THE ALIGNMENT BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models improve with scale, yet feedback-based alignment still exhibits systematic deviations from intended behavior. Motivated by bounded rationality in economics and cognitive science, we view judgment as resource-limited and feedback as a constrained channel; on this basis we model the loop as a two-stage cascade $U \to H \to Y$ given S, with cognitive capacity $C_{\mathrm{cog}|S}$ and average total capacity $\bar{C}_{\mathrm{tot}|S}$. Our main result is a capacity coupled Alignment Performance Interval. It pairs a data size independent Fano lower bound proved on a separable codebook mixture with a PAC–Bayes upper bound whose KL term is controlled by the same channel via $m\,\bar{C}_{\mathrm{tot}|S}$. The PAC–Bayes bound becomes an upper bound on the same true risk when the canonical observable loss is used and the dataset is drawn from the same mixture. Under these matched conditions both limits are governed by a single capacity. Consequences include that, with value complexity and capacity fixed, adding labels alone cannot cross the bound; attaining lower risk on more complex targets requires capacity that grows with $\log M$; and once useful signal saturates capacity, further optimization tends to fit channel regularities, consistent with reports of sycophancy and reward hacking. The analysis views alignment as interface engineering: measure and allocate limited capacity, manage task complexity, and decide where information is spent.

1 INTRODUCTION

Scaling laws continue to improve LLM capabilities (Kaplan et al., 2020; Wei et al., 2022; Hestness et al., 2017), but feedback-based alignment shows a tension: instruction-following improves on average, while systematic deviations from intended behavior persist. In practice, feedback-based alignment pipelines have substantially improved instruction following (Ouyang et al., 2022; Bai et al., 2022b; Ziegler et al., 2020; Zheng et al., 2023; Rafailov et al., 2024b; Lee et al., 2024). Nevertheless, models continue to exhibit sycophancy, reward hacking, and inverse scaling on truthfulness (Perez et al., 2023; Sharma et al., 2024; Pan et al., 2022; Denison et al., 2024; Lin et al., 2022; Amodei et al., 2016). A natural question is whether these patterns partly reflect a structural limit of the human–AI feedback loop.

Motivating evidence spans economics and cognitive science: bounded rationality views decisions as resource-limited and often satisficing (Simon, 1955); computational and information-theoretic models then show that people compress task representations and trade performance for cognitive cost (Lewis et al., 2014; Ho et al., 2022; 2020; Zénon et al., 2019; Zaslavsky et al., 2021). Rate–distortion and information-bottleneck perspectives connect these constraints to perception, control, and RL (Sims, 2016; Ortega and Braun, 2011; Lai and Gershman, 2021; Arumugam et al., 2023; Arumugam and Roy, 2022). These results motivate treating feedback as information passing through a bounded system rather than as a noiseless oracle.

With this empirical and theoretical background, we model the feedback loop as a two-stage cascade $U \to H \to Y$ given context S: latent human values U are first compressed into internal judgments H, then articulated as observable signals Y. We define the total conditional capacity $\bar{C}_{\mathrm{tot}|S} = \mathbb{E}[\min\{C_{\mathrm{cog}|S},\, C_{\mathrm{art}|S}\}]$ and highlight the cognitive capacity $C_{\mathrm{cog}|S}$ as the typical bottleneck through which value information must pass, and we connect rate–distortion and information-bottleneck ideas, focusing on what fidelity is possible under the human–AI channel's cognitive capacity (Tishby et al., 2000; Tishby and Zaslavsky, 2015; Alemi et al., 2017; Kolchinsky et al., 2019; Kawaguchi et al., 2023; Saxe et al., 2019; Shwartz-Ziv and Tishby, 2017; Goldfeld et al., 2019).

We establish a capacity link: the same capacity that limits value information entering the data also governs the statistical complexity needed for generalization. On the lower-bound side (Sec. 4), using separable codebooks and Fano, we obtain a data-size–independent information lower bound on true risk,

$$R_{\mathrm{mix}}(\pi) \;\geq\; (\varepsilon+\Delta)\left(1-\frac{\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S}+\log 2}{\log M}\right)_+.$$

On the upper-bound side (Sec. 5), via PAC–Bayes for bounded observable losses and the link between KL complexity and dataset–parameter mutual information (Xu and Raginsky, 2017; Russo and Zou, 2019; Rodríguez-Gálvez et al., 2024; Lotfi et al., 2024; Dziugaite and Roy, 2017), we show

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{KL}(P\|Q)\big] \;\leq\; m\,\bar{C}_{\mathrm{tot}|S} \;+\; m\,I(U;S) \;+\; \rho \;+\; \mathrm{KL}\big(p(\theta)\|Q\big),$$

which renders the upper bound explicit for the same channel that defines the converse. Taken together, under the canonical observable loss and under the same codebook mixture used in the converse, we obtain a capacity coupled interval:

$$(\varepsilon+\Delta)\Big(1-\frac{\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S}+\log 2}{\log M}\Big)_+ \;\leq\; R_{\mathrm{mix}}(\pi) \;\leq\; \mathbb{E}_{\theta\sim P}\big[\widehat{R}^{\mathrm{obs}}_m(\theta)\big] + \sqrt{\frac{\mathrm{KL}(P\|Q)+\log(1/\delta)}{2m}},$$

The KL term is further controlled in expectation by the same channel capacity. To our knowledge, prior analyses have not coupled a Fano-type lower bound and a PAC–Bayes upper bound through a single capacity term of the human–AI channel.

These bounds imply that increasing the dataset size m alone cannot overcome the lower bound when separability and capacity are fixed; achieving a target risk requires capacity that scales with value complexity, which constrains pluralistic or multi-objective alignment; and once a useful signal saturates capacity, powerful optimizers can continue to reduce empirical loss by fitting residual channel regularities, consistent with reports of sycophancy and related behaviors (Perez et al., 2023; Sharma et al., 2024).

2 Related Work

Feedback Alignment and Systematic Deviations Contemporary alignment trains policies to preference signals using feedback-driven pipelines that collect preference data and adjust behavior under varied supervision protocols (Ouyang et al., 2022; Christiano et al., 2017; Ziegler et al., 2020; Bai et al., 2022a; Zheng et al., 2023; Rafailov et al., 2024b; Lin et al., 2024b; Bai et al., 2022b; Ethayarajh et al., 2024; Guo et al., 2024; Lee et al., 2024; Mu et al., 2024). Despite gains, models display systematic deviations such as sycophancy, reward hacking, and inverse scaling on truthfulness, and raise sequential concerns including user tampering (Perez et al., 2023; Sharma et al., 2024; Pan et al., 2022; Denison et al., 2024; Lin et al., 2022; Evans and Kasirzadeh, 2023). Formal accounts situate these deviations in system-level incentives and representation/oversight mismatches across aggregation and interaction protocols (Ge et al., 2024; Irving et al., 2018; Everitt et al., 2021; Ngo et al., 2024; Rane et al., 2024). Related analyses emphasize dynamic optimization effects in which proxy-reward gains can diverge from target behavior under increasing optimization pressure (Gaikwad, 2025), with empirical scaling in model size and KL budgets and $\sqrt{\mathrm{KL}}$-type ceilings (Rafailov et al., 2024a; Mroueh and Nitsure, 2025). We instead take a static source–channel view: the same capacity term—typically governed by $C_{\mathrm{cog}|S}$—controls the Fano lower bound and the PAC–Bayes complexity, yielding an optimizer-agnostic interval.

Bounded Rationality and Cognitive Constraints Originating in economics and organizational theory as bounded rationality and satisficing (Simon, 1955), subsequent work in cognitive science and information-theoretic decision-making models judgment as resource-limited computation with explicit costs for processing and representation (Lewis et al., 2014; Ho and Griffiths, 2022; Gottwald and Braun, 2019; Ortega and Braun, 2011; Zénon et al., 2019; Ho et al., 2020). Empirically, people construct simplified task representations and plan under constrained internal state (Ho et al., 2022). Rate–distortion accounts capture these bottlenecks in perception and communication (Sims, 2016; Zaslavsky et al., 2021) and connect to RL and Bayesian decision-making to yield capacity-limited agents (Arumugam et al., 2023; Arumugam and Roy, 2022; Arumugam and Van Roy, 2021; Arumugam and Roy, 2021); policy compression frames action selection as an information bottleneck (Lai and Gershman, 2021). These lines motivate modeling the feedback pipeline as $U \to H \to Y$

with a cognitive-capacity term $C_{\mathrm{cog}|S}$ that often forms the binding bottleneck and upper-bounds $I(U;Y \mid S)$ in our analysis.

Information Theory in Machine Learning The Information Bottleneck program studies compression of irrelevant bits while preserving task-relevant information (Tishby et al., 2000; Tishby and Zaslavsky, 2015; Alemi et al., 2017; Peng et al., 2020; Kolchinsky et al., 2019), with continued debate about what "compression" measures in deterministic networks (Saxe et al., 2019; Goldfeld et al., 2019; Shwartz-Ziv and LeCun, 2023; Shwartz-Ziv and Tishby, 2017). Information-theoretic bounds relate hierarchical processing to generalization (Kawaguchi et al., 2023; He et al., 2025; Bartlett et al., 2017) and analyze self-supervised objectives (Shwartz-Ziv et al., 2024). Our use of packings and Fano to build a data-size independent wall follows classical converse techniques and rate–distortion thinking (Shannon, 1959), and we interpret residual information learned beyond the true value as channel overfitting (Ngampruetikorn and Schwab, 2022). The compression view is also relevant because recent work shows that LLMs are strong general-purpose compressors (Delétang et al., 2024). These threads link information budgets and performance. Departing from IB's focus on compressing internal representations, we instead parameterize alignment by the context-conditioned capacity of the human–AI channel ($\bar{C}_{\mathrm{tot}|S}$) and use it to couple a codebook–Fano wall with a PAC–Bayes ceiling.

PAC–Bayes and Mutual Information PAC–Bayes provides non-asymptotic generalization guarantees that can remain informative at scale (Lotfi et al., 2024; Rodríguez-Gálvez et al., 2024; Wu et al., 2025; Leblanc et al., 2025; Picard-Weibel et al., 2025; Dziugaite and Roy, 2017; Neyshabur et al., 2018; Langford and Caruana, 2001; Neyshabur et al., 2017). A key development ties the KL term to mutual information between data and parameters (Xu and Raginsky, 2017; Russo and Zou, 2019), and the PAC–Bayes Information Bottleneck makes this connection algorithmic by directly regularizing $I(\mathcal{D};\theta)$ (Wang et al., 2022). We ground the abstract KL complexity in a physical constraint of the learning environment: the finite human-feedback capacity $\bar{C}_{\mathrm{tot}|S}$, often dominated by $C_{\mathrm{cog}|S}$ in our $U \to H \to Y$ model. This yields a capacity-coupled interval in which the same capacity term both limits $I(U;Y \mid S)$ in the Fano floor and controls attainable KL complexity in the PAC–Bayes ceiling. Related mitigations, such as information-bottleneck style reward modeling and behavior-supported methods (Miao et al., 2024; Dai et al., 2025), and upper-bound–style results (Mroueh and Nitsure, 2025) are compatible with this view by reallocating or constraining where the limited information budget is spent. By externalizing the KL complexity into the environmental budget $m\,\bar{C}_{\mathrm{tot}|S} + m\,I(U;S) + \rho$ induced by the $U \to H \to Y$ channel, our bound aligns the PAC–Bayes term with the same capacity that limits $I(U;Y \mid S)$, a linkage not provided by PAC–Bayes-IB.

3 Problem Setup

To rigorously analyze the Alignment Bottleneck, we model alignment as resource-constrained inference and communication. Following bounded and computational rationality in cognitive science (Lewis et al., 2014; Ortega and Braun, 2011), we treat the human feedback provider as a two-stage communication channel, and use channel capacity to quantify the bottleneck. This connects our formulation to the Information Bottleneck and rate–distortion viewpoints (Tishby et al., 2000; Shannon, 1959) and to cognitive accounts that frame judgment/externalization as utility–information trade-offs (Sims, 2016; Zaslavsky et al., 2021). We next formalize the task, the two-stage channel, and the corresponding capacities.

3.1 Task, Loss, and Feedback Channel

Definition 1 (Task and Observable Loss). *Let S denote publicly observable context, U the latent task target ("what humans truly want"), and Y the human feedback emitted through a finite-capacity channel. A learner outputs an action $\hat{a} = \pi(Y,S) \in \mathcal{A}$ using a decoder π. The task loss is a* bounded *measurable function $\ell : \mathcal{U} \times \mathcal{A} \to [0,1]$ (such as 0–1 loss, pairwise ranking loss mapped to $[0,1]$, or a truncated and normalized MSE; see Appx. K). The (population) risk is*

$$R(\pi) \triangleq \mathbb{E}\big[\ell(U, \pi(Y,S))\big]. \tag{1}$$

Definition 2 (Human Channel Families). *We model the human-in-the-loop communication by a cascade $U \to H \to Y$ given S. The* cognitive stage *uses a conditional kernel $p(h \mid u,s) \in \mathcal{F}_{\mathrm{cog}}$, and the* articulation stage *uses $p(y \mid h,s) \in \mathcal{F}_{\mathrm{art}}$. The learner observes only (Y,S), not H.*

This two-stage cascade $U \to H \to Y$ treats the human as a finite-capacity communication channel. Evidence from cognitive science shows that human judgment is resource-bounded and therefore compressive rather than perfect retrieval (Lewis et al., 2014; Zénon et al., 2019; Ortega and Braun, 2011; Gottwald and Braun, 2019); people construct task-specific construals that trade representational complexity for utility (Ho et al., 2022), which provides a concrete mechanism for the $U \to H$ bottleneck and aligns with rate–distortion views of perception and communication (Sims, 2016; Zaslavsky et al., 2021). Beyond description, bounded-rationality formalisms operationalize these limits: policy selection itself can be cast as an information bottleneck (Lai and Gershman, 2021), and rate–distortion–constrained learning appears in bandits and then full RL, culminating in a common Bayesian/RL view of capacity-limited behavior (Arumugam and Van Roy, 2021; Arumugam and Roy, 2021; 2022; Arumugam et al., 2023). We adopt this source–channel lens and treat the finite cognitive capacity $C_{\mathrm{cog}|S}$ as an often binding bottleneck through which value information must pass.

Assumption 1 (Per-stage Feasibility)**.** *All admissible systems considered in this paper satisfy* $p(h \mid u, s) \in \mathcal{F}_{\mathrm{cog}}$ *and* $p(y \mid h, s) \in \mathcal{F}_{\mathrm{art}}$ *for almost every* (u, s)*, with the two stages independent across i.i.d. samples.*

3.2 Context-Conditional Capacities

Definition 3 (Cognitive Capacity)**.** *For each* s*, define*

$$C_{\mathrm{cog}|S}(s) \triangleq \sup_{p(h|u,s)\in\mathcal{F}_{\mathrm{cog}}} I(U; H \mid S = s)\,. \tag{2}$$

Definition 4 (Articulation Capacity)**.** *For each* s*, define*

$$C_{\mathrm{art}|S}(s) \triangleq \sup_{p(y|h,s)\in\mathcal{F}_{\mathrm{art}}} I(H; Y \mid S = s)\,. \tag{3}$$

Definition 5 (Total Capacity and Its Average)**.** *For each* s*, define the per-context total capacity*

$$C_{\mathrm{tot}|S}(s) \triangleq \min\Big\{C_{\mathrm{cog}|S}(s),\ C_{\mathrm{art}|S}(s)\Big\}, \tag{4}$$

and its average

$$\bar{C}_{\mathrm{tot}|S} \triangleq \mathbb{E}_S\big[C_{\mathrm{tot}|S}(S)\big]. \tag{5}$$

Proposition 1 (Cascade Upper Bound via Data Processing)**.** *Under Assumption 1, any admissible cascade* $U \to H \to Y$ *forms a Markov chain. By the data processing inequality (Shannon, 1948), it satisfies for every* s*,*

$$I(U; Y \mid S = s) \ \le\ \min\{I(U; H \mid S = s),\ I(H; Y \mid S = s)\} \ \le\ C_{\mathrm{tot}|S}(s), \tag{6}$$

and hence, averaging over S*,*

$$I(U; Y \mid S) \ \le\ \bar{C}_{\mathrm{tot}|S}. \tag{7}$$

We discuss technical details of these capacity definitions, including source-dependency and an extension to coarsened context variables, in Appendix C.

4 Information-Theoretic Lower Bounds

Having established our problem model, we derive the first component of the Alignment Performance Interval: an information-theoretic lower bound on the true risk. The bound exposes how task difficulty scales with value complexity and channel capacity. Following the classic minimax methodology from statistical decision theory and information theory (Shannon, 1948), we construct a family of hard but distinguishable tasks and apply Fano's inequality to show that any algorithm incurs nontrivial error in telling them apart. In our setting, this family is a "Δ-separable codebook" of value–action pairs.

4.1 Separable Codebooks and the Loss–Index Link

Definition 6 (Δ-Separable Codebook)**.** *A collection* $\{(u^{(i)}, a^{(i)})\}_{i=1}^{M} \subset \mathcal{U} \times \mathcal{A}$ *is called a* Δ-separable codebook *for loss* $\ell \in [0, 1]$ *if*

$$\ell\big(u^{(i)}, a^{(i)}\big) \le \varepsilon \quad \textit{for all } i, \tag{8}$$

$$\ell\big(u^{(j)}, a^{(i)}\big) \ge \varepsilon + \Delta \quad \textit{for all } j \ne i, \tag{9}$$

for some $\varepsilon \in [0, 1 - \Delta]$*. We write* $\mathcal{C}(M, \Delta, \varepsilon)$ *for the set of such codebooks.*

Standard constructions for such codebooks exist for common losses; we provide examples in Appendix E.

Assumption 2 (Loss–Index Link via a Measurable Partition)**.** *Given a Δ-separable codebook $\{(u^{(i)}, a^{(i)})\}_{i=1}^{M}$, there exists a measurable map $\phi : \mathcal{A} \times \mathcal{S} \to [M]$ (an "index decoder") such that for all i, all $s \in \mathcal{S}$, and all $a \in \mathcal{A}$,*

$$\phi(a, s) \neq i \implies \ell\big(u^{(i)}, a\big) \geq \varepsilon + \Delta. \tag{10}$$

This assumption holds for standard decoders under common losses; we provide examples and a high-probability variant in Appendix G.

4.2 Fano–Packing Converse Lower Bound

Lemma 1 (Risk ⇒ Index Error)**.** *Under Assumption 2, for any decoder π and ϕ as in equation 10, define $\hat{J} \triangleq \phi(\pi(Y, S), S)$. If J is uniform over $[M]$ and $U = U^{(J)}$ is the codebook target (measurable in (J, S)), then*

$$\mathbb{E}\big[\ell(U, \pi(Y, S))\big] \geq (\varepsilon + \Delta)\, \mathbb{P}\{\hat{J} \neq J\}. \tag{11}$$

Proof. By Assumption 2, for every i, s, a, $\phi(a, s) \neq i \Rightarrow \ell(u^{(i)}, a) \geq \varepsilon + \Delta$. Instantiate $i = J$, $a = \pi(Y, S)$, $s = S$ and note $U = U^{(J)}$ to obtain the pointwise bound

$$\ell\big(U, \pi(Y, S)\big) \geq (\varepsilon + \Delta)\, \mathbf{1}\{\hat{J} \neq J\} \qquad \text{a.s.}$$

Since $\ell \in [0, 1]$, all terms are integrable and π, ϕ are measurable by assumption; taking expectations yields equation 11. □

Lemma 2 (Information Reduction: $J \to U \to Y$)**.** *With $U = U^{(J)}$ measurable in (J, S), we have the Markov chain $J \to U \to Y$ given S, and hence*

$$I(J; Y \mid S) \leq I(U; Y \mid S). \tag{12}$$

Theorem 1 (Fano–Packing Lower Bound (Bayes/Minimax Semantics))**.** *Let $\ell \in [0, 1]$ and suppose there exists a Δ-separable codebook of size M satisfying Assumption 2. Let $J \sim \mathrm{Unif}[M]$, and define the mixture distribution over (U, S) by setting $U = U^{(J)}$, with $U^{(J)}$ measurable in (J, S), and $S \sim P(S)$. Assume $M \geq 2$. Write $R_{\mathrm{mix}}(\pi)$ for the risk under this mixture distribution. Then, for any decoder π,*

Then for any decoder π,

$$R_{\mathrm{mix}}(\pi) \geq (\varepsilon + \Delta) \Big(1 - \frac{I(U; Y \mid S) + \log 2}{\log M}\Big)_+. \tag{13}$$

In particular, using equation 7,

$$R_{\mathrm{mix}}(\pi) \geq (\varepsilon + \Delta) \Big(1 - \frac{\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S} + \log 2}{\log M}\Big)_+. \tag{14}$$

Equivalently, these yield a standard minimax lower bound over the family of sources supported on the codebook.

Proof. Let J be uniform on $[M]$, $U = U^{(J)}$. Since J is independent of S, we have $H(J \mid S) = \log M$. By Fano's inequality conditioned on S (Shannon, 1948),

$$\mathbb{P}\{\hat{J} \neq J\} \geq 1 - \frac{I(J; Y \mid S) + \log 2}{\log M}.$$

Combine with Lemma 1 and Lemma 2 to get equation 13. Then apply equation 7. □

4.3 Capacity–Limited Achievability and the Information Wall

Define the information wall for a problem class $\mathcal{P}$ (set of admissible codebooks) by

$$\mathsf{Wall}\big(\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S};\mathcal{P}\big) \;\triangleq \sup_{(M,\Delta,\varepsilon):\ \mathcal{C}(M,\Delta,\varepsilon)\in\mathcal{P}} (\varepsilon+\Delta)\left(1-\frac{\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S}+\log 2}{\log M}\right)_+. \tag{15}$$

Here $\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S}$ is evaluated under the codebook-induced mixture distribution used in Theorem 1.

Then, for every decoder π,

$$\sup_{\mathcal{C}\in\mathcal{P}} R^{\mathcal{C}}_{\mathrm{mix}}(\pi) \geq \mathsf{Wall}(\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S};\mathcal{P})$$

where $R^{(M,\Delta,\varepsilon)}_{\mathrm{mix}}(\pi)$ denotes the risk under the mixture induced by the chosen codebook (Bayes/minimax semantics from Thm. 1). Equivalently, this yields the standard minimax lower bound

$$\inf_{\pi}\sup_{\mathcal{C}\in\mathcal{P}} R^{\mathcal{C}}_{\mathrm{mix}}(\pi) \geq \mathsf{Wall}(\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S};\mathcal{P}).$$

This replaces log-loss/posterior-entropy converses and is invariant to reparameterizations.

5 PAC–Bayes Upper Bounds

With the floor in place, we now derive the upper bound. We use PAC–Bayes, a non-asymptotic framework suited to overparameterized learners (including LLMs) with non-vacuous guarantees at scale (Lotfi et al., 2024). Our aim is not to introduce a new bound but to make its complexity term, $\mathrm{KL}(P\|Q)$, capacity-explicit in the same $\bar{C}_{\mathrm{tot}|S}$ that drives the converse, closing the loop between the lower wall and the statistical ceiling.

5.1 PAC–Bayes Bounds

We recall a standard PAC–Bayes result for bounded losses as the basis of the ceiling. Recent variants tighten constants, cover heavier tails, and allow anytime validity, yielding non-vacuous bounds even for billion-parameter LLMs (Dziugaite and Roy, 2017; Rodríguez-Gálvez et al., 2024; Lotfi et al., 2024; Wu et al., 2025). Tightness is not automatic: strong guarantees require priors that put sufficient mass on high-performing predictors (Picard-Weibel et al., 2025). This interacts with the Alignment Bottleneck: finite human-feedback capacity limits how informative data-independent priors can be, and this constraint enters through the KL term that we bound via $\bar{C}_{\mathrm{tot}|S}$.

Theorem 2 (PAC–Bayes for Observable Loss)**.** *Let $\tilde{\ell}:\mathcal{Y}\times\mathcal{S}\times\mathcal{A}\to[0,1]$ be any bounded loss measurable with respect to the observed data (Y,S). For any $\delta\in(0,1)$, with probability at least $1-\delta$ over the i.i.d. draw of the dataset $\mathcal{D}=\{(Y_i,S_i)\}_{i=1}^m$,*

$$\mathbb{E}_{\theta\sim P}\big[R_{\mathrm{obs}}(\theta)\big] \;\leq\; \mathbb{E}_{\theta\sim P}\big[\widehat{R}^{\mathrm{obs}}_m(\theta)\big] \;+\; \sqrt{\frac{\mathrm{KL}(P\|Q)+\log(1/\delta)}{2m}}, \tag{16}$$

where

$$R_{\mathrm{obs}}(\theta) \;\triangleq\; \mathbb{E}\big[\tilde{\ell}\big(Y,S,\pi_\theta(Y,S)\big)\big], \qquad \widehat{R}^{\mathrm{obs}}_m(\theta) \;\triangleq\; \frac{1}{m}\sum_{i=1}^m \tilde{\ell}\big(Y_i,S_i,\pi_\theta(Y_i,S_i)\big).$$

Canonical choice. If we choose $\tilde{\ell}=\tilde{\ell}^\star$ as in Appendix M, then $R_{\mathrm{obs}}(\theta)=R(\pi_\theta)$ holds for the same data distribution.

5.2 KL Decomposition and Capacity Control

Lemma 3 (Expected KL Decomposition)**.** *Fix a prior Q that is independent of the dataset $\mathcal{D}$. Let $p(\theta)$ be the marginal of θ and $P(\cdot\mid\mathcal{D})$ be the posterior. Then*

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{KL}(P\|Q)\big] \;=\; I(\mathcal{D};\theta) \;+\; \mathrm{KL}\big(p(\theta)\,\|\,Q\big). \tag{17}$$

This identity underlies information-theoretic generalization bounds and is central to our analysis. Russo and Zou (2019) and Xu and Raginsky (2017) relate generalization directly to the mutual information $I(\mathcal{D};\theta)$ between the data and the learned hypothesis. We take this link as given and show that $I(\mathcal{D};\theta)$ is constrained by the capacity of the human-feedback channel.

Lemma 4 (From $\mathcal{D}$ to (U^m, S^m, Y^m))**.** *Assume samples (U_i, S_i) are i.i.d., and Y_i are drawn conditionally independently via the human channel given (U_i, S_i) as in Assumption 1. Let θ be any (possibly randomized) function of $\mathcal{D} \triangleq \{(Y_i, S_i)\}_{i=1}^m$. Then*

$$\begin{aligned} I(U^m;\theta) \;&\le\; I(U^m; Y^m, S^m) \;=\; I(U^m; S^m) \;+\; I(U^m; Y^m \mid S^m) \\ &=\; \sum_{i=1}^m I(U_i; S_i) \;+\; \sum_{i=1}^m I(U_i; Y_i \mid S_i) \;=\; m\, I(U;S) \;+\; \sum_{i=1}^m I(U_i; Y_i \mid S_i). \end{aligned} \tag{18}$$

Under the i.i.d. source and memoryless per-sample channel assumed in this paper, all equalities in equation 18 hold. If either cross-sample dependence in (U_i, S_i) or channel memory in $p(y_i \mid u^m, s^m)$ is allowed, replace the corresponding equalities by "$\le$" accordingly.

Proposition 2 (Capacity Control of $I(U^m;\theta)$)**.** *Using equation 7 and Lemma 4,*

$$I(U^m;\theta) \;\le\; m\,\bar{C}_{\mathrm{tot}|S} \;+\; m\, I(U;S). \tag{19}$$

Convention. *All mutual informations in this section are defined with respect to the underlying data-generating distribution (population quantities), and $\bar{C}_{\mathrm{tot}|S}$ is computed under the same source distribution; no averaging over the realized dataset is involved.*

5.3 Algorithmic Residual Information

Assumption 3 (Residual Information of the Algorithm)**.** *There exists $\rho \ge 0$ such that $I(\mathcal{D};\theta \mid U^m) \le \rho$. It can be reduced by algorithmic noise (SGD temperature), early stopping, or posterior smoothing; see Appx. I. This term measures information about the particular sample beyond the latent value U and parallels the "residual information" used in information-theoretic analyses of overfitting (Ngampruetikorn and Schwab, 2022).*

Practically, a data-independent randomized compression of the posterior enforces a finite residual, giving $\rho \le \log K$ for any chosen codebook size K without increasing $\mathrm{KL}(P\|Q)$ (see Appendix N). The idea of limiting information flow to improve generalization is widespread, though the causal link between compression and performance remains under debate (Kawaguchi et al., 2023; Saxe et al., 2019; Shwartz-Ziv et al., 2024). Here ρ isolates information learned from (Y, S) that is not about U, which is the target of such regularization.

Corollary 1 (A Capacity-Aware Upper Bound)**.** *Combining Lemma 3, Proposition 2, and Assumption 3, we have*

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{KL}(P\|Q)\big] \;\le\; m\,\bar{C}_{\mathrm{tot}|S} \;+\; m\, I(U;S) \;+\; \rho \;+\; \mathrm{KL}\big(p(\theta)\,\|\,Q\big). \tag{20}$$

Equation equation 20 controls the expectation of $\mathrm{KL}(P\|Q)$ over the draw of $\mathcal{D}$ and does not by itself yield a capacity-explicit high-probability bound. Appendix J gives a Markov-type lifting to high probability. Taking expectations in Thm. 2 and applying Jensen yields corresponding in-expectation variants.

Remark 1 (Conservative Interpretation)**.** *When ρ or $I(U;S)$ is large, capacity may not dominate the upper bound. Our statements should be read as: under Assumption 3 and moderate $I(U;S)$, both the converse (Thm. 1) and the PAC–Bayes upper bound are primarily driven by $\bar{C}_{\mathrm{tot}|S}$.*

6 The Alignment Performance Interval

The preceding sections developed two components: an information-theoretic error floor via Fano's inequality (Section 4) and a statistical error ceiling via PAC–Bayes theory (Section 5). We now establish the Alignment Performance Interval. The same capacity term (the channel capacity $\bar{C}_{\mathrm{tot}|S}$) determines the lower bound and, at the same time, limits the learnable model complexity that determines the generalization upper bound.

6.1 CAPACITY-COUPLED BOUNDS

Let $\mathcal{P}$ be a collection of codebooks $\mathcal{C}(M, \Delta, \varepsilon)$. For any learning algorithm (decoder) π, its worst-case true risk under a codebook-induced mixture distribution is bounded from below by the information-theoretic wall:

$$\textbf{Lower (Minimax):} \qquad \sup_{\mathcal{C}\in\mathcal{P}} R^{\mathcal{C}}_{\mathrm{mix}}(\pi) \;\geq\; \mathsf{Wall}\big(\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S}; \mathcal{P}\big) \quad \text{from Thm. 1.} \tag{21}$$

Simultaneously, for any prior Q and posterior P, the expected true risk is bounded from above. With probability $\geq 1-\delta$ over the draw of a dataset $\mathcal{D}$ from the same mixture, and using the canonical observable loss $\tilde{\ell}^\star$ (Appendix M) such that $R_{\mathrm{obs}}(\theta) = R_{\mathrm{mix}}(\pi_\theta)$, we have:

$$\textbf{Upper (High-probability):} \qquad \mathbb{E}_{\theta\sim P}\big[R_{\mathrm{mix}}(\pi_\theta)\big] \;\leq\; \mathbb{E}_{\theta\sim P}\big[\widehat{R}^{\mathrm{obs}}_m(\theta)\big] + \sqrt{\frac{\mathrm{KL}(P\|Q) + \log(1/\delta)}{2m}}. \tag{22}$$

This is a direct application of Theorem 2 to the true risk R_{mix}. As shown in Corollary 1, the expected KL-divergence term is controlled by the channel capacity, $\mathbb{E}_{\mathcal{D}}[\mathrm{KL}(P\|Q)] \leq m\,\bar{C}_{\mathrm{tot}|S} + \ldots$, thus explicitly coupling the ceiling to the same capacity term that defines the floor. Together, equations equation 21 and equation 22 yield two-sided bounds on the same risk quantity, R_{mix}, driven by $\bar{C}_{\mathrm{tot}|S}$.

Interpretation. The two bounds control different risks: the Bayes/minimax lower bound applies to the true risk under the mixture distribution R_{mix}, whereas the PAC–Bayes upper bound applies to the observable risk R_{obs} under the actual data distribution. Without an explicit link between ℓ and $\tilde{\ell}$ and without a distribution match, they should not be treated as an interval on the same quantity. Under the Loss–Observable Link (Assumption 4) in Appx. L and when $\mathcal{D}$ is drawn from the same codebook-induced mixture used in Thm. 1, we obtain the following direct upper bound on the true risk (by Appx. Lemma 6 combined with equation 22):

$$\mathbb{E}_{\theta\sim P}\big[R_{\mathrm{mix}}(\pi_\theta)\big] \;\leq\; \alpha\Big(\mathbb{E}_{\theta\sim P}\big[\widehat{R}^{\mathrm{obs}}_m(\theta)\big] \;+\; \sqrt{\tfrac{\mathrm{KL}(P\|Q)+\log(1/\delta)}{2m}}\Big) \;+\; \beta\,.$$

Together with equation 21, this yields two-sided bounds on the same quantity R_{mix}, with explicit constants (α, β) coming from the link assumption.

Finally, if the dataset $\mathcal{D}$ is drawn from the same codebook-induced mixture as in Theorem 1 and we take the canonical observable loss $\tilde{\ell} = \tilde{\ell}^\star$ (Appendix M), then $R_{\mathrm{obs}}(\theta) = R_{\mathrm{mix}}(\pi_\theta)$ and Eq. equation 22 becomes a high-probability upper bound on the same risk R_{mix} as in the converse; together with Eq. equation 21, this yields a two-sided bound without additional link assumptions. We detail the practical assumptions and limitations of this framework in Appendix B.

7 IMPLICATIONS FOR ALIGNMENT DESIGN

The Alignment Performance Interval (Sec. 6) is operational: it explains practical alignment limits and suggests design levers. We highlight three implications that follow directly from the lower and upper bounds established earlier.

7.1 IMPLICATION I: DATA SIZE INDEPENDENT LOWER BOUND

Corollary 2 (Information-theoretic lower bound independent of m)**.** *Let $\mathcal{C}(M, \Delta, \varepsilon)$ be any Δ-separable codebook with $M \geq 2$, and let R_{mix} denote the risk under its mixture distribution (as in Thm. 1). For any decoder π,*

$$R_{\mathrm{mix}}(\pi) \;\geq\; (\varepsilon + \Delta)\left(1 - \frac{\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S} + \log 2}{\log M}\right)_+, \tag{23}$$

which is exactly equation 14. The bound equation 23 does not depend on m, hence the lower bound is independent of dataset size.

Eq. equation 23 shows a lower bound that does not depend on m: for fixed value complexity ($\log M$) and channel capacity ($\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S}$), more samples alone cannot lower the risk. This helps interpret the

empirical alignment tax as an information constraint (Lin et al., 2024a; Korkmaz et al., 2025) and may help explain inverse-scaling effects on truthfulness/safety when models more tightly fit the feedback channel (Lin et al., 2022).

Note. Using the canonical observable loss $\tilde{\ell}^\star$ (Appx. M) and sampling $\mathcal{D}$ from the same mixture as in Thm. 1, the PAC–Bayes upper bound equation 22 applies to the *same* R_{mix}, yielding a two-sided bound together with equation 23.

7.2 Implication II: Capacity Requirements for Target Risk

Proposition 3 (Necessary capacity for a target risk)**.** *Fix a codebook $\mathcal{C}(M, \Delta, \varepsilon)$ and a target risk $r \in [0, 1]$. If a decoder π satisfies $R_{\mathrm{mix}}(\pi) \le r$, then necessarily*

$$\bar{C}^{\mathrm{mix}}_{\mathrm{tot}|S} \;\ge\; \Big(1 - \frac{r}{\varepsilon + \Delta}\Big)\,\log M \;-\; \log 2. \tag{24}$$

Proof. Rearrange equation 14; the $(\cdot)_+$ can be dropped once $r < \varepsilon + \Delta$ (otherwise the inequality is vacuous but true). □

The required channel capacity scales linearly with $\log M$—a proxy for value-system complexity—thus, aligning on more complex, pluralistic targets (Sorensen et al., 2024; Guo et al., 2024; Fisher et al., 2025) demands proportionally higher fidelity, mirroring the rate-distortion trade-offs in communication theory (Shannon, 1959).

7.3 Implication III: Capacity Controlled Complexity and Channel Overfitting

Theorem 3 (Capacity-controlled PAC–Bayes complexity)**.** *Under the i.i.d. source and memoryless channel, with a prior Q independent of $\mathcal{D}$ and any learning algorithm whose residual satisfies Assumption 3, the expected PAC–Bayes complexity obeys*

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{KL}(P\|Q)\big] \;\le\; m\,\bar{C}_{\mathrm{tot}|S} \;+\; m\,I(U;S) \;+\; \rho \;+\; \mathrm{KL}\big(p(\theta)\|Q\big), \tag{25}$$

as given in Cor. 1. Combining equation 25 with the Markov lift in Appx. J and equation 22 yields a high-probability capacity-explicit *upper bound on the (observable or, under the canonical choice, true) risk.*

An Information-Theoretic View of Overfitting to the Channel. When $\widehat{R}^{\mathrm{obs}}_m(\theta) \approx 0$ but a small $\bar{C}_{\mathrm{tot}|S}$ imposes a strong lower bound, the KL term must grow. Decomposing

$$I(\mathcal{D};\theta) \;=\; \underbrace{I(U^m;\theta)}_{\text{signal about true value}} \;+\; \underbrace{I(\mathcal{D};\theta \mid U^m)}_{\text{residual: channel noise/bias}}\;,$$

the useful signal is capped by capacity (Prop. 2), so further optimization fits residual channel regularities (Ngampruetikorn and Schwab, 2022). This mechanism aligns with observations of goal misgeneralization, sycophancy, and reward hacking under strong optimization pressure (Langosco et al., 2023; Sharma et al., 2024; Pan et al., 2022; Gaikwad, 2025; Lin et al., 2024b).

8 Conclusion

Motivated by bounded rationality, we model the human-AI loop as a capacity-limited channel, yielding an Alignment Performance Interval where Fano and PAC-Bayes bounds are coupled by the same channel capacity, $\bar{C}_{\mathrm{tot}|S}$. This framework explains why simply scaling data is insufficient, quantifies how fidelity must grow with value complexity, and frames reward hacking as overfitting to the channel's limits. Practical applications, current limitations, and future work all center on managing this information budget—from engineering the interface and verifying assumptions to developing capacity-aware protocols. Ultimately, our work reframes alignment from a search for optimal rewards to the engineering of systems robust to the fundamental information-theoretic limits of the human-AI interface.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. DEEP VARIATIONAL INFORMATION BOTTLENECK. 2017.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

Dilip Arumugam and Benjamin Van Roy. The Value of Information When Deciding What to Learn. 2021.

Dilip Arumugam and Benjamin Van Roy. Deciding What to Model: Value-Equivalent Sampling for Reinforcement Learning, October 2022. URL http://arxiv.org/abs/2206.02072. arXiv:2206.02072 [cs].

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 373–382. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/arumugam21a.html.

Dilip Arumugam, Mark K. Ho, Noah D. Goodman, and Benjamin Van Roy. Bayesian Reinforcement Learning with Limited Cognitive Load, May 2023. URL http://arxiv.org/abs/2305.03263. arXiv:2305.03263 [cs].

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022b. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. 2017.

Juntao Dai, Taiye Chen, Yaodong Yang, Qian Zheng, and Gang Pan. MITIGATING REWARD OVER-OPTIMIZATION IN RLHF VIA BEHAVIOR-SUPPORTED REGULARIZATION. 2025.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. LANGUAGE MODELING IS COMPRESSION. 2024.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to Subterfuge: Investigating Reward-Tampering

in Large Language Models, June 2024. URL http://arxiv.org/abs/2406.10162. arXiv:2406.10162 [cs].

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data, 2017. URL https://arxiv.org/abs/1703.11008.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model Alignment as Prospect Theoretic Optimization, November 2024. URL http://arxiv.org/abs/2402.01306. arXiv:2402.01306 [cs].

Charles Evans and Atoosa Kasirzadeh. User Tampering in Reinforcement Learning Recommender Systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 58–69, August 2023. doi: 10.1145/3600211.3604669. URL http://arxiv.org/abs/2109.04083. arXiv:2109.04083 [cs].

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective, March 2021. URL http://arxiv.org/abs/1908.04734. arXiv:1908.04734 [cs].

Jillian Fisher, Ruth E Appel, Chan Young Park, Yujin Potter, Liwei Jiang, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret E Roberts, Jennifer Pan, Dawn Song, and Yejin Choi. Position: Political Neutrality in AI Is Impossible — But Here Is How to Approximate It. 2025.

Madhava Gaikwad. Murphys Laws of AI Alignment: Why the Gap Always Wins, September 2025. URL http://arxiv.org/abs/2509.05381. arXiv:2509.05381 [cs].

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI Alignment from Human Feedback. 2024.

Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2299–2308. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/goldfeld19a.html.

Sebastian Gottwald and Daniel A. Braun. Bounded rational decision-making from elementary computations that reduce uncertainty. *Entropy*, 21(4):375, April 2019. ISSN 1099-4300. doi: 10.3390/e21040375. URL http://arxiv.org/abs/1904.03964. arXiv:1904.03964 [cs].

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1454, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.85. URL https://aclanthology.org/2024.emnlp-main.85.

Haiyun He, Christina Lee Yu, and Ziv Goldfeld. Information-Theoretic Generalization Bounds for Deep Neural Networks. 2025.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017. URL https://arxiv.org/abs/1712.00409.

Mark K. Ho and Thomas L. Griffiths. Cognitive Science as a Source of Forward and Inverse Models of Human Decisions for Robotics and Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):33–53, May 2022. ISSN 2573-5144, 2573-5144. doi: 10.1146/annurev-control-042920-015547. URL https://www.annualreviews.org/doi/10.1146/annurev-control-042920-015547.

Mark K Ho, David Abel, Jonathan D Cohen, Michael L Littman, and Thomas L Griffiths. The Efficiency of Human Cognition Reflects Planned Information Processing. 2020.

Mark K. Ho, David Abel, Carlos G. Correa, Michael L. Littman, Jonathan D. Cohen, and Thomas L. Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, June 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-022-04743-9. URL http://arxiv.org/abs/2105.06948. arXiv:2105.06948 [cs].

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, October 2018. URL http://arxiv.org/abs/1805.00899. arXiv:1805.00899 [stat].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How Does Information Bottleneck Help Deep Learning?, May 2023. URL http://arxiv.org/abs/2305.18887. arXiv:2305.18887 [cs].

Artemy Kolchinsky, Brendan D. Tracey, and David H. Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, November 2019. ISSN 1099-4300. doi: 10.3390/e21121181. URL http://dx.doi.org/10.3390/e21121181.

Buse Sibel Korkmaz, Rahul Nair, Elizabeth M. Daly, and Antonio del Rio Chanona. Paying Alignment Tax with Contrastive Learning, May 2025. URL http://arxiv.org/abs/2505.19327. arXiv:2505.19327 [cs].

Le Lai and Samuel J. Gershman. Policy compression: An information bottleneck in action selection. In Kara D. Federmeier, editor, *The psychology of learning and motivation*, pages 195–232. Elsevier Academic Press, 2021. doi: 10.1016/bs.plm.2021.02.004.

John Langford and Rich Caruana. (not) bounding the true error. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/98c7242894844ecd6ec94af67ac8247d-Paper.pdf.

Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, and David Krueger. Goal Misgeneralization in Deep Reinforcement Learning. 2023.

Benjamin Leblanc, Mathieu Bazinet, Nathaniel D'Amours, Alexandre Drouin, and Pascal Germain. Generalization Bounds via Meta-Learned Model Representations:PAC-Bayes and Sample Compression Hypernetworks. 2025.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=uydQ2W41KO.

Richard L. Lewis, Andrew Howes, and Satinder Singh. Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization. *Topics in Cognitive Science*, 6(2):279–311, April 2014. ISSN 1756-8757, 1756-8765. doi: 10.1111/tops.12086. URL https://onlinelibrary.wiley.com/doi/10.1111/tops.12086.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the Alignment Tax of RLHF. 2024a.

Yong Lin, Skyler Seto, Maartje Ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, and Tong Zhang. On the Limited Generalization Capability of the Implicit Reward Model Induced by Direct Preference Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16015–16026, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.940. URL https://aclanthology.org/2024.findings-emnlp.940.

Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-Vacuous Generalization Bounds for Large Language Models, July 2024. URL http://arxiv.org/abs/2312.17173. arXiv:2312.17173 [stat].

Yuchun Miao, Sen Zhang, Liang Ding, Lefei Zhang, and Dacheng Tao. InfoRM: Mitigating Reward Hacking in RLHF via Information-Theoretic Reward Modeling. 2024.

Youssef Mroueh and Apoorva Nitsure. Information theoretic guarantees for policy alignment in large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Uz9J77Riul.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 108877–108901. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c4e380fb74dec9da9c7212e834657aa9-Paper-Conference.pdf.

Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/10ce03a1ed01077e3e289f3e53c72813-Paper.pdf.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Vudtiwat Ngampruetikorn and David J Schwab. Information bottleneck theory of high-dimensional regression: relevancy, efficiency and optimality. 2022.

Richard Ngo, Lawrence Chan, and Sören Mindermann. THE ALIGNMENT PROBLEM FROM A DEEP LEARNING PERSPECTIVE. 2024.

Pedro A Ortega and Daniel A Braun. Information, Utility & Bounded Rationality. 2011.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. THE EFFECTS OF REWARD MISSPECIFICATION: MAPPING AND MITIGATING MISALIGNED MODELS. 2022.

Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow, August 2020. URL http://arxiv.org/abs/1810.00821. arXiv:1810.00821 [cs].

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang,

Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847.

Antoine Picard-Weibel, Eugenio Clerico, Roman Moscoviz, and Benjamin Guedj. How good is PAC-Bayes at explaining generalisation?, March 2025. URL http://arxiv.org/abs/2503.08231. arXiv:2503.08231 [stat].

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=pf4OuJyn4Q.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, July 2024b. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs].

Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L. Griffiths. Concept Alignment, January 2024. URL http://arxiv.org/abs/2401.08672. arXiv:2401.08672 [cs].

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity, June 2024. URL http://arxiv.org/abs/2306.12214. arXiv:2306.12214 [stat].

Daniel Russo and James Zou. How much does your data exploration overfit? Controlling bias via information usage, October 2019. URL http://arxiv.org/abs/1511.05219. arXiv:1511.05219 [stat].

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning*. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, December 2019. ISSN 1742-5468. doi: 10.1088/1742-5468/ab3985. URL https://iopscience.iop.org/article/10.1088/1742-5468/ab3985.

Claude Shannon. Coding Theorems for a Discrete Source With a Fidelity Criterion. 1959.

Claude Elwood Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. TOWARDS UNDERSTANDING SYCOPHANCY IN LANGUAGE MODELS. 2024.

Ravid Shwartz-Ziv and Yann LeCun. To Compress or Not to Compress- Self-Supervised Learning and Information Theory: A Review, November 2023. URL http://arxiv.org/abs/2304.09355. arXiv:2304.09355 [cs].

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information, 2017. URL https://arxiv.org/abs/1703.00810.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G J Rudner, and Yann LeCun. An Information-Theoretic Perspective on Variance-Invariance-Covariance Regularization. 2024.

Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69 (1):99–118, 1955.

Chris R. Sims. Rate–distortion theory and human perception. *Cognition*, 152:181–198, July 2016. ISSN 00100277. doi: 10.1016/j.cognition.2016.03.020. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027716300750.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A Roadmap to Pluralistic Alignment, August 2024. URL http://arxiv.org/abs/2402.05070. arXiv:2402.05070 [cs].

Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle, March 2015. URL http://arxiv.org/abs/1503.02406. arXiv:1503.02406 [cs].

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, April 2000. URL http://arxiv.org/abs/physics/0004057. arXiv:physics/0004057.

Zifeng Wang, Shao-Lun Huang, Ercan E Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. PAC-BAYES INFORMATION BOTTLENECK. 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

Yi-Shan Wu, Yijie Zhang, Badr-Eddine Chérief-Abdellatif, and Yevgeny Seldin. Recursive PAC-Bayes: A Frequentist Approach to Sequential Prior Updates with No Information Loss. 2025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. 2017.

Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. A Rate–Distortion view of human pragmatic reasoning? In Allyson Ettinger, Ellie Pavlick, and Brandon Prickett, editors, *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348, Online, February 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.scil-1.32/.

Alexandre Zénon, Oleg Solopchuk, and Giovanni Pezzulo. An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123:5–18, February 2019. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2018.09.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0028393218306328.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences, January 2020. URL http://arxiv.org/abs/1909.08593. arXiv:1909.08593 [cs].

A Notation and Cross-References

Throughout, $\log$ denotes the natural logarithm. Key references: single capacity inequality equation 7; Fano–packing converse Thm. 1; PAC–Bayes Thm. 2; expected KL identity Lemma 3; dataset information Lemma 4; capacity-control Proposition 2; Posterior Bayes–Loss Identity Lemma 7; Assumptions 2, 3; final statements in Sec. 6.

B Limitations and Guidance

The loss–index link (Assump. 2) must be validated for each task; templates are given in Appx. F. Capacity $\bar{C}_{\mathrm{tot}|S}$ enters only via equation 7, so deployment should report how $\mathcal{F}_{\mathrm{cog}}$ and $\mathcal{F}_{\mathrm{art}}$ instantiate per context S. The residual ρ (Assump. 3) should be promoted by algorithmic choices, or reported if uncontrolled.

C Remarks on Capacity Definitions and Context Coarsening

Remark 2 (Source-dependent min and achievability)**.** *Definitions 3–5 impose per-stage feasibility (Assumption 1), so equation 6 follows from data processing without additional compatibility assumptions. The conditional quantities $C_{\mathrm{cog}|S}(s)$ and $C_{\mathrm{art}|S}(s)$ are evaluated under the given source $P(U \mid S = s)$ rather than maximized over inputs; they are therefore analogous to rate–distortion quantities computed for a fixed source (Shannon, 1959) and depend on $P(U \mid S)$. Finally, $\min\{\sup I(U; H \mid S = s), \sup I(H; Y \mid S = s)\}$ is in general a conservative upper bound for the cascade because the per-stage optimizers need not be mutually compatible, so equality (achievability) should not be expected.*

All statements also hold if one replaces the context S by any measurable coarsening $S' = T(S)$: then the interference term $I(U; S)$ is replaced by $I(U; S') \le I(U; S)$ and capacities are recomputed as $\bar{C}_{\mathrm{tot}|S'}$.

D Conditional Capacities and the Cascade

Proof of Proposition 1. The cascade $U \to H \to Y$ given S forms a Markov chain. The result follows directly from the data processing inequality (Shannon, 1948), which states that for such a chain, $I(U; Y \mid S = s) \le I(U; H \mid S = s)$ and $I(U; Y \mid S = s) \le I(H; Y \mid S = s)$. Taking suprema over the respective families yields $I(U; Y \mid S = s) \le C_{\mathrm{tot}|S}(s)$. Averaging over S proves equation 7.

E Codebook Construction Examples

Remark 3 (How to Build $\mathcal{C}(M, \Delta, \varepsilon)$ in Practice)**.** *The construction of such codebooks (packings) is standard for minimax lower bounds in statistical decision theory and information theory (Shannon, 1948). For 0–1 classification, choose $a^{(i)}$ predicting class i, giving $\varepsilon = 0$ and $\Delta = 1$. For pairwise ranking with 0–1* pairwise *loss averaged over all $\binom{n}{2}$ pairs, choose $a^{(i)}$ realizing ranking i, so $\varepsilon = 0$ and any two total orders differ on at least one pair, yielding $\Delta = 1/\binom{n}{2}$ after normalization to $[0, 1]$. For truncated MSE $\ell(u, a) = \min\{\|u - a\|^2/\tau^2, 1\}$, take $a^{(i)} = u^{(i)}$ on an r-separated packing of $\mathcal{U}$, then $\varepsilon = 0$ and it is consistent with Assump. 2 to use the common margin $\Delta = r^2/(4\tau^2)$ (the prototype cross-loss is $\ge r^2/\tau^2 \ge \Delta$, while the Voronoi misclassification loss is $\ge r^2/(4\tau^2)$).*

F Packing Constructions for Common Losses

F.1 Binary Classification

Let $\mathcal{U} = [M]$ and $\mathcal{A}$ the set of labels. Take $a^{(i)} = i$. Then $\varepsilon = 0$ and for any $j \ne i$, $\ell(u^{(j)}, a^{(i)}) = 1$, giving $\Delta = 1$. Let ϕ be the predicted label; Assump. 2 holds with margin 1.

G Details on the Loss-Index Link

Remark 4 (When Assumption 2 Holds). *For 0–1 classification and pairwise ranking, let ϕ return the predicted class/ranking (allowing dependence on S if needed); then equation 10 holds immediately. For truncated MSE, let ϕ be the nearest-prototype Voronoi partition under $\|\cdot\|$ (prototypes may depend on S). With an r-separated packing, any misclassification implies $\|a - u^{(i)}\| \geq r/2$, hence $\ell(u^{(i)}, a) \geq r^2/(4\tau^2)$, so equation 10 holds with the common choice $\Delta = r^2/(4\tau^2)$; equivalently, the separation condition is $r \geq 2\tau\sqrt{\Delta}$. See Appx. F.*

A soft high-probability variant of the loss–index link that yields a correspondingly slackened converse is provided in Appendix P.

G.1 Pairwise Ranking with 0–1 Loss

Let $u^{(i)}$ encode a total order over items and ℓ be the fraction of misordered pairs. Use $a^{(i)} = u^{(i)}$ (predict that order). Then $\varepsilon = 0$ and for any $j \neq i$, at least one pair flips, so $\Delta \geq 1/\binom{n}{2}$; with standard $\{0,1\}$ pairwise loss averaged over all $\binom{n}{2}$ pairs and normalized to $[0,1]$, the minimal separation is $\Delta = 1/\binom{n}{2}$. Let ϕ output the predicted order; Assump. 2 holds.

G.2 Truncated and Normalized MSE

Let $\ell(u, a) = \min\{\|u-a\|^2/\tau^2, 1\}$. Choose an r-separated packing $\{u^{(i)}\}_{i=1}^M$ in $\mathcal{U}$ (under $\|\cdot\|$), and set $a^{(i)} = u^{(i)}$. Then $\varepsilon = 0$ and, for $j \neq i$, the prototype cross-loss satisfies $\ell(u^{(j)}, a^{(i)}) \geq r^2/\tau^2$. To make Definition 6 and Assumption 2 hold with a single margin, take the common choice

$$\Delta \;=\; \frac{r^2}{4\tau^2}\,.$$

Indeed, for any a misclassified by the nearest-prototype Voronoi rule, one has $\|a - u^{(i)}\| \geq r/2$, so $\ell(u^{(i)}, a) \geq r^2/(4\tau^2) = \Delta$. Since $r^2/\tau^2 \geq \Delta$, the prototype cross-loss condition in Definition 6 also holds.

H Fano–Packing Converse Details

We expand the proof of Thm. 1. Let J be uniform on $[M]$, $U = U^{(J)}$. With $\hat{J} = \phi(\pi(Y, S), S)$, Lemma 1 gives $R(\pi) \geq (\varepsilon + \Delta)\,\mathbb{P}\{\hat{J} \neq J\}$. The standard form of Fano's inequality (Shannon, 1948), when conditioned on S, implies that $H(J \mid Y, S) \leq \mathbb{P}\{\hat{J} \neq J\}\log(M-1) + h_2(\mathbb{P}\{\hat{J} \neq J\})$, which gives the more convenient bound

$$\mathbb{P}\{\hat{J} \neq J\} \;\geq\; 1 - \frac{I(J; Y \mid S) + \log 2}{\log M}.$$

Using $J \to U \to Y$ given S (Lemma 2), we get equation 13; then apply equation 7 for equation 14.

I PAC–Bayes Details and Residual Control

I.1 Proofs of Lemma 3 and Lemma 4

Lemma 3. The identity is a foundational result in information-theoretic learning theory (Xu and Raginsky, 2017; Russo and Zou, 2019). The proof is as follows: with Q independent of $\mathcal{D}$, $\mathbb{E}_{\mathcal{D}}[\mathrm{KL}(P\|Q)] = \mathbb{E}_{\mathcal{D},\theta\sim P}\big[\log \frac{P(\theta|\mathcal{D})}{Q(\theta)}\big] = I(\mathcal{D};\theta) + \mathrm{KL}(p(\theta)\|Q)$.

Lemma 4. Data processing gives $I(U^m;\theta) \leq I(U^m;\mathcal{D})$. Then $I(U^m; Y^m, S^m) = I(U^m; S^m) + I(U^m; Y^m \mid S^m)$, with $I(U^m; S^m) = \sum_i I(U_i; S_i) = mI(U; S)$ by i.i.d. For the conditional term, under the i.i.d. source and the memoryless channel $p(y_i \mid u_i, s_i)$, we have $p(u^m \mid s^m) = \prod_i p(u_i \mid s_i)$ and hence $p(y^m \mid s^m) = \prod_i \int p(y_i \mid u_i, s_i)\, p(u_i \mid s_i)\, du_i = \prod_i p(y_i \mid s_i)$. Therefore $H(Y^m \mid S^m) = \sum_i H(Y_i \mid S_i)$ and $H(Y^m \mid U^m, S^m) = \sum_i H(Y_i \mid U_i, S_i)$, which gives $I(U^m; Y^m \mid S^m) = \sum_i I(U_i; Y_i \mid S_i)$.

I.2 Controlling the Residual Term

We list standard mechanisms to enforce Assump. 3. These methods all serve to regularize the information that the learned parameters θ contain about the specific training dataset $\mathcal{D}$. *Algorithmic noise*: inject Gaussian noise into updates or use high-temperature posteriors; *early stopping*: bound the mutual information by limiting the number of optimization steps; *posterior smoothing*: mix the learned posterior with the prior. The general goal of controlling information flow, often framed as a form of compression, is a central theme in understanding deep learning generalization, although its precise role and benefits are still actively debated (Kawaguchi et al., 2023; Saxe et al., 2019; Shwartz-Ziv et al., 2024; He et al., 2025).

J From Expectation to High Probability

This section provides a simple method to convert our expectation-based capacity bound on the KL-divergence into a high-probability statement. This type of conversion from expectation to high-probability bounds is a common step in applying learning-theoretic results. More sophisticated techniques can yield tighter, anytime-valid bounds that hold uniformly over time (Rodríguez-Gálvez et al., 2024). A direct application of Markov's inequality suffices.

Lemma 5 (Markov Lift for the KL Term)**.** *Let $X \triangleq \mathrm{KL}(P\|Q) \geq 0$ denote the (dataset-dependent) PAC–Bayes KL term. For any $\eta \in (0,1)$, with probability at least $1-\eta$ (over the draw of $\mathcal{D}$),*

$$X \;\leq\; \frac{\mathbb{E}_{\mathcal{D}}[X]}{\eta}.$$

Proof. Since $X \geq 0$ and $\mathbb{E}_{\mathcal{D}}[X] < \infty$ under the conditions of Theorem 2, Markov's inequality gives

$$\mathbb{P}\Big\{\, X > \frac{\mathbb{E}_{\mathcal{D}}[X]}{\eta} \,\Big\} \;\leq\; \eta.$$

Equivalently, with probability at least $1-\eta$ we have $X \leq \mathbb{E}_{\mathcal{D}}[X]/\eta$, as claimed. □

Corollary 3 (A Capacity-Aware High-Probability Upper Bound)**.** *Fix $\delta, \eta \in (0,1)$. With probability at least $1-\delta-\eta$ (over the draw of $\mathcal{D}$), the PAC–Bayes bound of Thm. 2 implies*

$$\mathbb{E}_{\theta\sim P}\big[R_{\mathrm{obs}}(\theta)\big] \;\leq\; \mathbb{E}_{\theta\sim P}\big[\widehat{R}^{\mathrm{obs}}_m(\theta)\big] \;+\; \sqrt{\frac{\mathbb{E}_{\mathcal{D}}[\mathrm{KL}(P\|Q)]/\eta + \log(1/\delta)}{2m}}.$$

Combining with Cor. 1 and applying a union bound yields, with the same probability,

$$\mathbb{E}_{\theta\sim P}\big[R_{\mathrm{obs}}(\theta)\big] \;\leq\; \mathbb{E}_{\theta\sim P}\big[\widehat{R}^{\mathrm{obs}}_m(\theta)\big] \;+\sqrt{\frac{m\,\bar{C}_{\mathrm{tot}|S} + m\,I(U;S) + \rho + \mathrm{KL}(p(\theta)\|Q)}{2m\,\eta} \;+\; \frac{\log(1/\delta)}{2m}}\,.$$

All constants are explicit; the price of eliminating the dataset randomness in $\mathrm{KL}(P\|Q)$ is the slack parameter η.

K Loss Truncation and Normalization

For unbounded losses such as MSE, define $\ell(u,a) = \min\{\|u-a\|^2/\tau^2, 1\}$ for a scale $\tau > 0$ (report τ when plotting). All PAC–Bayes statements and Thm. 1 require only $\ell \in [0,1]$; truncation ensures this and keeps statements coordinate-free.

L Loss–Observable Link and Risk Transfer

Assumption 4 (Loss–Observable Link)**.** *There exist constants $\alpha \geq 0$ and $\beta \geq 0$ such that for all measurable actions $a \in \mathcal{A}$ and all (y,s) in the support of (Y,S),*

$$\mathbb{E}\big[\ell(U,a)\,\big|\,Y=y,\;S=s\big] \;\leq\; \alpha\,\tilde{\ell}(y,s,a) \;+\; \beta\,. \tag{26}$$

Lemma 6 (Risk Transfer). *Under Assumption 4, for any (possibly randomized) decoder π_θ,*

$$\mathbb{E}\big[\ell\big(U, \pi_\theta(Y,S)\big)\big] \;\le\; \alpha\, \mathbb{E}\big[\tilde{\ell}\big(Y, S, \pi_\theta(Y,S)\big)\big] \;+\; \beta\,. \tag{27}$$

In particular, when $\mathcal{D}$ is drawn from the same codebook-induced mixture distribution used in Theorem 1, taking $\theta \sim P(\cdot \mid \mathcal{D})$ and expectation over both θ and the data gives

$$\mathbb{E}_{\theta\sim P}\big[R_{\mathrm{mix}}(\pi_\theta)\big] \;\le\; \alpha\, \mathbb{E}_{\theta\sim P}\big[R_{\mathrm{obs}}(\theta)\big] \;+\; \beta\,.$$

Proof. By the tower property and equation 26,

$$\mathbb{E}\big[\ell\big(U, \pi_\theta(Y,S)\big)\big] = \mathbb{E}\Big[\,\mathbb{E}\big[\ell(U, \pi_\theta(Y,S)) \mid Y, S\big]\,\Big] \le \alpha\, \mathbb{E}\big[\tilde{\ell}(Y, S, \pi_\theta(Y,S))\big] + \beta.$$

Averaging over $\theta \sim P$ yields the stated forms. □

M Posterior Bayes Loss Identity

Lemma 7 (Posterior Bayes–Loss Identity). *Fix any bounded loss $\ell \in [0,1]$ and define $\tilde{\ell}^\star(y,s,a) \triangleq \mathbb{E}[\ell(U,a) \mid Y=y,\ S=s]$. Then for any (possibly randomized) decoder π_θ and any data distribution over (U,S,Y),*

$$\mathbb{E}\big[\tilde{\ell}^\star(Y,S,\pi_\theta(Y,S))\big] \;=\; \mathbb{E}\big[\ell(U,\pi_\theta(Y,S))\big].$$

Proof. By the tower property of conditional expectation, $\mathbb{E}[\tilde{\ell}^\star(Y,S,\pi_\theta(Y,S))] = \mathbb{E}\{\mathbb{E}[\ell(U,\pi_\theta(Y,S)) \mid Y,S]\} = \mathbb{E}[\ell(U,\pi_\theta(Y,S))]$. □

N Residual Control via Posterior Compression

Let $\theta \sim P(\cdot \mid \mathcal{D})$ be the (possibly randomized) learner parameter. Let W be an auxiliary random seed, independent of (U^m, S^m, Y^m). Consider a data-independent randomized quantizer T that maps θ to $\tilde{\theta} = T(\theta, W)$ taking at most K distinct values. Let P_c and Q_c be the pushforwards of P and Q through T. Then:

Lemma 8 (Residual control by compression). $I(\mathcal{D}; \tilde{\theta} \mid U^m) \;\le\; H(\tilde{\theta}) \;\le\; \log K$.

Proof. $I(\mathcal{D}; \tilde{\theta} \mid U^m) \le H(\tilde{\theta})$, and $H(\tilde{\theta}) \le \log K$ since $\tilde{\theta}$ takes at most K values. □

Lemma 9 (KL does not increase under post-processing). $\mathrm{KL}(P_c \| Q_c) \;\le\; \mathrm{KL}(P\|Q)$.

Using P_c, Q_c in Theorem 2 and Lemma 3 yields the capacity-aware bound of Corollary 1 with $\rho \le \log K$. This approach is conceptually related to other works that leverage model compression or selection of small representative subsets to derive non-vacuous generalization bounds for overparameterized models (Leblanc et al., 2025; Lotfi et al., 2024).

O Context Coarsening

Let $S' = T(S)$ for a measurable (data-release) channel T such that $U \to S \to S'$ forms a Markov chain (that is, S' is generated from S without direct access to U). Then by data processing $I(U;S') \le I(U;S)$. All definitions and bounds in the paper hold verbatim with S' in place of S, with capacities recomputed as $\bar{C}_{\mathrm{tot}|S'}$ and dataset information term $m\, I(U;S')$ replacing $m\, I(U;S)$. Thus, for any preprocessor T, Corollary 1 becomes

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{KL}(P\|Q)\big] \;\le\; m\, \bar{C}_{\mathrm{tot}|S'} \;+\; m\, I(U;S') \;+\; \rho \;+\; \mathrm{KL}(p(\theta)\|Q).$$

Choosing T to enforce $I(U;S') \le \kappa$ makes the interference term $m\,\kappa$ explicit.

P SOFT LOSS–INDEX LINK

Assume there exists a measurable $\phi : \mathcal{A} \times \mathcal{S} \to [M]$ and parameters $\varepsilon, \Delta \geq 0, \zeta \in [0, 1)$ such that for all i and all a,

$$\mathbb{P}\big\{\, \mathbb{E}[\ell(U^{(i)}, a) \mid S] \;\geq\; \varepsilon + \Delta \;\big|\; \phi(a, S) \neq i \,\big\} \;\geq\; 1 - \zeta.$$

Then for any decoder π with $\hat{J} = \phi(\pi(Y, S), S)$, writing $E \triangleq \{\hat{J} \neq J\}$ and $G \triangleq \big\{\mathbb{E}[\ell(U^{(J)}, \pi(Y, S)) \mid S] \geq \varepsilon + \Delta\big\}$, we have

$$\begin{aligned}\mathbb{E}\big[\ell(U, \pi(Y, S))\big] \;&=\; \mathbb{E}\Big[\,\mathbb{E}\big[\ell(U^{(J)}, \pi(Y, S)) \mid S\big] \cdot \mathbf{1}_E\Big] \;+\; \mathbb{E}\Big[\,\mathbb{E}\big[\ell(U^{(J)}, \pi(Y, S)) \mid S\big] \cdot \mathbf{1}_{E^c}\Big] \\ &\geq\; (\varepsilon + \Delta)\,\mathbb{P}(E \cap G),\end{aligned}$$

hence

$$\mathbb{E}\big[\ell(U, \pi(Y, S))\big] \;\geq\; (\varepsilon + \Delta)\,\mathbb{P}(E) \;-\; (\varepsilon + \Delta)\,\mathbb{P}(E \cap G^c).$$

By the assumption, $\mathbb{P}(G^c \mid E) \leq \zeta$, so $\mathbb{P}(E \cap G^c) \leq \zeta\,\mathbb{P}(E) \leq \zeta$ and consequently

$$\mathbb{E}\big[\ell(U, \pi(Y, S))\big] \;\geq\; (\varepsilon + \Delta)\,\mathbb{P}\{\hat{J} \neq J\} \;-\; \zeta.$$

A slightly tighter but equivalent multiplicative form also holds:

$$\mathbb{E}\big[\ell(U, \pi(Y, S))\big] \;\geq\; (\varepsilon + \Delta)\,(1 - \zeta)\,\mathbb{P}\{\hat{J} \neq J\}.$$

Consequently, Theorem 1 holds with an additive $-\zeta$ (or multiplicative $(1 - zeta)$) slack in the lower bound.

Q USE OF LARGE LANGUAGE MODELS

The author utilized Large Language Models as assistive tools in preparing this manuscript. Their applications included literature discovery, language refinement, and the formal review of mathematical derivations. The author directed the entire process and takes full responsibility for the final content and the accuracy of all theoretical claims.