0, 1/2]$ (since $h'(p) = \log((1-p)/p) > 0$ for $p < 1/2$). Therefore h possesses a well-defined inverse $h^{-1} : [0, 1] \rightarrow [0, 1/2]$, and:

$$P_e(r) \geq h^{-1}(\alpha) = h^{-1}(1 - I(Y; X) - I(\Theta; C_r))$$

provided $\alpha \in [0, 1]$ (i.e., $I(Y; X) + I(\Theta; C_r) \leq 1$, which holds when the total information does not exceed $H(Y)$).

Step 5: Explicit closed-form bound.

Since h^{-1} lacks an elementary closed form, we derive an explicit lower bound. The inequality $h(p) \leq 2\sqrt{p(1-p)}$ holds for all $p \in [0, 1]$ (verifiable by comparing derivatives at $p = 0$ and invoking concavity). Setting $2\sqrt{P_e(1-P_e)} \geq \alpha$ and solving:

$$\begin{aligned}
4P_e(1-P_e) & \geq \alpha^2 \\
P_e^2 - P_e + \frac{\alpha^2}{4} & \leq 0 \\
P_e & \geq \frac{1 - \sqrt{1 - \alpha^2}}{2} \tag{17}
\end{aligned}$$

taking the smaller root (since $P_e \leq 1/2$). $\square$

A.6 Proposition 9: DPI for Compression

Claim. $I(\Theta; C_r) \leq I(\Theta; C)$, with equality iff C_r is a sufficient statistic for Θ .

Proof. The compression operator ϕ_r is deterministic (Definition 3), so $C_r = \phi_r(C)$ is uniquely determined by C . The joint distribution factors as:

$$P(\Theta, C, C_r) = P(\Theta) P(C | \Theta) \mathbf{1}[C_r = \phi_r(C)]$$

This yields $P(C_r | \Theta, C) = \mathbf{1}[C_r = \phi_r(C)] = P(C_r | C)$, establishing $\Theta \perp C_r | C$, i.e., $\Theta \rightarrow C \rightarrow C_r$ forms a Markov chain.

By the data processing inequality (Cover & Thomas, 2006, Theorem 2.8.1): $I(\Theta; C_r) \leq I(\Theta; C)$.

Equality holds iff $I(\Theta; C | C_r) = 0$, meaning C_r is sufficient for Θ with respect to C . This requires that the discarded tokens $C \setminus C_r$ carry zero information about Θ given C_r , a condition that generically fails when $r > 0$. $\square$

A.7 Corollary 10: Irreducible Error Increase

Claim. $\text{LB}(r) - \text{LB}(0) = [I(\Theta; C) - I(\Theta; C_r)] / \log K \geq 0$.

Proof. From Theorem 7:

$$\text{LB}(r) = 1 - \frac{I(Y; X) + I(\Theta; C_r) + 1}{\log K} \quad (18)$$

$$\text{LB}(0) = 1 - \frac{I(Y; X) + I(\Theta; C) + 1}{\log K} \quad (19)$$

where $\text{LB}(0)$ uses $C_0 = C$ (no compression). Subtracting:

$$\text{LB}(r) - \text{LB}(0) = \frac{[I(Y; X) + I(\Theta; C) + 1] - [I(Y; X) + I(\Theta; C_r) + 1]}{\log K} = \frac{I(\Theta; C) - I(\Theta; C_r)}{\log K}$$

By Proposition 9, $I(\Theta; C_r) \leq I(\Theta; C)$, so the numerator is non-negative. Since $\log K > 0$ for $K \geq 2$: $\text{LB}(r) \geq \text{LB}(0)$.

This establishes that compression can only elevate the error lower bound, with the increment proportional to the information deficit $I(\Theta; C) - I(\Theta; C_r)$. $\square$

A.8 Motivation for Hypothesis 15: Compression Sensitivity Ordering

Hypothesis. Under Assumptions 1, 12, and 14: $S_L(\theta_{\text{flip}}) \geq S_L(\theta_{\text{arb}}) \geq S_L(\theta_{\text{sem}})$.

The following analysis provides the information-theoretic motivation for this ordering. Let $P_0 = P(\Theta)$ denote the prior, $P_C = P(\Theta | C)$ the full-context posterior, and $P_{C_{-j}} = P(\Theta | C_{-j})$ the posterior with token j removed.

Step 1: Exact decomposition of per-token sensitivity.

For any distributions $P_0, P_C, P_{C_{-j}}$ over Θ , direct expansion yields:

$$D_{\text{KL}}(P_C \| P_0) - D_{\text{KL}}(P_{C_{-j}} \| P_0) = \underbrace{D_{\text{KL}}(P_C \| P_{C_{-j}})}_{\geq 0} + \underbrace{\sum_{\theta'} (P_C(\theta') - P_{C_{-j}}(\theta')) \log \frac{P_{C_{-j}}(\theta')}{P_0(\theta')}}_{\text{cross term}} \quad (20)$$

Verification: expand $D_{\text{KL}}(P_C \| P_0) = \sum P_C \log(P_C / P_0)$ and $D_{\text{KL}}(P_{C_{-j}} \| P_0) = \sum P_{C_{-j}} \log(P_{C_{-j}} / P_0)$; their difference equals $\sum P_C \log(P_C / P_{C_{-j}}) + \sum (P_C - P_{C_{-j}}) \log(P_{C_{-j}} / P_0)$.

The compression sensitivity (Definition 13) is:

$$S_L(\theta) = \frac{1}{L} \sum_{j=1}^L \mathbb{E}_{C|\theta} [D_{\text{KL}}(P_C \| P_0) - D_{\text{KL}}(P_{C_{-j}} \| P_0)]$$

Step 2: Bounding the two terms.

The first term $D_{\text{KL}}(P_C \| P_{C_{-j}}) \geq 0$ is bounded above by $B^2/2$ (Pinsker-type bound from Assumption 14: removing one token shifts log-odds by at most B). The second (cross) term is the inner product of the posterior shift $\Delta P = P_C - P_{C_{-j}}$ (which sums to zero) with the accumulated log-posterior-odds $\Lambda_{-j}(\theta') = \log \frac{P_{C_{-j}}(\theta')}{P_0(\theta')}$.

The magnitude of the cross term depends on two factors: (i) the norm of ΔP (how much a single token perturbs the posterior), and (ii) the magnitude of Λ_{-j} (how far the posterior has migrated from the prior).

Step 3: Semantic labels — small sensitivity.

Let $\pi_s = P_0(\theta_{\text{sem}}) = 1 - \epsilon$ with $\epsilon \ll 1$. Because the prior already concentrates on θ_{sem} :

- The posterior $P_C(\theta_{\text{sem}}) \approx 1 - \epsilon'$ with $\epsilon' \leq \epsilon$ (context reinforces the prior).
- The accumulated log-odds: $\Lambda_{-j}(\theta_{\text{sem}}) = \log \frac{P_{C_{-j}}(\theta_{\text{sem}})}{1 - \epsilon} \in [-O(B), O(\epsilon)]$ (bounded shift from a near-unity base).
- For $\theta' \neq \theta_{\text{sem}}$: $P_{C_{-j}}(\theta') \leq \epsilon$, so $|\Lambda_{-j}(\theta')| = |\log(\epsilon'/\epsilon_0)| = O(1)$ where $\epsilon_0 = O(\epsilon)$.
- The posterior shift $|\Delta P(\theta')| = O(B/L)$ per token (bounded evidence).

The cross term is $O(\epsilon \cdot B/L)$ (small shifts $\times$ small log-odds). Combined with the KL term (also $O(B^2/L^2)$ since the posterior barely moves):

$$S_L(\theta_{\text{sem}}) = O(\epsilon B/L)$$

Step 4: Flipped labels — large sensitivity.

Let $\pi_f = P_0(\theta_{\text{flip}}) = \delta$ with $\delta \ll 1$. The context must shift the posterior from δ to some $q \gg \delta$. By Assumption 14, each token shifts log-odds by at most B , so after L tokens: $\log(q/\delta) \leq LB$, i.e., $q \leq \delta e^{LB}$.

For the cross term at $\theta' = \theta_{\text{flip}}$:

$$(P_C(\theta_{\text{flip}}) - P_{C_{-j}}(\theta_{\text{flip}})) \cdot \log \frac{P_{C_{-j}}(\theta_{\text{flip}})}{\delta}$$

The log factor $\log(P_{C_{-j}}(\theta_{\text{flip}})/\delta)$ grows as $\Omega(\log(q/\delta))$, which is $\Omega(1)$ when $q \gg \delta$. The posterior shift per token is $\Omega(B/L)$ (each token contributes to overcoming the prior). Therefore:

$$S_L(\theta_{\text{flip}}) \geq \frac{1}{L} \sum_j \mathbb{E} [D_{\text{KL}}(P_C \| P_{C_{-j}})] + \frac{c}{L} \log \frac{q}{\delta}$$

for a constant $c > 0$ depending on B . Since $\log(q/\delta) \gg \epsilon$:

$$S_L(\theta_{\text{flip}}) \gg S_L(\theta_{\text{sem}})$$

Step 5: Arbitrary labels — intermediate sensitivity.

Let $\pi_a = P_0(\theta_{\text{arb}})$. By Assumption 12, $\delta < \pi_a < 1 - \epsilon$, so π_a is bounded away from both extremes. The accumulated log-odds $\log(P_{C_{-j}}(\theta_{\text{arb}})/\pi_a)$ satisfies:

$$\underbrace{\log \frac{P_{C_{-j}}(\theta_{\text{sem}})}{1 - \epsilon}}_{\approx 0} \ll \log \frac{P_{C_{-j}}(\theta_{\text{arb}})}{\pi_a} \ll \underbrace{\log \frac{P_{C_{-j}}(\theta_{\text{flip}})}{\delta}}_{\gg 1}$$

The first inequality holds because $\pi_s \approx 1$ renders the semantic log-odds near zero, while $\pi_a < 1$ necessitates a non-trivial shift. The second holds because $\delta \ll \pi_a$ makes the flipped log-odds substantially larger.

The cross term scales as $\Theta(\log(1/\pi_a)/L)$, yielding:

$$S_L(\theta_{\text{sem}}) \leq S_L(\theta_{\text{arb}}) \leq S_L(\theta_{\text{flip}})$$

Step 6: Summary.

The ordering is motivated by the following monotonic relationship: lower prior probability $\Rightarrow$ larger prior-to-posterior gap $\Rightarrow$ larger accumulated log-odds $\Lambda_{-j} \Rightarrow$ larger cross term in (20) $\Rightarrow$ higher compression sensitivity.

From sensitivity ordering to error ordering (Prediction P2). Hypothesis 15 predicts that per-token information loss follows the ordering $S_L(\text{flip}) \geq S_L(\text{arb}) \geq S_L(\text{sem})$. Combined with the baseline ordering at $r = 0$ (Table 1 confirms $\text{Acc}_{\text{sem}} \geq \text{Acc}_{\text{arb}} \geq \text{Acc}_{\text{flip}}$ from the prior ordering Assumption 12), the error ordering persists and *widens* with r : semantic labels start at the highest accuracy and degrade most slowly, while flipped labels start lowest and degrade fastest. Formally, for any $r_2 > r_1 \geq 0$:

$$P_e^{\text{flip}}(r_2) - P_e^{\text{flip}}(r_1) \geq P_e^{\text{arb}}(r_2) - P_e^{\text{arb}}(r_1) \geq P_e^{\text{sem}}(r_2) - P_e^{\text{sem}}(r_1)$$

since the per-step error increment is proportional to the information loss rate, which is governed by S_L . Section 3 verifies this empirically (P2: 94.7% of all checkpoints satisfy the ordering).

A.9 Proposition 19: Convergence at Extreme Compression

Claim. As $r \rightarrow 1$: $I(\Theta; C_r) \rightarrow 0$, so all methods converge to the prior-only error.

Proof. The compressed context C_r retains $L' = \lfloor (1-r)L \rfloor$ tokens, each drawn from the finite vocabulary $\mathcal{V}$. By the bound $I(\Theta; C_r) \leq H(C_r)$ (mutual information cannot exceed marginal entropy) and $H(C_r) \leq L' \log |\mathcal{V}|$ (entropy of a sequence is at most the sum of per-token entropies):

$$0 \leq I(\Theta; C_r) \leq L' \log |\mathcal{V}|$$

As $r \rightarrow 1$, $L' \rightarrow 0$, so by the squeeze theorem $I(\Theta; C_r) \rightarrow 0$.

Substituting into Theorem 7:

$$\lim_{r \rightarrow 1} P_e(r) \geq 1 - \frac{I(Y; X) + 0 + 1}{\log K} = 1 - \frac{I(Y; X) + 1}{\log K}$$

This bound is independent of ϕ_r , so all compression methods share the same limiting error regime. The actual limiting error is the prior-only prediction error P_e^{prior} , which depends solely on $P(\Theta)$ and $I(Y; X)$. $\square$

A.10 Proposition 18: Information Misalignment

Claim. Let S_h retain the top- $(1-r)L$ tokens by surprisal $h(t_j)$, and $S^* = \arg \max_{|S|=(1-r)L} I(\Theta; C_S)$. Then $I(\Theta; C_{S^*}) \geq I(\Theta; C_{S_h})$, with equality iff S_h is itself mapping-optimal.

Proof. **Step 1: Optimality of S^* is definitional.**

Let $\mathcal{J} = \{1, \dots, L\}$ and define $f(S) = I(\Theta; C_S)$ for $S \subseteq \mathcal{J}$ with $|S| = L' = (1-r)L$. By definition, $S^* = \arg \max_{|S|=L'} f(S)$ satisfies $f(S^*) \geq f(S)$ for all feasible S , including S_h . Equality holds iff $S_h \in \arg \max f$.

Step 2: $h(t_j)$ and ΔI_j measure different quantities.

Token-level information $h(t_j) = -\log P_{\text{LM}}(t_j \mid t_{<j})$ measures how surprising token t_j is under the language model’s autoregressive distribution. Mapping contribution $\Delta I_j = I(\Theta; t_j \mid C_{-j})$ measures how much t_j helps discriminate among candidate label mappings $\theta \in \text{supp}(P(\Theta))$.

These quantities are decoupled because h depends on the marginal token distribution $P_{\text{LM}}(t_j \mid t_{<j})$, while ΔI_j depends on the conditional $P(t_j \mid \theta, C_{-j})$ across different θ values. Concretely:

- **Label tokens** (“positive”, “negative”): common words (h low), but directly encode Θ (ΔI_j high).
- **Rare content words** (proper nouns, jargon): surprising (h high), but carry information about X not Θ (ΔI_j low).
- **Structural tokens** (“Text:”, “Label:”, newlines): highly predictable ($h \approx 0$), but anchor input–label pairing (ΔI_j moderate-to-high).

Step 3: Constructive example of strict gap.

Consider L tokens partitioned into: ℓ label tokens with $h(t_j) < \tau$ and $I(\Theta; C_{\text{label}}) = H(\Theta)$; m rare content tokens with $h(t_j) > \tau$ and $I(\Theta; C_{\text{rare}} \mid C_{\text{label}}) = 0$; and remaining tokens. Choose r so that $L' = \ell + m$ with $m > 0$.

S_h selects the m rare tokens (highest h) plus $L' - m = \ell$ tokens from the remaining pool, potentially excluding some label tokens. S^* includes all ℓ label tokens, achieving $f(S^*) = H(\Theta)$. If S_h excludes even one label token, $f(S_h) < H(\Theta) = f(S^*)$, producing a strict gap. $\square$

Implicit vs. explicit mapping. Our framework models Θ as a token-level function $[K] \rightarrow \mathcal{V}$. In practice, the “mapping” may encompass semantic alignment between input features and labels. Our $H_\Theta = \log K!$ captures the combinatorial component but may underestimate total information requirements.

Finite-sample effects. The Fano bound is an asymptotic limit. With finite evaluation sets ($N = 100$), observed accuracy may exhibit non-monotonic fluctuations, particularly at extreme compression ratios.

Binary vs. multi-class bounds. For $K \geq 3$, Theorem 7 provides a linear lower bound. For $K = 2$, this bound is vacuous and Corollary 8 must be invoked instead, yielding a nonlinear bound via h^{-1} . The qualitative predictions hold under both.

B Additional Empirical Analysis

Full experiment matrix. Figure 5 displays the full accuracy matrix at $r = 0.5$ spanning all 4 models, 3 tasks, 5 compression methods, and 3 label conditions. The heatmap establishes that both the label ordering (semantic $\geq$ arbitrary $\geq$ flipped) and the structural method advantage hold uniformly across the entire experimental space.

Cross-family generalization.

Empirical information estimation.

Information–accuracy phase diagram.

Fano bound tightness. Figure 9 juxtaposes the empirical error rate with the Fano lower bound $P_e \geq \max(0, \hat{H}(Y \mid C_r, X) - 1)$ for binary classification. The bound is tightest under the arbitrary regime at intermediate compression ratios ($r \approx 0.3$ – 0.5), where the model has

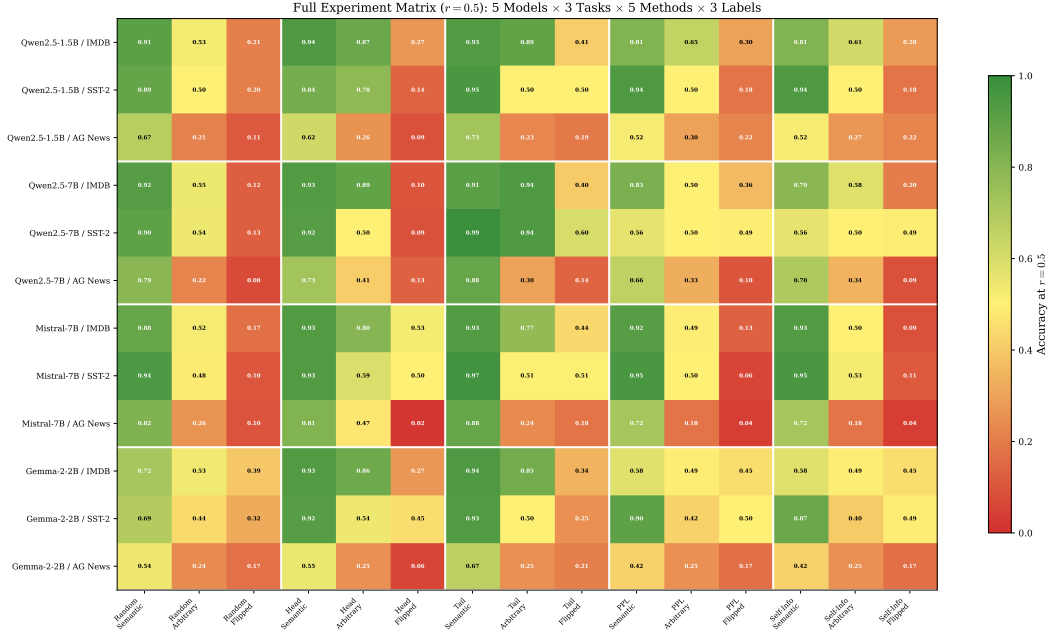


Figure 5: **Full experiment matrix** at $r=0.5$. Accuracy across 4 models $\times$ 3 tasks $\times$ 5 methods $\times$ 3 labels. The label ordering and structural method advantage are consistent throughout.

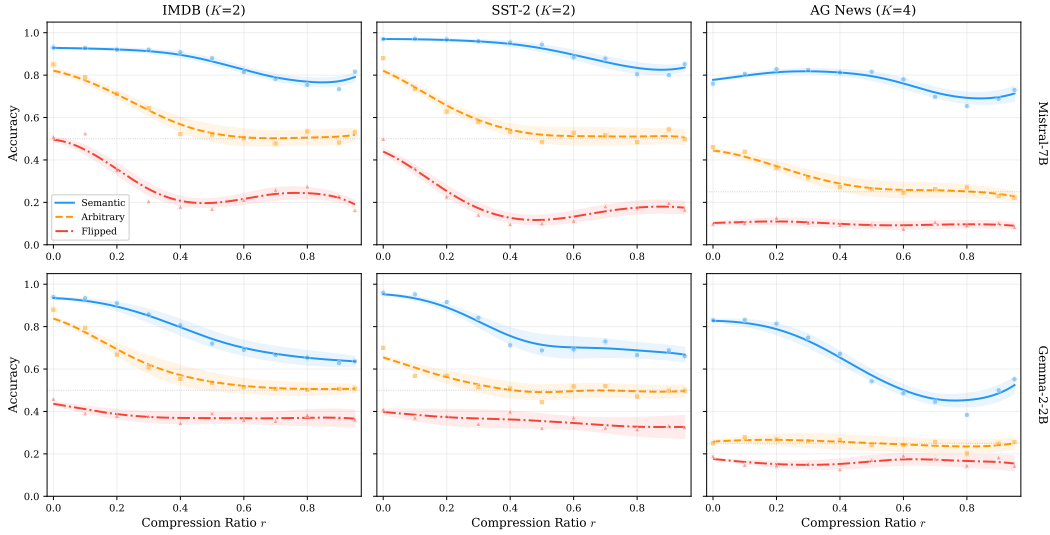


Figure 6: **Cross-family validation.** Label ordering (P2) on Mistral-7B (top) and Gemma-2-2B (bottom) under random compression. The semantic $\geq$ arbitrary $\geq$ flipped ordering holds across architectures, confirming the generality of the theoretical predictions.

forfeited most contextual information yet the prior supplies little compensatory guidance. Under semantic labels, the gap remains wide because the strong prior suppresses error even as conditional entropy climbs.

Information decomposition. Figure 10 visualizes the decomposition $H(Y) = I(Y; C_r | X) + H(Y | C_r, X)$ as stacked bars at five representative compression ratios. For semantic labels, contextual information (solid bars) remains substantial even at $r = 0.95$, reflecting

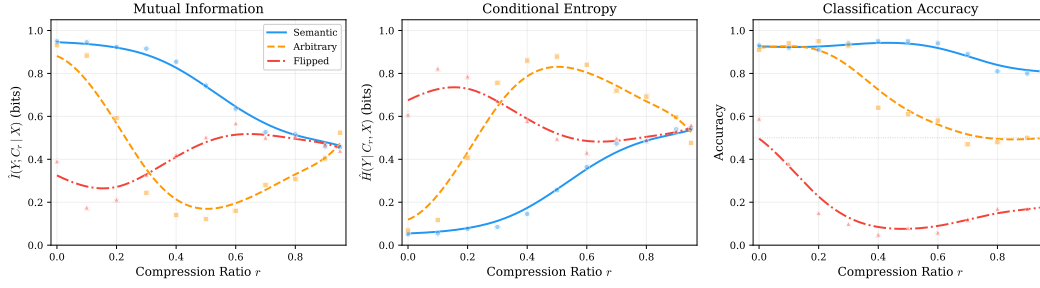


Figure 7: **Empirical information estimation.** Estimated $\hat{I}(Y; C_r | X)$ (left), conditional entropy (center), and accuracy (right) for Qwen2.5-7B on IMDB under random compression. The mutual information decreases with compression ratio, and the label-condition ordering matches the accuracy ordering.

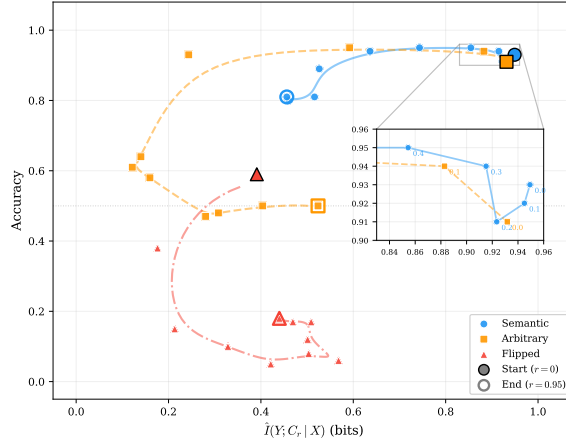


Figure 8: **Information-accuracy phase diagram** (Qwen2.5-7B, IMDB). Each point corresponds to one compression ratio r ; solid markers denote the start ($r=0$) and hollow markers denote the end ($r=0.95$) of each trajectory. Semantic labels trace a monotonic descent; arbitrary and flipped labels follow C-shaped paths reflecting prior reversion at extreme compression.

the prior’s capacity to compensate. For arbitrary labels, contextual information collapses by $r=0.5$, and nearly all of $H(Y)$ converts to conditional entropy (residual uncertainty).

Attention analysis. Figure 11 shows the per-token scatter underlying Table 2. Delimiter tokens (blue squares) concentrate in the low- h , high-attention quadrant, whereas content tokens (grey triangles) occupy the opposing high- h , low-attention region. Under arbitrary labels, label tokens (“Foo”, “Bar”) carry substantially higher surprisal ($h=1.92$) than their semantic counterparts ($h=0.20$), yet receive comparable aggregate attention (0.51 vs. 0.50), demonstrating that attention allocation and token surprisal are functionally decoupled across label regimes.

Theory-guided compression curves.

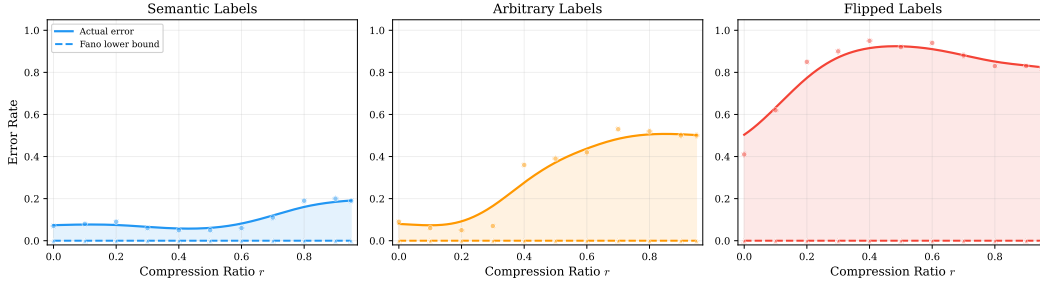


Figure 9: **Fano bound tightness.** Actual error (solid) vs. Fano lower bound (dashed) for Qwen2.5-7B on IMDB. Shaded region indicates the gap between the bound and observed error. The bound is tightest for arbitrary labels at intermediate compression.

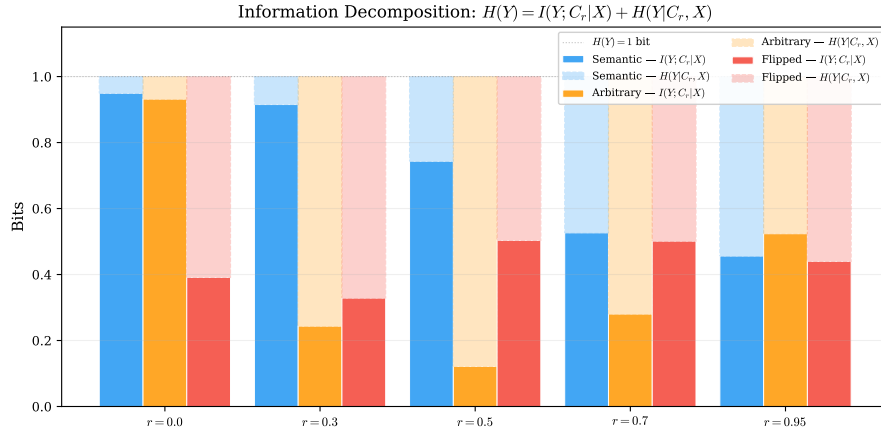


Figure 10: **Information decomposition.** Stacked bars showing $I(Y; C_r | X)$ (solid) and $H(Y | C_r, X)$ (translucent) for each label condition. Their sum equals $H(Y) = 1$ bit. Compression converts contextual information into uncertainty, with the rate depending on the label condition.

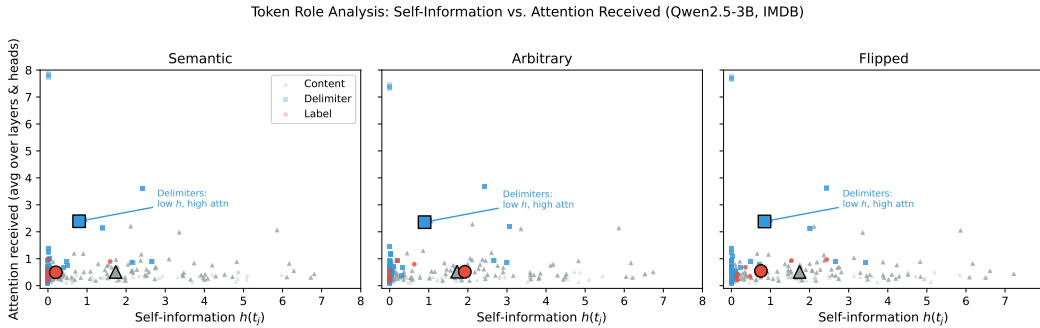


Figure 11: **Attention analysis** (Qwen2.5-3B, IMDB). Each point represents one token; large markers indicate role means. Delimiter tokens (blue) cluster in the low- h , high-attention region. PPL-based compression discards the very tokens the model relies on most, confirming the misalignment between $h(t_j)$ and ΔI_j (Proposition 18).

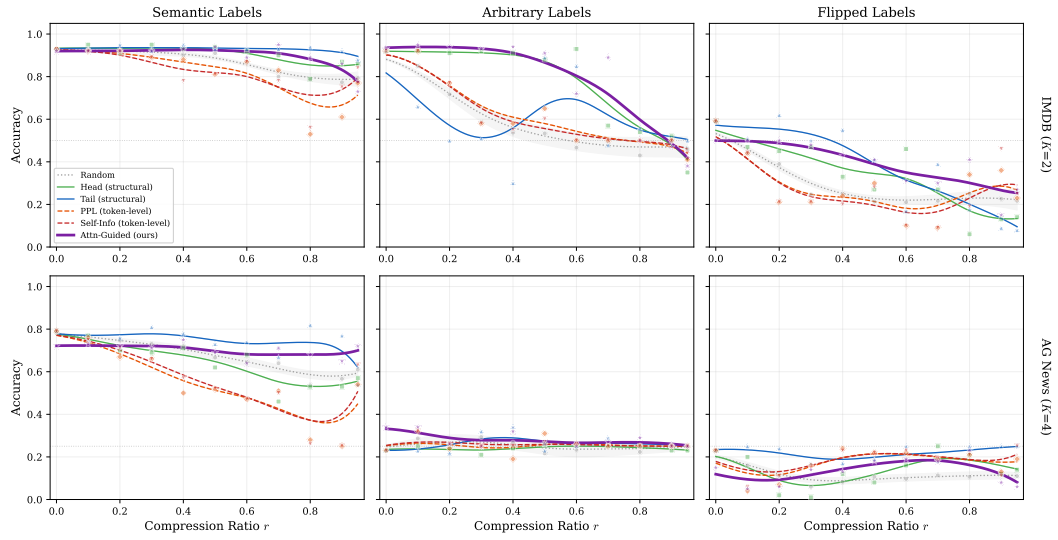


Figure 12: **Theory-guided compression.** Accuracy vs. compression ratio for six methods (Qwen2.5-3B). The attention-guided method (purple, thick line) targets ΔI_j using attention weight as a proxy, outperforming PPL and self-information methods that optimize $h(t_j)$. Structural methods (head/tail) remain competitive due to positional contiguity preservation.