LEARNING RULES, MEMORIZING FACTS: AN INFORMATION-THEORETIC VIEW OF HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the relationship between generalization and hallucination requires clarifying the role of memorization. We study this relationship from an information-theoretic perspective through a Selector-Density Conditional Mutual Information (CMI) framework. In the sparse random-fact limit, a uniformly bounded information budget per fact forces a nonvanishing false-positive rate at recall bounded away from zero, even when all facts are observed during training. In a rules-and-facts setting, we decompose model–data mutual information into rule learning, atomic-fact memory, and excess memorization, revealing competition between rule learning and factual memory even without excess memorization. A joint budget bound quantifies the resulting constraint on atomic hallucination under perfect atomic recall. Moreover, an additive Gaussian model shows that maintaining high recall for atomic and composite facts incurs false positives even under optimal data mixing. Finally, controlled relational experiments hint how training exposure, rule complexity, and model capacity shape these trade-offs.

1 INTRODUCTION

Large language models (LLMs) exhibit broad capabilities across language, coding, and reasoning tasks, along with impressive generalization abilities such as zero-shot (Sanh et al., 2022), few-shot (Brown et al., 2020), and in-context learning (Garg et al., 2022). Yet this ability to generalize does not guarantee factuality: a high-confidence claim about an unseen case may be either a valid inference or a hallucination, namely a plausible but false statement. Such hallucinations can propagate false premises into downstream reasoning and decisions. Recent theoretical work connects hallucination to Is-It-Valid (IIV) classification (Kalai et al., 2025; 2026) or membership testing (Guo & Li, 2026). These perspectives raise a fundamental question about the information retained during training: *what must a model learn or memorize to distinguish facts from plausible non-facts?*

Memorization plays a central, but nuanced, role in this question. Generalization theory often controls the gap between training and population risks by limiting dependence on the particular training sample. At the same time, memorization can be necessary for accurate prediction, especially for rarely observed subpopulations and facts with little reusable structure (Feldman, 2020; Brown et al., 2021). Learning a general rule cannot replace retaining the arbitrary facts to which that rule applies. This motivates distinguishing atomic-fact memory, which preserves factual information not determined by reusable rules, from excess memorization, which records which examples happened to be observed after the underlying factual world is fixed. *How do these distinct forms of retained information jointly constrain generalization and factuality under a limited information budget?*

We address this question through a Selector-Density Conditional Mutual Information (CMI) framework. CMI measures how much a learned model reveals about the training-set selection when conditioned on a supersample of candidate examples (Steinke & Zakynthinou, 2020). Using a uniform selection of n examples from N candidates, we connect selector information to the distinguishability of evaluation outputs on selected and unselected examples: the information an individual output reveals about selection equals a weighted Jensen-Shannon divergence between its two conditional laws. The selector and evaluator determine the interpretation. When selection denotes training membership and the evaluator is a loss, small selector information controls the expected generalization

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

gap. When selection denotes truth and the evaluator is a factuality score, accurate discrimination requires sufficient selector information to separate facts from non-facts.

We first make the latter role explicit in a sparse random-fact setting, following the membership-testing formulation of Guo & Li (2026). A uniformly selected subset of candidate claims constitutes the entire factual world and is fully observed during training, while all remaining claims are false. Since truth has no inferable structure, factual discrimination depends on retaining information about this subset. We derive lower bounds on the false-positive hallucination rate at prescribed fact recall. In the sparse-fact limit, a uniformly bounded information budget per fact prevents false positives from vanishing whenever recall remains bounded away from zero. Thus, post-hoc thresholding alone cannot overcome insufficient factual memory. The contrast with generalization concerns the information being measured, that is, indistinguishability of training and held-out losses controls a generalization gap, whereas indistinguishability of factual and non-factual scores obstructs factual discrimination. In this sense, hallucination can be viewed as an "unwanted generalization" of factual confidence from observed facts to unseen non-facts.

To incorporate knowledge that supports inference beyond observed examples, we introduce a rules-and-facts setting in which reusable rules and atomic facts jointly determine the factual world. Here, truth and training membership are distinct: an unobserved claim may still be true. We decompose model-data mutual information into rule learning, atomic-fact memory conditional on the rules, and excess memorization of the training selection. This separates the information supporting factual knowledge from residual dependence on the observed sample. Under a shared information budget, information committed to rule recovery and training selection limits what remains for atomic-fact memory. Consequently, even accurate rule learning and a small generalization gap do not by themselves ensure factual discrimination; namely, rule-independent facts still require memory. Our bounds quantify the constraint on atomic false positives under perfect atomic recall.

Our contributions are summarized as follows:

- We connect the CMI generalization framework with information requirements for factual discrimination. In the sparse random-fact setting, we show that a bounded information budget per fact forces a nonvanishing hallucination rate at recall bounded away from zero, even when every true fact is observed during training (Theorem 3.3). Post-hoc thresholding therefore cannot compensate for insufficient factual memory. To the best of our knowledge, this is the first application of the CMI framework to hallucination analysis.
- In a rules-and-facts setting, we separate model-data mutual information into rule learning, atomic-fact memory and excess memorization (Lemma 4.1). In the sparse atomic-fact regime, our joint budget bound quantifies how information committed to rule recovery and training selection constrains atomic hallucination suppression under perfect recall (Theorem 4.1). Thus, accurate rule learning and a small generalization gap do not by themselves ensure factuality.
- In a shared-location Gaussian model, we show how the training mixture redistributes recall between atomic and composite facts (Theorem 4.2), while their separation and dispersion impose a hallucination cost even under optimal mixing (Corollary 4.1). Controlled relational experiments show that adding composite facts during training need not improve atomic factual discrimination, and that greater rule complexity can increase its capacity requirements. Larger models ease these trade-offs, providing qualitative evidence that the factual benefits of additional structured data depend on the capacity available to retain atomic facts.

Other Related Literature. Building on CMI (Steinke & Zakynthinou, 2020), functional CMI (Harutyunyan et al., 2021), evaluated CMI (Hellström & Durisi, 2022), and loss-difference CMI (Wang & Mao, 2023b) use successive post-processing to obtain tighter generalization bounds. Leave-one-out (Haghifam et al., 2022; Rammal et al., 2022) and leave-m-out CMI (Lu et al., 2026) further vary the sampling mechanism. These works establish CMI as a well-developed framework for generalization. However, its implications for hallucination remain unexplored. Existing analyses of hallucination instead draw on Good–Turing estimation (Kalai & Vempala, 2024), Kolmogorov complexity (Pan et al., 2025), and statistical learning theory (Kalavasis et al., 2025). Our distinction between factual knowledge and residual sample dependence also relates to excess memorization (Feldman et al., 2025) and unintended memorization (Morris et al., 2026). Moreover, knowledge acquisition under finite capacity has been studied through data pruning (Ye et al., 2026) and data mixing (Gu et al., 2025). Appendix A provides further discussion on related work.

2 PRELIMINARIES

Notations. Throughout this paper, unless otherwise stated, capital letters (e.g., X) denote random variables, while the corresponding lowercase letters (e.g., x) denote their realizations. Expectations are expressed as $\mathbb{E}_X[\cdot]$. All logarithms are natural and all information quantities are in nats. Let the binary entropy be $h(a) \coloneqq -a\log a - (1-a)\log(1-a)$. For probability measures P, Q on a common measurable space, $\mathrm{D_{KL}}\left(P\|Q\right)$ denotes the KL divergence.

Supersample Setting. Let μ be an unknown distribution over the data instance space $\mathcal{Z}$. Let $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_N) \overset{i.i.d.}{\sim} \mu^{\otimes N}$ be a supersample of candidate instances. Each $\widetilde{Z}_j$ may represent a data–label pair in supervised learning or a plausible claim in a generative-model setting. For $n \in [1, N)$, we draw a selector $K \sim \mathrm{Unif}\{A \subseteq [N] : |A| = n\}$ independently of $\widetilde{Z}$ and form the training sample $S \coloneqq \widetilde{Z}_K$. The remaining $m \coloneqq N - n$ instances constitute the holdout sample, and we define the selector density as $\rho \coloneqq \frac{n}{N}$. Note that this construction differs from the original CMI framework (Steinke & Zakynthinou, 2020), which arranges $2n$ instances into n pairs and independently selects one instance from each pair with equal probability. The distinction lies in the selector's support, that is, the original selector is uniform over the 2^n subsets containing exactly one instance per pair, whereas ours is uniform over all $\binom{N}{n}$ subsets of size n. Thus, even when $N = 2n$, the two constructions generally induce different selector laws. Our construction is closer to the leave-one-out or leave-m-out setting (Haghifam et al., 2022; Rammal et al., 2022; Lu et al., 2026).

CMI and eCMI. Let $\mathcal{W}$ be the hypothesis space, and let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ be a learning algorithm that maps the training sample S to a model $W \coloneqq \mathcal{A}(S)$. Given a measurable evaluator $q : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, define the evaluation outputs $T^N \coloneqq (T_1, \ldots, T_N)$ where $T_i \coloneqq q(W, \widetilde{Z}_i)$. We define the *selector-density CMI* and its *evaluated counterpart* as $\mathrm{CMI}_\rho(\mathcal{A}) \coloneqq I(W; K \mid \widetilde{Z})$, and $\mathrm{eCMI}^q_\rho(\mathcal{A}) \coloneqq I(T^N; K \mid \widetilde{Z})$. These quantities measure how much the model and its evaluated behavior, respectively, reveal about the training-set selection given the supersample. Since T^N is a deterministic function of $(W, \widetilde{Z})$, the data-processing inequality gives $\mathrm{eCMI}^q_\rho(\mathcal{A}) \le \mathrm{CMI}_\rho(\mathcal{A})$.

At the individual level, let $U_i \coloneqq \mathbf{1}\{i \in K\}$ denote the membership indicator of $\widetilde{Z}_i$. Uniform subset selection implies $U_i \sim \mathrm{Bern}(\rho)$. The mutual information $I(U_i; T_i)$ quantifies how much the evaluation output T_i reveals about whether $\widetilde{Z}_i$ was selected for training, providing an information-theoretic perspective on membership testing (Steinke & Zakynthinou, 2020).

For a fixed coordinate $i \in [N]$, define the selected-item and unselected-item laws $P \coloneqq \mathrm{Law}(T_i \mid U_i = 1)$ and $Q \coloneqq \mathrm{Law}(T_i \mid U_i = 0)$, where we omit their dependence on i for notational simplicity. The mixture $M_\rho \coloneqq \rho P + (1-\rho)Q$ is the marginal law of T_i. The corresponding weighted Jensen–Shannon (JS) divergence is $J_\rho(P, Q) \coloneqq \rho \mathrm{D_{KL}}\left(P\|M_\rho\right) + (1-\rho)\mathrm{D_{KL}}\left(Q\|M_\rho\right)$. The following lemma connects this divergence to individual selector information.

Lemma 2.1. *Under the preceding notation, for any fixed $i \in [N]$, we have*

$$I(U_i; T_i) = J_\rho(P, Q). \tag{1}$$

Remark 2.1. *Eq. (1) identifies the information that T_i reveals about membership with the prior weighted JS divergence between the two conditional output laws. Equivalently, $I(U_i; T_i) = h(\rho) - H(U_i \mid T_i)$, where h denotes binary entropy. Thus, selector information measures the expected reduction in membership uncertainty after observing T_i. If $P = Q$, then $I(U_i; T_i) = 0$, that is, the output provides no membership information beyond the prior. At the opposite extreme, if P and Q are mutually singular, membership can be recovered perfectly from T_i, and $I(U_i; T_i) = h(\rho)$. The selector density therefore determines the maximum information that an individual output can reveal.*

At the dataset level, CMI and eCMI share the same uncertainty-reduction interpretation. Conditioning on the supersample fixes the candidate instances, so the remaining uncertainty concerns which instances were selected. This membership-testing perspective underlies the analyses that follow.

3 CMI VIEW OF GENERALIZATION AND HALLUCINATION

This section studies generalization and random-fact hallucination through the membership-testing interpretation of selector information.

3.1 Generalization Bounds

We revisit CMI generalization analysis, deriving individual and global bounds that illustrate how the membership information revealed by model losses controls the expected generalization gap.

Let the evaluator q be a bounded loss function $\ell : \mathcal{W} \times \mathcal{Z} \to [0,1]$, so that $T_i = \ell(W, \widetilde{Z}_i)$. Let $K^c \coloneqq [N] \setminus K$, and define the empirical and held-out risks as $\hat{L}_n \coloneqq \frac{1}{n}\sum_{i\in K} \ell(W, \widetilde{Z}_i)$ and $\hat{L}_m \coloneqq \frac{1}{m}\sum_{i\in K^c} \ell(W, \widetilde{Z}_i)$, respectively. Let $L_\mu(w) \coloneqq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$ denote the population risk. Under the i.i.d. supersample construction, the held-out observations are independent of W conditional on K, and hence $\mathbb{E}\left[\hat{L}_m\right] = \mathbb{E}\left[L_\mu(W)\right]$. The expected generalization error can therefore be written as $\mathcal{E}(\mathcal{A}) \coloneqq \mathbb{E}_{\widetilde{Z},K,W}\left[\hat{L}_m - \hat{L}_n\right] = \mathbb{E}\left[L_\mu(W) - \hat{L}_n\right]$. Our first result bounds this gap through the selector information associated with an individual evaluation output.

Theorem 3.1. *Let $J \sim \mathrm{Unif}([N])$ be independent of $(\widetilde{Z}, K, W)$, and define $U \coloneqq \mathbf{1}\{J \in K\}$ and $T \coloneqq T_J$. Then*

$$|\mathcal{E}(\mathcal{A})| \le \sqrt{\frac{I(U; T \mid \widetilde{Z}, J)}{2\rho(1-\rho)}}. \tag{2}$$

The next result instead uses the global evaluated CMI of the entire loss vector, $\mathrm{eCMI}^{\ell}_{\rho}(\mathcal{A})$.

Theorem 3.2. *Under the construction above, we have*

$$|\mathcal{E}(\mathcal{A})| \le \min\left\{ \frac{N}{m}\sqrt{\frac{\mathrm{eCMI}^{\ell}_{\rho}(\mathcal{A})}{2n}}, \frac{N}{n}\sqrt{\frac{\mathrm{eCMI}^{\ell}_{\rho}(\mathcal{A})}{2m}} \right\}. \tag{3}$$

The two theorems quantify selector information at different levels. Theorem 3.1 uses an average of coordinate-wise CMI, whereas Theorem 3.2 uses the information that the full loss vector reveals about the selected subset. Both provide a membership-testing interpretation, namely when the losses reveal little about which instances were selected for training, the expected training and held-out risks cannot differ significantly. Importantly, this controls the generalization gap rather than the population risk itself, and does not require the entire model to carry little information about the training sample.

Our bounds use a fixed-cardinality selector with general density, so the information quantities need not coincide with those of the original CMI framework. Our objective is not to improve upon existing CMI bounds, but to establish the selector-information perspective that also enables our hallucination analysis. We defer a comparison with existing CMI bounds to Appendix (see Remark D.1).

3.2 Hallucination in the Random-Fact Setting

We now discuss generative models and consider a random-fact setting. For this subsection, let $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_N)$ be a fixed universe of N distinct plausible claims, rather than an i.i.d. supersample. Draw $K \sim \mathrm{Unif}\{A \subseteq [N] : |A| = n\}$, and declare precisely the claims indexed by K to be true. The training sample $S = \widetilde{Z}_K$ therefore contains all facts, while $\widetilde{Z}_{K^c}$ contains all nonfacts. Thus, in this closed-world abstraction, K serves both as the training-set selector and as the truth selector, following the random-fact membership-testing model of Guo & Li (2026). Since truth is assigned uniformly at random, the claim identities provide no additional structure from which membership can be inferred, and factual discrimination must instead rely on information retained about K. Recall that $J \sim \mathrm{Unif}([N])$, $U \coloneqq \mathbf{1}\{J \in K\}$ and $T \coloneqq q(W, \widetilde{Z}_J)$. Here, $U \sim \mathrm{Bern}(\rho)$ indicates whether the queried claim is true, and $T \in [0,1]$ is a factuality score, with larger values indicating greater confidence in the claim's truth. Accordingly, $P \coloneqq \mathrm{Law}(T \mid U = 1)$ and $Q \coloneqq \mathrm{Law}(T \mid U = 0)$ are now score laws on facts and non-facts, respectively, rather than loss distributions on training and held-out samples. Since the claim universe is fixed, conditioning on $\widetilde{Z}$ is implicit throughout this subsection (i.e. $\widetilde{Z} = \tilde{z}$).

We focus first on the sparse regime $\rho = n/N \ll 1$, in which true facts constitute a small fraction of the candidate universe (Guo & Li, 2026; Kalai et al., 2025; 2026). For intuition, a birthday query

with 365 candidate dates and exactly one correct answer has truth density $1/365$ under uniform candidate querying. This illustrates a small truth base rate. The limit $\rho \to 0$ idealizes an increasingly large set of plausible alternatives relative to the number of facts.

The following lemma identifies the normalized selector information in this regime.

Lemma 3.1. *Under the setting of Lemma 2.1, as* $\rho = \frac{n}{N} \to 0$, $\frac{I(U;T)}{\rho} = \frac{J_\rho(P,Q)}{\rho} \longrightarrow \mathrm{D}_{\mathrm{KL}}(P\|Q)$.

Lemma 3.1, similar to Guo & Li (2026, Lemma 3.5), shows that directional KL is the limiting selector information per rare true claim, with Q serving as the nonfact background. This hints at a *capacity-per-unit-fact* interpretation, namely for the stationary memoryless extension of the binary channel $U \mapsto T$, assign cost zero to $U = 0$ and cost one to $U = 1$. Its capacity per unit cost is then $\mathrm{D}_{\mathrm{KL}}(P\|Q)$ by the zero-cost-symbol formula (Polyanskiy & Wu, 2023).

Under their membership-testing assumptions, Guo & Li (2026) establish that, for prescribed error constraints, the minimum asymptotic memory rate per fact is the minimum of $\mathrm{D}_{\mathrm{KL}}(P\|Q)$ over admissible fact and non-fact score laws, along sequences with $n \to \infty$ and $n/N \to 0$. This result connects the score laws' distinguishability to the memorization required for random facts.

Generative models generally assign nonzero probability to unseen outputs, including non-facts, thereby incurring prediction error; for example, softmax assigns positive probability to every candidate token. Hallucination, however, is the special case in which a non-fact receives sufficient confidence to be treated as factual. In the random-fact setting, facts with $U = 1$ form the positive class, and non-facts with $U = 0$ form the negative class. We consider a threshold-based notion of factuality: a candidate claim is accepted as factual when its score reaches a prescribed threshold. For $A_\tau = \mathbf{1}\{T \geq \tau\}$, define the fact recall and false-positive rate as $\alpha_\tau := \mathbb{P}(A_\tau = 1 \mid U = 1)$ and $\beta_\tau := \mathbb{P}(A_\tau = 1 \mid U = 0)$, respectively. Thus, accepting a nonfact is a false positive, and we refer to β_τ as the *false-positive hallucination rate*. A natural question is whether a suitable threshold can eliminate false-positive hallucinations while retaining factual claims. We characterize their relationship in the sparse-fact regime below.

Theorem 3.3. *In the random-fact setting, suppose that* $\frac{1}{n} I(T^N; K) \leq R$. *Then in the sparse limit* $\rho \to 0$ *and* $n \to \infty$, *for every* τ *s.t.* $\alpha_\tau > 0$, *we have* $\beta_\tau \geq \exp\left(-\frac{R + h(\alpha_\tau)}{\alpha_\tau}\right)$. *In particular, perfect recall* $\alpha_\tau = 1$ *implies* $\beta_\tau \geq e^{-R}$.

Theorem 3.3 identifies an information-limited recall–hallucination trade-off, illustrated in Figure 7 in Appendix D. Under a uniformly bounded score-level information budget per fact, recall bounded away from zero prevents the hallucination rate from vanishing in the sparse limit. Changing the threshold can move the operating point, but cannot bypass this constraint for the same score array. In addition, in the random-fact setting, facts contain no reusable structure, so factual discrimination reduces to memorizing selector K. Since $I(T^N; K) \leq I(W; K) \leq H(W)$, finite storage therefore provides one sufficient source of the information bottleneck. As a by-product of our analysis, we also study the relationship between calibration and hallucination in Appendix B.

Remark 3.1 (Revisiting Lemma 2.1: two roles of selector information)**.** *Lemma 2.1 gives both (generalization and hallucination) analyses a membership-testing interpretation through* $I(U;T) = J_\rho(P,Q)$. *Small selector information in* losses *controls the generalization gap, whereas high fact recall with few false positives requires sufficient selector information in* factuality scores. *Indeed,* $P = Q$ *implies* $\alpha_\tau = \beta_\tau$ *for every threshold. In the random-fact setting, hallucination can be viewed as* unwanted generalization of factual confidence *from observed facts to unseen nonfacts.*

The random-fact abstraction isolates the role of memorization. Real knowledge, however, also contains reusable structure that supports inference beyond observed examples. In the next section, we introduce a rules-and-facts setting to study how rule learning and atomic-fact memory jointly constrain factual discrimination under a shared information budget.

4 Rules-and-Facts Setting

We now extend the random-fact setting to a rules-and-facts model in which atomic facts and reusable rules jointly generate the factual world. This construction distinguishes rule learning from atomic-fact memorization at the level of the generative model and its information-theoretic analysis.

Formally, let $\mathscr{B} = (b_1, \ldots, b_{N_1})$ be a fixed pool of N_1 different candidate atomic claims. An atomic fact selector $K_1 \sim \mathrm{Unif}\{A \subseteq [N_1] : |A| = n_1\}$ determines the true atomic facts $B \coloneqq \mathscr{B}_{K_1}$, with $\rho_1 \coloneqq \frac{n_1}{N_1}$. Let $\Theta \sim P_\Theta$ represent the reusable rule structure, independently of K_1. Given (B, Θ), a deterministic mapping Γ generates the full factual supersample $\widetilde{Z} = \Gamma(B, \Theta) = (\widetilde{Z}_1, \ldots, \widetilde{Z}_N)$, which contains all true atomic and composite facts within the modeled universe. Unlike the earlier i.i.d. setting, the entries $\widetilde{Z}_i$ need not be independent, since they are generated from the shared latent variables (B, Θ). Independently, let $K \sim \mathrm{Unif}\{A \subseteq [N] : |A| = n\}$ select the training set $S \coloneqq \widetilde{Z}_K$, with sampling density $\rho \coloneqq \frac{n}{N}$. The learner then outputs $W = \mathcal{A}(S)$. Thus, (K_1, Θ) determines which claims are facts, whereas K determines which facts are observed during training.

The mutual information (MI) $I(W; S)$ is a central quantity in information-theoretic generalization analysis (Xu & Raginsky, 2017). This quantity and its conditional variants have also been used to quantify information retention and memorization in learning (Bassily et al., 2018; Brown et al., 2021). More recently, related MI measures have been used to formalize the effective capacity of LLMs (Gu et al., 2025). In our setting, $I(W; S)$ quantifies how much information the learned model retains, on average, about the realized training set and serves as a data-dependent information budget. The following lemma decomposes this budget into three components.

Lemma 4.1. *In the rules-and-facts setting, we have*

$$I(W; S) = \underbrace{I(W; \Theta)}_{\textit{Learning}} + \underbrace{I(W; B \mid \Theta)}_{\textit{Atomic fact memory}} + \underbrace{I(W; K \mid \widetilde{Z})}_{\textit{Excess memorization}} . \tag{4}$$

Remark 4.1. *Here, $\widetilde{Z}$ contains all atomic and composite facts, and $S = \widetilde{Z}_K$ contains only the selected observations. However, S need not contain all atomic facts, so complete atomic-fact recovery is not guaranteed even with sufficient model capacity. To guarantee direct access to every atomic fact, one can instead let $\widetilde{Z} = \Gamma(B, \Theta)$ contain only composite facts and define $S = (B, \widetilde{Z}_K)$, where K only selects composite observations for rule learning. The exact decomposition then has the third term $I(W; K \mid B, \Theta) \geq I(W; K \mid \widetilde{Z})$, giving $I(W; S) \geq I(W; \Theta) + I(W; B \mid \Theta) + I(W; K \mid \widetilde{Z})$. Note that the joint information-budget bound (i.e. Theorem 4.1) below also applies to this variant under the corresponding information budget, with the generalization gap evaluated over the selected and held-out composite observations.*

The decomposition separates three roles of model information. The first term, $I(W; \Theta)$, measures the reusable rule information retained by the model. Learning Θ from $\widetilde{Z}_K$ enables inference on unseen $\widetilde{Z}_i$ and compresses multiple related facts into shared structure. The second term, $I(W; B \mid \Theta)$, measures atomic fact information not explained by the rules. In other words, to capture all facts, the model must store these atomic exceptions and compose them with the rules. The third term, $I(W; K \mid \widetilde{Z})$, is selector-density CMI: conditioned on the full factual supersample, it is a record not of truth, but of which facts the model happened to observe. As above, it controls the generalization gap, and larger values may indicate sample-specific overfitting, i.e., undesirable memorization.

The first two terms jointly form the world information $I(W; B, \Theta)$: rules compactly encode transferable structure, and applying them obtains new conclusions rather than new independent information about the world. Learning such structure, however, competes with memorizing arbitrary facts for finite capacity, echoing the rule–memory competition discussed in Guo & Li (2026, Appendix A.3). Our further decomposition separates factual content from training selection, i.e. $I(W; B, K \mid \Theta) = I(W; B \mid \Theta) + I(W; K \mid \widetilde{Z})$, indicating that selection-specific memory can be discarded to promote effective learning and generalization. This separation is closely related to the "average excess memorization" studied by Feldman et al. (2025), who measure training data information conditional on a latent parameter drawn from a meta-distribution and specifying the data-generating distribution. Another related distinction underlies the notion of "unintended memorization" in Morris et al. (2026), who separate dataset-specific information from information about the data-generating process. In our setting, the corresponding conditioning object is the full latent world (B, Θ). Thus, atomic-fact information belongs to the world-information component, whereas selector information captures residual sample dependence after that world is fixed.

Theorem 4.1. *Suppose that $I(W; S) \leq C$ and that $\Theta \sim \mathrm{Unif}([N_\Theta])$, where $N_\Theta \geq 2$. Assume that there exists a decoder $\widehat{\Theta}(W)$ satisfying $\mathbb{P}(\widehat{\Theta} \neq \Theta) \leq \varepsilon_\Theta$. Define*

$$C_{\mathrm{rule}} \coloneqq \log N_\Theta - h(\varepsilon_\Theta) - \varepsilon_\Theta \log(N_\Theta - 1), \quad \textit{and } C_{\mathrm{gen}} \coloneqq \frac{2nm \max\{n, m\}}{N^2} |\mathcal{E}(\mathcal{A})|^2.$$

Then $I(W;\Theta) \geq C_{\text{rule}}$ *and* $I(W; K \mid \widetilde{Z}) \geq C_{\text{gen}}$.

Moreover, for a threshold decision on atomic claims, let α_τ *denote fact recall and* β_τ *the false-positive hallucination rate. In the sparse atomic-fact limit* $\rho_1 = \frac{n_1}{N_1} \to 0$*, perfect atomic recall,* $\alpha_\tau = 1$*, implies*

$$\beta_\tau \geq \exp\left(-\frac{C - C_{\text{rule}} - C_{\text{gen}}}{n_1}\right).$$

Theorem 4.1 places rule recovery, sample dependence, and hallucination suppression under the total information budget C, unlike Theorem 3.3, which assumes a score-level information budget R per fact. At perfect atomic recall, its sparse-limit constraint can be written as $C_{\text{rule}} + C_{\text{gen}} + n_1 \log \frac{1}{\beta_\tau} \leq C$. Suppressing false positives to rate β_τ therefore requires at least $\log(1/\beta_\tau)$ nats per atomic fact on average. The remaining terms C_{rule} and C_{gen} quantify the minimum information committed to rule recovery and certified by the generalization gap, respectively. At fixed C, these commitments reduce the budget available for atomic factual discrimination. This perspective also motivates retrieval-augmented generation (RAG) (Lewis et al., 2020). As also discussed by Guo & Li (2026), external non-parametric memory can alleviate the need to encode all factual knowledge in model parameters.

4.1 Data Mixing and the Cost of Factual Recall

Theorem 4.1 identifies a competition between reusable rule information and atomic-fact memory under a shared information budget. It leaves open an operational question: *how does the training mixture affect the factuality behavior of a particular learner?* We study an additive Gaussian setting with a restricted generative model that represents both fact types through a single location parameter, allowing explicit characterization of fact recall and false-positive rates.

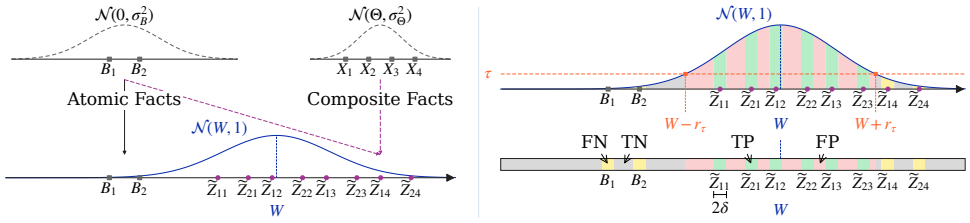


Figure 1: Left: atomic and composite fact generation and fitting with $\mathcal{N}(W,1)$. Right: factuality tolerance and the acceptance threshold jointly define classification outcomes.

Atomic facts and rule-derived facts. Let atomic facts be $B = (B_1, \ldots, B_{n_1})$ with $B_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_B^2)$, where $\sigma_B > 0$. Draw $\Theta \sim P_\Theta$ on $[0,\infty)$ and, conditional on Θ, generate additive rule instances $X = (X_1, \ldots, X_{n_2})$ with $X_j \mid \Theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\Theta, \sigma_\Theta^2)$, where $\sigma_\Theta^2 \in (0,\infty)$ is a prescribed function of Θ. We assume $B \perp\!\!\!\perp \Theta$. Each rule instance is shared across atomic facts, producing $\widetilde{Z} := \Gamma(B, X) \in \mathbb{R}^{n_1 \times n_2}$ with $\widetilde{Z}_{ij} := B_i + X_j$. Thus the composite pool contains $N = n_1 n_2$ facts. Independently of (Θ, B, X), draw K uniformly over the size-n subsets of $[n_1] \times [n_2]$. As in Remark 4.1, training uses all atomic facts and the selected composite facts, namely $S := (B, \widetilde{Z}_K)$ with $|S| = n_1 + n$, where $K_{ij} := \mathbf{1}\{(i,j) \in K\}$. Define the composite proportion in the training sample by $a_N := \frac{n}{n_1+n}$.

Shared generative representation. Let ϕ and Φ denote the standard Gaussian density and distribution functions. We fit the unit-variance Gaussian density $p_w(z) = \phi(z - w)$ by minimizing the average negative log-likelihood. In fact, this is a simple analogue of LLM pretraining, which learns a distribution over sequences. Up to an additive constant, the per-example loss is $\ell(w, z) = \frac{1}{2}(z-w)^2$. The empirical objective and its minimizer are therefore

$$\widehat{L}_S(w) := \frac{\sum_i \ell(w, B_i) + \sum_{i,j} K_{ij} \ell(w, B_i + X_j)}{n_1 + n}, \quad \widehat{W}_N := \frac{\sum_i B_i + \sum_{i,j} K_{ij}(B_i + X_j)}{n_1 + n}.$$

Let W_N denote the learned parameter, allowing for an approximate rather than exact minimizer.

Recall and hallucination rate. Let $q(W_N, z) = p_{W_N}(z)$, namely using the likelihood $p_{W_N}(z)$ as a factuality score. For $r \geq 0$, set $\tau(r) = \phi(r)$, so that $A_{\tau(r)}(z) = \mathbf{1}\{p_{W_N}(z) \geq \tau(r)\} =$

$\mathbf{1}\{|z - W_N| \le r\}$. Positive queries are sampled uniformly from the atomic or composite fact pool. Their conditional acceptance probabilities are

$$\alpha_N^{\mathrm{ato}}(r) \coloneqq \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{1}\{|B_i - W_N| \le r\}, \quad \alpha_N^{\mathrm{comp}}(r) \coloneqq \frac{1}{N}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \mathbf{1}\{|\widetilde{Z}_{ij} - W_N| \le r\}.$$

The composite recall is evaluated on the full composite pool, including facts not selected for training.

To define nonfact queries in this continuous setting, introduce a factuality tolerance $\delta_N > 0$ and let $\mathcal{T}_N \coloneqq \bigcup_{i=1}^{n_1}[B_i - \delta_N, B_i + \delta_N] \ \cup \ \bigcup_{i=1}^{n_1}\bigcup_{j=1}^{n_2}[\widetilde{Z}_{ij} - \delta_N, \widetilde{Z}_{ij} + \delta_N]$. Conditional on the realized (B, X, W_N), draw a fresh candidate $Z \sim \mathcal{N}(W_N, 1)$ and define $\beta_N(r) \coloneqq \mathbb{P}\big(|Z - W_N| \le r \,\big|\, Z \notin \mathcal{T}_N\big)$. Thus $\beta_N(r)$ is the false-positive (hallucination) rate, i.e., the probability of accepting a model-proposed candidate conditioned on it being a nonfact.

Theorem 4.2. *Fix $\theta > 0$ and condition on $\Theta = \theta$. Let $\sigma_{\mathrm{c}} \coloneqq \sqrt{\sigma_B^2 + \sigma_\theta^2}$. Suppose that $n_1, n_2, n \to \infty$, $a_N \to a \in [0,1]$, $N\delta_N \to 0$, and that $W_N - \widehat{W}_N \xrightarrow{\mathbb{P}} 0$. Then $W_N \xrightarrow{\mathbb{P}} a\theta$. Moreover, for every fixed $r \ge 0$,*

$$\alpha_N^{\mathrm{ato}}(r) \xrightarrow{\mathbb{P}} \alpha^{\mathrm{ato}}(r; a) \coloneqq \Phi\left(\frac{r - a\theta}{\sigma_B}\right) + \Phi\left(\frac{r + a\theta}{\sigma_B}\right) - 1, \tag{5}$$

$$\alpha_N^{\mathrm{comp}}(r) \xrightarrow{\mathbb{P}} \alpha^{\mathrm{comp}}(r; a) \coloneqq \Phi\left(\frac{r - (1-a)\theta}{\sigma_{\mathrm{c}}}\right) + \Phi\left(\frac{r + (1-a)\theta}{\sigma_{\mathrm{c}}}\right) - 1, \tag{6}$$

$$\beta_N(r) \xrightarrow{\mathbb{P}} \beta(r) \coloneqq 2\Phi(r) - 1. \tag{7}$$

Notably, at any fixed radius $r > 0$, increasing the composite proportion a moves the learned center $a\theta$ from the atomic population toward the composite population. Atomic recall therefore decreases, whereas composite recall increases. Importantly, the limiting false-positive rate at a fixed radius is independent of a, hence the mixture affects false positives only through the radius required to maintain the desired recall. This motivates choosing the mixture jointly with the acceptance threshold.

Corollary 4.1. *Under the conditions of Theorem 4.2, fix a target recall $\alpha_0 \in (1/2, 1)$. For $a \in [0,1]$, define $r_g(a) \coloneqq \inf\{r \ge 0 : \alpha^g(r; a) \ge \alpha_0\}$ for $g \in \{\mathrm{ato}, \mathrm{comp}\}$, the smallest radius at which the limiting recall of population g reaches α_0. Let $r_{\alpha_0}(a) \coloneqq \max\{r_{\mathrm{ato}}(a), r_{\mathrm{comp}}(a)\}$, and let $\beta_{\alpha_0}(a) \coloneqq \beta(r_{\alpha_0}(a))$ be its limiting false-positive rate. Then*

$$\min_{a \in [0,1]} \beta_{\alpha_0}(a) \ge 2\Phi\left(\frac{\theta}{2} + \frac{\sigma_B + \sigma_{\mathrm{c}}}{2}\Phi^{-1}(\alpha_0)\right) - 1. \tag{8}$$

Remark 4.2. *If $r_{\mathrm{ato}}(1) > r_{\mathrm{comp}}(1)$, there is a unique $a_* \in (0,1)$ satisfying $r_{\mathrm{ato}}(a_*) = r_{\mathrm{comp}}(a_*)$. In this case, $\beta_{\alpha_0}(a)$ is strictly decreasing on $[0, a_*]$ and strictly increasing on $[a_*, 1]$. Otherwise, its unique minimizer is $a_* = 1$. Moreover, $\sigma_{\mathrm{c}} > \sigma_B$ and $\alpha_0 > 1/2$ imply $r_{\mathrm{comp}}(1/2) > r_{\mathrm{ato}}(1/2)$, and hence $a_* > 1/2$. Thus, equal recall targets do not imply equal training proportions, rather, the optimal mixture favors composite facts to compensate for their greater dispersion.*

Corollary 4.1 shows that an interior optimal mixture balances the acceptance radii required by the two fact populations, rather than their sample counts. An overly atomic-heavy mixture requires a wide acceptance region to retain composite facts, and an overly composite-heavy mixture does the opposite. Mixture selection can reposition the shared center, but cannot remove the separation or dispersion of the two populations. Consequently, even the optimal mixture incurs a false-positive cost when both recall requirements must be met. For fixed variances, greater separation θ raises the lower bound in Eq. (8); increasing either standard deviation or the recall target also raises this bound.

5 Experiments

In the preceding analysis, we lower-bound atomic hallucination under a shared information budget, revealing competition between rule learning and atomic-fact memory (Theorem 4.1). The Gaussian analysis further shows that the cost of recalling both fact types depends on the training mixture and the underlying distributional structure (Corollary 4.1). Motivated by these results, we design

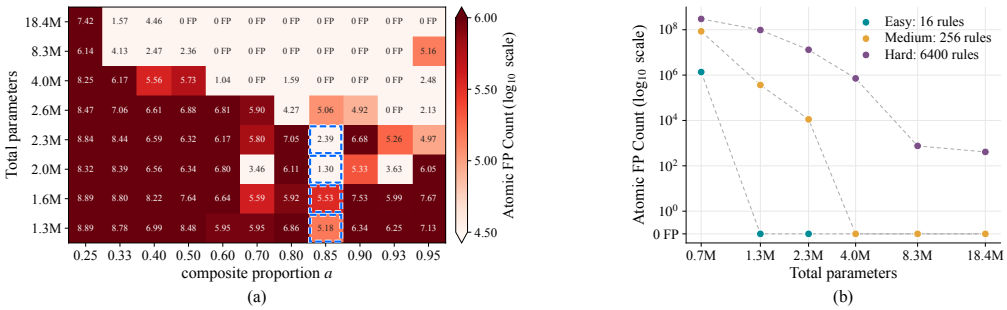


Figure 2: AtomicFP at perfect fact recall across model sizes: (a) varying composite training proportion a; (b) varying rule count at fixed $a = 0.5$.

a controlled relational task in the rules-and-facts setting, where composite-fact exposure adjusts the training mixture and rule count varies the structural learning burden. Across model sizes, we examine how these two factors interact with capacity to shape atomic hallucination.

Task setup. We construct a controlled rules-and-facts knowledge graph, inspired by Wang et al. (2024); Yao et al. (2026), to test whether a model can retain atomic facts and reuse composition rules. Entities are nodes, and each fact $z = (e_h, r_i, e_t)$ is a directed edge. Under the rule $r_a \circ r_b = r_c$, the atomic facts (e_1, r_a, e_2) and (e_2, r_b, e_3) imply the composite fact (e_1, r_c, e_3). In a real-world analogue, Alice's father is Bob and Bob's mother is Carol, so Carol is Alice's grandmother.

Training and evaluation. We train randomly initialized Transformers autoregressively with varying parameter counts on symbolic three-token sequences (e_h, r_i, e_t), mirroring language-model pretraining. The training set includes all n_1 atomic facts in B and n composite facts from $\widetilde{Z}_{\text{comp}}$. We fix n_1 and vary n to adjust the composite proportion $a = n/(n_1 + n)$ over $[0.25, 0.95]$. To evaluate hallucination for a trained model w, we use the predicted tail probability $q(w, z) := p_w(e_t \mid e_h, r_i)$ as the factuality score. We construct the negative set $B^- := \{(e_h, r_i, e) : (e_h, r_i, e_t) \in B,\ e \neq e_t\}$ and report $\text{AtomicFP} := \sum_{z \in B^-} \mathbf{1}\{q(w, z) \geq \tau_*\}$, where $\tau_* := \min_{z \in B \cup \widetilde{Z}_{\text{comp}}} q(w, z)$. With $|B|$ and entity vocabulary fixed, $|B^-|$ is constant, so AtomicFP is proportional to the empirical atomic hallucination rate at perfect fact recall. We also fix n_1 and a and vary the number of rules across 16, 256, and 6400 to examine how the rule-learning burden changes model capacity requirements.

Results and analysis. Figure 2(a) shows that small and medium-sized models generally require intermediate mixtures to achieve low AtomicFP, while larger models are less sensitive to this choice. Thus, additional composite training facts need not improve atomic factual discrimination. Meanwhile, Figure 2(b) shows that more composite rules increase hallucination at a fixed model size. Although the atomic and composite training counts remain unchanged, richer rule structure demands more bits to encode the factual world, thereby imposing a tighter capacity-per-unit-fact constraint. Overall, greater rule-related exposure or higher rule complexity can incur a cost in atomic factual discrimination. Scaling model capacity helps ease this trade-off, consistent with our theory in the rules-and-facts setting. We provide more experimental details and results in Appendix C.

6 Conclusion and Limitations

In this work, we introduce a selector-density CMI framework that unifies generalization, memorization, and hallucination, and distinguish the factual memory needed for reliable discrimination from the training-selection information associated with sample dependence. Our information-theoretic analysis uses uniformly selected fact sets and a specified separation between atomic facts and independent latent rules. Hallucination is measured through threshold-based false positives, and the joint budget bound assumes perfect atomic recall. The bounds characterize information requirements but do not determine how training dynamics allocate capacity or whether the bounds are attainable. The Gaussian model restricts both fact types to a single shared location parameter, while the synthetic relational experiments provide qualitative evidence without directly measuring the information components. Extending the joint analysis to partial recall and more general fact distributions, and connecting information allocation to training dynamics and factuality in open-ended language generation, remain important directions for future work.

AI USE STATEMENT

We used GPT-6 to polish language, correct grammar, and improve the manuscript's readability. It also assisted with parts of the theoretical proofs, primarily in the Gaussian example, and with implementing the experiments through Codex. The authors take responsibility for the final proofs, code, results, and text.

ETHICS STATEMENT

Our work investigates hallucination in generative models, which may have societal consequences when generated claims inform consequential decisions. We adhere to the ICLR Code of Ethics. However, this paper focuses solely on theoretical understanding and controlled experimental validation. We do not propose or evaluate real-world deployments. The experiments use synthetic relational data and involve no human subjects or personally identifiable information. We confirm that there are no conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

Code for the controlled experiments, including scripts for generating the experimental data, is available at `https://anonymous.4open.science/r/review-code-5c8a/`. Additional experimental details are provided in Appendix C. Appendix D restates the theoretical results presented in the main text and provides the proofs omitted there.

REFERENCES

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*. PMLR, 2018.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pp. 123–132, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yuxin Dong, Haoran Guo, Tieliang Gong, Wen Wen, and Chen Li. Exactly tight information-theoretic generalization bounds via binary jensen-shannon divergence. In *Forty-second International Conference on Machine Learning*, 2025.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, pp. 954–959, 2020.

Vitaly Feldman, Guy Kornowski, and Xin Lyu. Trade-offs in data memorization via strong data processing inequalities. In *The Thirty Eighth Annual Conference on Learning Theory*, volume 291, pp. 1935–1973. PMLR, 2025.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xinran Gu, Kaifeng Lyu, Jiazheng Li, and Jingzhao Zhang. Data mixing can induce phase transitions in knowledge acquisition. In *Advances in Neural Information Processing Systems*, volume 38, pp. 163200–163244, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/eee7ae5cf0c4356c2aeca400771791aa-Abstract-Conference.html.

Anxin Guo and Jingwei Li. Hallucination is a consequence of space-optimality: A rate-distortion theorem for membership testing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Aj7lNFaY0M.

Mahdi Haghifam, Shay Moran, Daniel M Roy, and Gintare Karolina Dziugaite. Understanding generalization via leave-one-out conditional mutual information. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2487–2492. IEEE, 2022.

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, 2021.

Fredrik Hellström and Giuseppe Durisi. A new family of generalization bounds using samplewise evaluated CMI. In *Advances in Neural Information Processing Systems*, 2022.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and pac-bayes. *Foundations and Trends in Machine Learning*, 18(1):1–223, 2025. doi: 10.1561/2200000112. URL https://doi.org/10.1561/2200000112.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171, 2024.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653:1047–1051, 2026. doi: 10.1038/s41586-026-10549-w. URL https://www.nature.com/articles/s41586-026-10549-w.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. On the limits of language generation: Trade-offs between hallucination and mode collapse. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1732–1743, 2025. doi: 10.1145/3717823.3718108. URL https://arxiv.org/abs/2411.09642.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yang Lu, Matthias Frey, Margreta Kuijper, and Jingge Zhu. On unified and sharpened cmi bounds for generalization errors, 2026. URL https://arxiv.org/abs/2605.21056.

John Xavier Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much can language models memorize? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bA6BgSbaUi.

Zhixuan Pan, Shaowen Wang, Pengfei Liao, and Jian Li. Understanding LLM behaviors via compression: Data generation, knowledge acquisition and scaling laws. In *Advances in Neural Information Processing Systems*, volume 38, pp. 186604–186654, 2025. doi: 10.52202/085713-5613. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f61d7778e89b9221d1ea0ce8428b7014-Abstract-Conference.html.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge university press, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Mohamad Rida Rammal, Alessandro Achille, Aditya Golatkar, Suhas Diggavi, and Stefano Soatto. On leave-one-out conditional mutual information for generalization. In *Advances in Neural Information Processing Systems*, 2022.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3437–3452. PMLR, 2020. URL https://proceedings.mlr.press/v125/steinke20a.html.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95238–95265, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad217e0c7fecc71bdf48660ad6714b07-Abstract-Conference.html.

Ziqiao Wang and Yongyi Mao. Sample-conditioned hypothesis stability sharpens information-theoretic generalization bounds. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. In *International Conference on Machine Learning*. PMLR, 2023b.

Ziqiao Wang and Yongyi Mao. Generalization bounds via conditional f-information. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017.

Xinhao Yao, Ruifeng Ren, Yun Liao, Lizhong Ding, and Yong Liu. Compositional generalization from learned skills via CoT training: A theoretical and structural analysis for reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VLjTqLB0J9.

Jiayuan Ye, Vitaly Feldman, and Kunal Talwar. Cram less to fit more: Training data pruning improves memorization of facts, 2026. URL https://arxiv.org/abs/2604.08519.

Appendix

A Related works

A.1 CMI Bounds for Generalization

Conditional mutual information provides a well-established framework for controlling generalization through information about training membership (Steinke & Zakynthinou, 2020; Hellström et al., 2025). Conditioning on a supersample isolates the information a model reveals about the training selection, connecting generalization to compression, stability, and privacy. Subsequent work measures this information in predictions through functional CMI (f-CMI) (Harutyunyan et al., 2021), in losses through evaluated CMI (e-CMI) (Hellström & Durisi, 2022), and in loss differences through ld-CMI (Wang & Mao, 2023b). These refinements discard information irrelevant to the evaluated loss and can sharpen parameter-level bounds. Further improvements use sample-conditioned hypothesis stability (Wang & Mao, 2023a), conditional f-information based on f-divergences (Wang & Mao, 2024), and binary Jensen–Shannon divergence (Dong et al., 2025).

The selection mechanism also matters. Leave-one-out CMI connects generalization to cross-validation and stability (Haghifam et al., 2022; Rammal et al., 2022); the leave-m-out framework unifies a broader family of supersample constructions (Lu et al., 2026). Our uniform selection of n samples from N candidates is its unpartitioned case. We combine this variable selection density with evaluator-level information: losses yield generalization bounds, whereas factuality scores yield limits on factual discrimination. The common object is the information revealed about selection; its statistical role depends on what the selector and evaluator represent.

A.2 Theoretical Perspectives on Hallucination

Theoretical accounts of hallucination use several complementary frameworks: calibration and Good–Turing missing-mass estimation (Kalai & Vempala, 2024), Is-It-Valid classification and evaluation incentives (Kalai et al., 2025; 2026), and rate–distortion theory for membership testing (Guo & Li, 2026). Compression-based analyses explain why limited capacity can prioritize frequent knowledge over rare facts (Pan et al., 2025); formal language generation exposes a trade-off between validity and coverage (Kalavasis et al., 2025). The membership-testing formulation is closest to our random-fact analysis: in the sparse-fact limit, the minimum KL divergence between admissible fact and nonfact score laws characterizes the optimal memory cost (Guo & Li, 2026). Our Selector-Density CMI framework provides a unified view of factual discrimination and membership information.

A.3 Memorization and Capacity Allocation

Memorization can be necessary for generalization. Feldman (2020) shows that near-optimal prediction on long-tailed subpopulation mixtures can require memorizing rarely observed labels. This connects to singleton-based hallucination bounds (Kalai & Vempala, 2024): rare observations create a need for memorization in the former and reveal missing factual evidence in the latter. Brown et al. (2021) further construct learning problems where high accuracy requires storing substantial training information, including task-irrelevant details. Using strong data processing inequalities, Feldman et al. (2025) characterize how the excess memorization required for accurate learning decreases as more samples become available. Their meta-distribution formulation conditions on the data-generating parameter, separating distributional knowledge from residual sample information. Morris et al. (2026) make a related distinction through compression-based measures of distributional knowledge and sample-specific memorization.

For factual knowledge, Ye et al. (2026) measure information about world facts, relate it to factual accuracy, and show that data pruning can improve the use of limited memorization capacity. Gu et al. (2025) show that mixing random facts with a structurally complex data source can reallocate finite capacity and induce phase transitions in knowledge acquisition. The mixing ratio changes the loss reduction available per unit of capacity, so the optimal allocation across sources can shift abruptly. In our decomposition, world information is resolved into transferable rules and atomic facts, while selector CMI measures training-choice information conditional on the factual world.

B CALIBRATION AT FINITE TRUTH DENSITY

As a by-product result, we next consider perfectly calibrated scores, without requiring the sparse limit. Let $T \in [0,1]$ be a reported probability of truth. Perfect calibration means $\mathbb{P}(U = 1 \mid T) = T$ almost surely. Since U is binary, this condition is equivalently written as $\mathbb{E}[U \mid T] = T$. We have the following bounds.

Corollary B.1. *Under perfect calibration, for $\tau \in (0,1]$ with $\mathbb{P}(T \geq \tau) > 0$, the precision of the accepted claims satisfies $\mathbb{P}(U = 1 \mid T \geq \tau) \geq \tau$, and therefore*

$$\left[\alpha_\tau - \sqrt{\frac{h(\rho) - \mathbb{E}_T[h(T)]}{2\rho(1-\rho)}}\right]_+ \leq \beta_\tau \leq \frac{\rho(1-\tau)}{(1-\rho)\tau}\alpha_\tau.$$

Proof. We first establish the change-of-measure identities used below. For any measurable set $A \subseteq [0,1]$, calibration gives

$$\rho P(A) = \mathbb{E}[\mathbf{1}\{T \in A\}U] = \mathbb{E}[\mathbf{1}\{T \in A\}\mathbb{E}[U \mid T]] = \int_A t\, M_\rho(dt),$$
$$(1-\rho)Q(A) = \mathbb{E}[\mathbf{1}\{T \in A\}(1-U)] = \mathbb{E}[\mathbf{1}\{T \in A\}(1-T)] = \int_A (1-t)\, M_\rho(dt).$$

Thus $\frac{dP}{dM_\rho}(t) = t/\rho$ and $\frac{dQ}{dM_\rho}(t) = (1-t)/(1-\rho)$, so

$$\alpha_\tau = \frac{1}{\rho}\int_{[\tau,1]} t\, M_\rho(dt), \qquad \beta_\tau = \frac{1}{1-\rho}\int_{[\tau,1]} (1-t)\, M_\rho(dt).$$

Perfect calibration and the tower property give

$$\begin{aligned}\mathbb{P}(U = 1 \mid T \geq \tau) &= \mathbb{E}[U \mid T \geq \tau] \\ &= \mathbb{E}[\mathbb{E}[U \mid T] \mid T \geq \tau] \\ &= \mathbb{E}[T \mid T \geq \tau] \geq \tau.\end{aligned}$$

Bayes' rule then yields

$$\mathbb{P}(U = 1 \mid T \geq \tau) = \frac{\rho\alpha_\tau}{\rho\alpha_\tau + (1-\rho)\beta_\tau} \geq \tau,$$

which gives the stated upper bound on β_τ.

For the lower bound, perfect calibration implies $U \mid T \sim \mathrm{Bern}(T)$. Therefore,

$$I(U;T) = H(U) - H(U \mid T) = h(\rho) - \mathbb{E}_T[h(T)].$$

We also need the weighted Pinsker inequality. For $M_\rho = \rho P + (1-\rho)Q$, Pinsker's inequality gives

$$\mathrm{TV}(P, M_\rho) = (1-\rho)\mathrm{TV}(P,Q) \leq \sqrt{\frac{\mathrm{D_{KL}}\left(P \| M_\rho\right)}{2}},$$
$$\mathrm{TV}(Q, M_\rho) = \rho\mathrm{TV}(P,Q) \leq \sqrt{\frac{\mathrm{D_{KL}}\left(Q \| M_\rho\right)}{2}},$$

where TV denotes total variation distance. Multiplying the squared inequalities by 2ρ and $2(1-\rho)$, respectively, and adding gives

$$J_\rho(P,Q) \geq 2\rho(1-\rho)^2\mathrm{TV}^2(P,Q) + 2(1-\rho)\rho^2\mathrm{TV}^2(P,Q) = 2\rho(1-\rho)\mathrm{TV}^2(P,Q).$$

For $A_\tau = \mathbf{1}\{T \geq \tau\}$, data processing therefore implies

$$I(U;T) \geq I(U;A_\tau) = J_\rho\big(\mathrm{Bern}(\alpha_\tau), \mathrm{Bern}(\beta_\tau)\big) \geq 2\rho(1-\rho)(\alpha_\tau - \beta_\tau)^2.$$

Combining this with the entropy identity yields

$$|\alpha_\tau - \beta_\tau| \leq \sqrt{\frac{h(\rho) - \mathbb{E}_T[h(T)]}{2\rho(1-\rho)}}.$$

The stated lower bound follows because $\beta_\tau \geq 0$. □

The two sides of the bound above serve different purposes. The lower bound quantifies how limited truth-selector information restricts the separation between recall and hallucination rate. The upper bound follows from the precision guarantee associated with calibration, but does not guarantee substantial recall. Our result is complementary to Kalai & Vempala (2024), who establish statistical hallucination lower bounds under a semantic generative-calibration condition and assumptions on arbitrary facts whose truth cannot be inferred from the observed data. Here, we instead study binary correctness calibration, $\mathbb{P}(U = 1 \mid T) = T$, and the false-positive rate of a claim-acceptance rule. These differences in calibration and error definitions preclude a direct identification of the two results. In particular, the calibrated oracle $T = U$ achieves zero false positives and perfect recall. Note that its existence as a joint distribution does not imply that it is learnable under the information constraints of a given training problem.

C Experimental Details and Additional Results

C.1 Data Construction and Splits

We build the knowledge graph around centers e_c, each connecting d_{in} incoming and d_{out} outgoing entities. The atomic facts $(e_{h,i}, r_i^{\text{in}}, e_c)$ and $(e_c, r_j^{\text{out}}, e_{t,j})$ generate the composite fact $(e_{h,i}, r_{ij}^{\text{comp}}, e_{t,j})$ through the shared rule $r_i^{\text{in}} \circ r_j^{\text{out}} = r_{ij}^{\text{comp}}$ (Figure 3). Different centers use disjoint entities but share the same rules, so each rule applies across multiple factual instances. Each (e_h, r_i) query has a unique true tail. We use $d_{\text{in}} = d_{\text{out}} = d$, giving d^2 composition rules and $2d$ atomic facts per complete center.

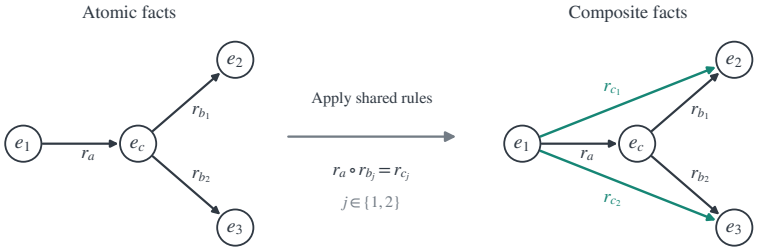


Figure 3: Shared composition rules map atomic facts to composite facts.

All atomic facts enter training. We distinguish three center types by their composite facts (Figure 4):

- **Full-training centers**: all composite facts enter training.
- **Atomic-only centers**: no composite facts enter training or testing.
- **Split-test centers**: half the composite facts enter training and the other half enter testing, with balanced rule coverage; each test fact is unseen during training, but both atomic premises are observed.

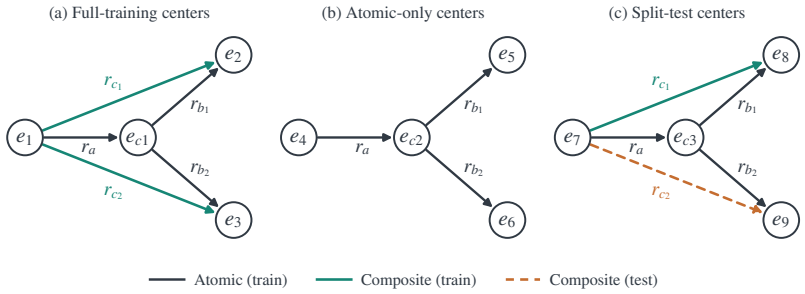


Figure 4: Training and test splits across the three center types. Solid edges enter training; dashed edges enter testing.

C.2 Models and Training

We use randomly initialized GPT-style decoder-only Transformers (Radford et al., 2019) with the configurations in Table 1. Entities and relations are represented as tokens, and each fact $z = (e_h, r_i, e_t)$ is encoded as $(\texttt{BOS}, e_h, r_i, e_t)$ for autoregressive training, with loss

$$\begin{aligned}\mathcal{L}(w, z) &= -\frac{1}{3} \log p_w(e_h, r_i, e_t \mid \texttt{BOS}) \\ &= -\frac{1}{3}\Big[\log p_w(e_h \mid \texttt{BOS}) + \log p_w(r_i \mid \texttt{BOS}, e_h) \\ &\qquad + \log p_w(e_t \mid \texttt{BOS}, e_h, r_i)\Big].\end{aligned}$$

Unless otherwise specified, the vocabulary contains 40,961 tokens: 32,768 entity tokens, 8,192 relation tokens, and BOS. The 1,024-rule graph uses 1,280 relation tokens instead, giving a vocabulary of 34,049 tokens. We train using AdamW with learning rate 10^{-3}, batch size 10,000, and weight decay 0.1. Composite test accuracy is the fraction of held-out composite facts for which the highest-probability entity is the true tail.

Table 1: Transformer configurations with a 40,961-token vocabulary.

Parameters (M)	Layers	Hidden size	Attention heads	FFN size
0.66	1	8	1	32
1.31	1	16	1	64
1.65	2	20	4	80
1.98	2	24	4	96
2.31	2	28	4	112
2.65	2	32	4	128
4.02	3	48	4	192
8.31	4	96	4	384
18.40	6	192	8	768

C.3 Experimental Setup

Following Section 5, we examine how composite-fact exposure and rule-learning burden interact with model capacity to shape atomic hallucination. The two experiments vary the training mixture and rule count, respectively, across model sizes.

Training mixture. With all n_1 atomic facts in training, we vary the number n of composite training facts to adjust $a = n/(n_1 + n)$, keeping the atomic facts, rules, and composite test set fixed. The total number of atomic-only and full-training centers remains constant: adding an atomic-only center's composite edges to training converts it into a full-training center and increases a, while split-test centers remain unchanged. We consider graphs with 4,096 and 1,024 rules, with a ranging over $[0.25, 0.95]$ and $[0.23, 0.94]$, respectively.

Rule count. At fixed n_1 and a, we vary the number of composition rules across 16, 256, and 6,400 while matching the atomic and composite training counts and the composite test count. This changes the structural learning burden without adding training facts. We compare these settings at training proportions $a \in [0.40, 0.55]$. Table 2 summarizes the datasets.

Table 2: Datasets for the training-mixture and rule-count experiments.

Experiment	Rules	Atomic train n_1	Composite test
Training mixture	4,096	24,704	8,192
	1,024	3,520	1,024
Rule count	16 / 256 / 6,400	24,704	6,400

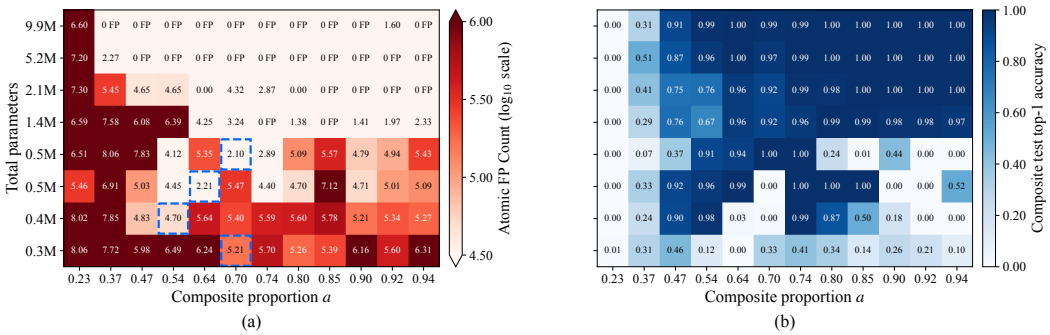


Figure 5: Training-mixture results with 1,024 rules: (a) AtomicFP at perfect fact recall; (b) composite test accuracy.

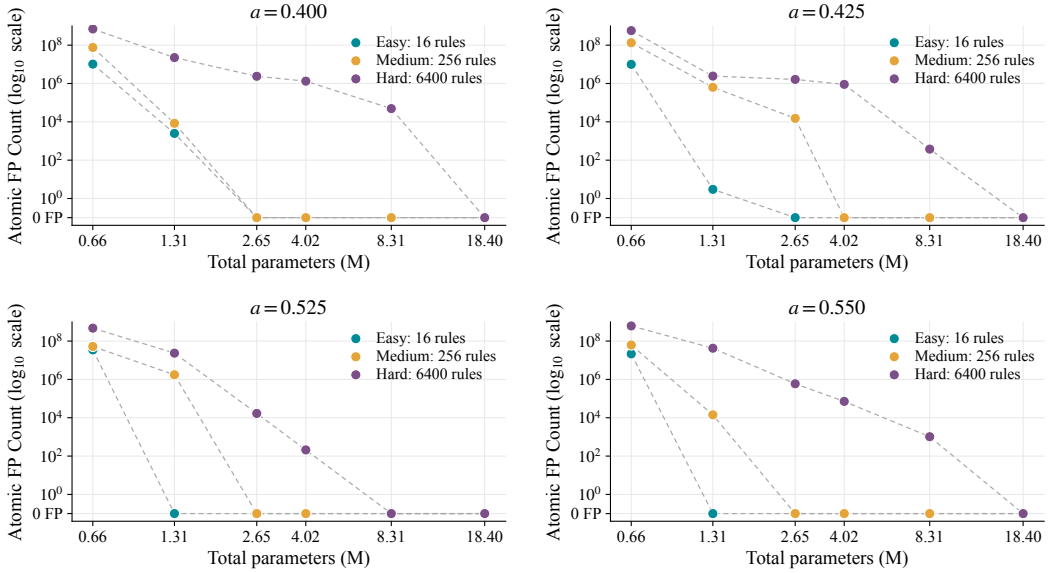


Figure 6: AtomicFP across model sizes and rule counts at four additional training mixtures.

C.4 Additional Results

Figure 2(a) in the main text uses 4,096 rules, $n_1 = 24{,}704$, and $a \in [0.25, 0.95]$. Figure 5 reports AtomicFP alongside composite test accuracy for 1,024 rules, $n_1 = 3{,}520$, and $a \in [0.23, 0.94]$. Smaller models favor intermediate mixtures for low AtomicFP, while larger models combine low FP with near-perfect composite accuracy over a broader range. The mixture thus affects both atomic factual discrimination and inference on unseen composite facts.

Figure 2(b) in the main text compares 16, 256, and 6,400 rules at $n_1 = 24{,}704$ and $a = 0.5$. Figure 6 extends this comparison to other training proportions $a \in [0.40, 0.55]$, keeping n_1 fixed. Across the displayed mixtures, models learning 6,400 rules generally produce more atomic false positives than equally sized models learning 16 or 256 rules. Larger models ease this trade-off: the capacity needed to reach zero FP depends on both rule count and training mixture.

D Omitted Proofs

Lemma 2.1. *Under the preceding notation, for any fixed $i \in [N]$, we have*

$$I(U_i; T_i) = J_\rho(P, Q). \tag{1}$$

Proof. Since $U_i \sim \mathrm{Bern}(\rho)$, the marginal law of T_i is $M_\rho = \rho P + (1-\rho)Q$. By the definition of mutual information,

$$\begin{aligned} I(U_i;T_i) &= \mathbb{E}_{U_i}\mathbb{E}_{T_i|U_i}\left[\log \frac{dP_{T_i|U_i}}{dP_{T_i}}(T_i)\right] \\ &= \rho\,\mathbb{E}_{T_i|U_i=1}\left[\log \frac{dP}{dM_\rho}(T_i)\right] + (1-\rho)\,\mathbb{E}_{T_i|U_i=0}\left[\log \frac{dQ}{dM_\rho}(T_i)\right] \\ &= \rho \mathrm{D}_{\mathrm{KL}}\left(P\|M_\rho\right) + (1-\rho)\mathrm{D}_{\mathrm{KL}}\left(Q\|M_\rho\right) \\ &= J_\rho(P,Q). \end{aligned}$$

□

For $a, b \in [0,1]$, let $d_{\mathrm{JS},\rho}(a,b) \coloneqq J_\rho\big(\mathrm{Bern}(a), \mathrm{Bern}(b)\big)$ be the binary JS divergence, which is also used in Dong et al. (2025). We then have the following theorem.

Theorem D.1 (Full statement of Theorem 3.1)**.** *Let $J \sim \mathrm{Unif}([N])$ be independent of $(\widetilde{Z}, K, W)$, and define $U \coloneqq \mathbf{1}\{J \in K\}$ and $T \coloneqq T_J$. Then*

$$d_{\mathrm{JS},\rho}\Big(\mathbb{E}\left[\hat{L}_n\right], \mathbb{E}\left[\hat{L}_m\right]\Big) \le I(U;T) \le I(U;T \mid \widetilde{Z}, J) = \frac{1}{N}\sum_{i=1}^{N} I(U_i;T_i \mid \widetilde{Z}). \tag{9}$$

In particular,

$$|\mathcal{E}(\mathcal{A})| \le \sqrt{\frac{I(U;T \mid \widetilde{Z}, J)}{2\rho(1-\rho)}}. \tag{10}$$

Proof. A direct calculation gives

$$\begin{aligned} \mathbb{E}[\hat{L}_n] &= \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}[U_i T_i] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[T_i \mid U_i = 1] = \mathbb{E}[T \mid U = 1], \\ \mathbb{E}[\hat{L}_m] &= \frac{1}{m}\sum_{i=1}^{N}\mathbb{E}[(1-U_i)T_i] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[T_i \mid U_i = 0] = \mathbb{E}[T \mid U = 0]. \end{aligned}$$

To obtain the binary projection, given $T \in [0,1]$, output one with probability T and zero otherwise, using randomness independent of U. The conditional output laws are $\mathrm{Bern}(\mathbb{E}[T \mid U = 1])$ and $\mathrm{Bern}(\mathbb{E}[T \mid U = 0])$. By Lemma 2.1, the mutual information between U and this binary output is the weighted JS divergence between these laws. Since this randomization uses only T, data processing gives

$$d_{\mathrm{JS},\rho}\big(\mathbb{E}[\hat{L}_n], \mathbb{E}[\hat{L}_m]\big) \le I(U;T).$$

Although U is a function of (J,K), uniform subset selection and $K \perp\!\!\!\perp \widetilde{Z}$ imply $\mathbb{P}(U = 1 \mid \widetilde{Z}, J) = n/N = \rho$. Thus $U \perp\!\!\!\perp (\widetilde{Z}, J)$, and

$$I(U;T) \le I(U;T,\widetilde{Z},J) = I(U;T \mid \widetilde{Z}, J).$$

Conditioning on the independent uniform index J gives

$$I(U;T \mid \widetilde{Z}, J) = \frac{1}{N}\sum_{i=1}^{N} I(U_i;T_i \mid \widetilde{Z}),$$

completing Eq. (9).

For the final bound, we establish the weighted JS–TV inequality. For any P, Q and $M_\rho = \rho P + (1-\rho)Q$, Pinsker's inequality gives

$$\begin{aligned} \mathrm{TV}(P, M_\rho) &= (1-\rho)\mathrm{TV}(P,Q) \le \sqrt{\frac{\mathrm{D}_{\mathrm{KL}}\left(P\|M_\rho\right)}{2}}, \\ \mathrm{TV}(Q, M_\rho) &= \rho\,\mathrm{TV}(P,Q) \le \sqrt{\frac{\mathrm{D}_{\mathrm{KL}}\left(Q\|M_\rho\right)}{2}}, \end{aligned}$$

where TV denotes total variation distance. Consequently,

$$\begin{aligned} J_\rho(P,Q) &= \rho \mathrm{D}_{\mathrm{KL}}\left(P \| M_\rho\right) + (1-\rho)\mathrm{D}_{\mathrm{KL}}\left(Q \| M_\rho\right) \\ &\geq 2\rho(1-\rho)^2 \mathrm{TV}^2(P,Q) + 2(1-\rho)\rho^2 \mathrm{TV}^2(P,Q) \\ &= 2\rho(1-\rho)\mathrm{TV}^2(P,Q). \end{aligned}$$

Applying this inequality to the two Bernoulli risk laws yields

$$d_{\mathrm{JS},\rho}\big(\mathbb{E}[\hat{L}_n], \mathbb{E}[\hat{L}_m]\big) \geq 2\rho(1-\rho)\big(\mathbb{E}[\hat{L}_n] - \mathbb{E}[\hat{L}_m]\big)^2.$$

Combining this with Eq. (9) and $\mathcal{E}(\mathcal{A}) = \mathbb{E}[\hat{L}_m] - \mathbb{E}[\hat{L}_n]$ proves Eq. (10) and Eq. (2). □

Theorem 3.2. *Under the construction above, we have*

$$|\mathcal{E}(\mathcal{A})| \leq \min\left\{ \frac{N}{m}\sqrt{\frac{\mathrm{eCMI}_\rho^\ell(\mathcal{A})}{2n}}, \frac{N}{n}\sqrt{\frac{\mathrm{eCMI}_\rho^\ell(\mathcal{A})}{2m}} \right\}. \tag{3}$$

Proof. For $t^N \in [0,1]^N$ and $k \subseteq [N]$ with $|k| = n$, let

$$G(t^N, k) = \frac{1}{m}\sum_{i \notin k} t_i - \frac{1}{n}\sum_{i \in k} t_i.$$

Fix $\widetilde{Z} = \tilde{z}$ and compare the true conditional joint law of (T^N, K) with the decoupled law

$$Q_{\tilde{z}} = P_{T^N|\widetilde{Z}=\tilde{z}} \otimes P_K.$$

Under $Q_{\tilde{z}}$, K is a uniform subset independent of T^N, and $\mathbb{E}_{Q_{\tilde{z}}}[G \mid T^N] = 0$. The sampling-without-replacement comparison and Hoeffding's lemma (Hoeffding, 1963) give, uniformly in t^N,

$$\begin{aligned} \log \mathbb{E}_K[e^{\lambda G(t^N,K)}] &\leq \frac{\lambda^2 N^2}{8nm^2}, \\ \log \mathbb{E}_K[e^{\lambda G(t^N,K)}] &\leq \frac{\lambda^2 N^2}{8mn^2}. \end{aligned}$$

Indeed,

$$G(t^N,K) = \frac{N}{m}\left(\frac{1}{N}\sum_i t_i - \frac{1}{n}\sum_{i\in K} t_i\right) = \frac{N}{n}\left(\frac{1}{m}\sum_{i\notin K} t_i - \frac{1}{N}\sum_i t_i\right),$$

so the two bounds follow by applying the sample-mean exponential-moment bound to the selected and unselected subsets, respectively.

Since $K \perp\!\!\!\perp \widetilde{Z}$, the KL divergence between the true conditional law and $Q_{\tilde{z}}$ equals $I(T^N; K \mid \widetilde{Z} = \tilde{z})$. The Donsker–Varadhan variational inequality (Donsker & Varadhan, 1983), applied to λG and $-\lambda G$, gives, for $\lambda > 0$,

$$\left|\mathbb{E}[G \mid \widetilde{Z} = \tilde{z}]\right| \leq \frac{I(T^N; K \mid \widetilde{Z} = \tilde{z})}{\lambda} + \frac{\lambda N^2}{8nm^2}.$$

Optimizing over $\lambda > 0$ yields

$$\left|\mathbb{E}[G \mid \widetilde{Z} = \tilde{z}]\right| \leq \frac{N}{m}\sqrt{\frac{I(T^N; K \mid \widetilde{Z} = \tilde{z})}{2n}}.$$

The second exponential-moment bound yields the complementary inequality with n, m interchanged. Averaging each bound over $\widetilde{Z}$ and applying Jensen's inequality gives

$$|\mathcal{E}(\mathcal{A})| \leq \min\left\{ \frac{N}{m}\sqrt{\frac{\mathrm{eCMI}_\rho^\ell(\mathcal{A})}{2n}}, \frac{N}{n}\sqrt{\frac{\mathrm{eCMI}_\rho^\ell(\mathcal{A})}{2m}} \right\},$$

as claimed. □

Remark D.1 (Relation to existing CMI bounds)**.** *Our fixed-cardinality selector is close to the leave-m-out supersample setting (Lu et al., 2026). When $N = 2n$ and $m = n$, Theorem 3.2 reduces to $|\mathcal{E}(\mathcal{A})| \le \sqrt{\frac{2\mathrm{eCMI}^{\ell}_{1/2}(\mathcal{A})}{n}}$, which has the same numerical form as the evaluated pairwise-CMI bound (Harutyunyan et al., 2021; Hellström & Durisi, 2022; Wang & Mao, 2023b). The underlying selector laws, however, are different. The original CMI construction selects exactly one instance from each of n prescribed pairs using independent Bernoulli bits (Steinke & Zakynthinou, 2020), whereas our selector is uniform over all size-n subsets of $[2n]$. It may therefore select both instances from one pair and neither from another. Thus, $\rho = 1/2$ recovers the balanced JS weighting, but not the same joint selector law or, in general, the same CMI quantity. Our purpose here is to establish a common selector-information perspective, rather than to improve upon existing generalization bounds. For a comprehensive treatment, including tighter bounds, fast-rate results, and extensions beyond bounded losses, we refer to Hellström et al. (2025).*

Lemma 3.1. *Under the setting of Lemma 2.1, as $\rho = \frac{n}{N} \to 0$, $\frac{I(U;T)}{\rho} = \frac{J_\rho(P,Q)}{\rho} \longrightarrow \mathrm{D_{KL}}\left(P\|Q\right)$.*

Proof. Hold the conditional score laws P, Q fixed as $\rho \downarrow 0$, and let $M_\rho = \rho P + (1-\rho)Q$. Since

$$\mathrm{TV}(M_\rho, Q) = \rho \mathrm{TV}(P, Q) \longrightarrow 0,$$

lower semicontinuity of KL divergence gives

$$\mathrm{D_{KL}}\left(P\|Q\right) \le \liminf_{\rho\downarrow 0} \mathrm{D_{KL}}\left(P\|M_\rho\right).$$

By nonnegativity of KL divergence,

$$\frac{J_\rho(P,Q)}{\rho} = \mathrm{D_{KL}}\left(P\|M_\rho\right) + \frac{1-\rho}{\rho}\mathrm{D_{KL}}\left(Q\|M_\rho\right) \ge \mathrm{D_{KL}}\left(P\|M_\rho\right).$$

Consequently,

$$\liminf_{\rho\downarrow 0} \frac{J_\rho(P,Q)}{\rho} \ge \mathrm{D_{KL}}\left(P\|Q\right). \tag{11}$$

If $\mathrm{D_{KL}}\left(P\|Q\right) = \infty$, this already proves the claim in the extended-real sense.

Now suppose that $\mathrm{D_{KL}}\left(P\|Q\right) < \infty$, so $P \ll Q$. Expanding the divergences and using $M_\rho = \rho P + (1-\rho)Q$ yields

$$\begin{aligned}
\rho \mathrm{D_{KL}}\left(P\|Q\right) - J_\rho(P,Q) &= \rho \int \log \frac{dM_\rho}{dQ}\, dP \\
&\quad + (1-\rho) \int \log \frac{dM_\rho}{dQ}\, dQ \\
&= \int \log \frac{dM_\rho}{dQ}\, dM_\rho \\
&= \mathrm{D_{KL}}\left(M_\rho\|Q\right).
\end{aligned}$$

Thus

$$\frac{J_\rho(P,Q)}{\rho} = \mathrm{D_{KL}}\left(P\|Q\right) - \frac{\mathrm{D_{KL}}\left(M_\rho\|Q\right)}{\rho} \le \mathrm{D_{KL}}\left(P\|Q\right).$$

Together with Eq. (11), this proves

$$\lim_{\rho\to 0} \frac{J_\rho(P,Q)}{\rho} = \mathrm{D_{KL}}\left(P\|Q\right).$$

Lemma 2.1 gives $I(U;T) = J_\rho(P,Q)$, completing the stated limit. In the finite-KL case, the same identity also implies

$$\frac{\mathrm{D_{KL}}\left(M_\rho\|Q\right)}{\rho} = \mathrm{D_{KL}}\left(P\|Q\right) - \frac{J_\rho(P,Q)}{\rho} \longrightarrow 0,$$

so $\mathrm{D_{KL}}\left(M_\rho\|Q\right) = o(\rho)$ follows from the limit rather than being assumed. □

Theorem 3.3. *In the random-fact setting, suppose that* $\frac{1}{n} I(T^N; K) \le R$. *Then in the sparse limit* $\rho \to 0$ *and* $n \to \infty$, *for every* τ *s.t.* $\alpha_\tau > 0$, *we have* $\beta_\tau \ge \exp\left(-\frac{R + h(\alpha_\tau)}{\alpha_\tau}\right)$. *In particular, perfect recall* $\alpha_\tau = 1$ *implies* $\beta_\tau \ge e^{-R}$.

Proof. The claim universe $\widetilde{Z}$ is fixed, so conditioning on it is implicit. Recall that $U_i = \mathbf{1}\{i \in K\}$, $U = U_J$, and $T = T_J$. Since K is uniform over the size-n subsets of $[N]$, $\mathbb{P}(U_i = 1) = \binom{N-1}{n-1} / \binom{N}{n} = \rho$ for every i. Thus $U \perp\!\!\!\perp J$, and the independence of J from (K, T^N) gives

$$\begin{aligned} J_\rho(P, Q) &= I(U; T) \le I(U; T, J) = I(U; T \mid J) \\ &= \frac{1}{N} \sum_{i=1}^{N} I(U_i; T_i). \end{aligned}$$

To bound this sum by the information in the full score vector, note that $(U_1, \dots, U_N)$ determines K and conversely. Conditional entropy subadditivity and the fact that T_i is a coordinate of T^N therefore imply

$$H(K \mid T^N) = H(U_1, \dots, U_N \mid T^N) \le \sum_{i=1}^{N} H(U_i \mid T^N) \le \sum_{i=1}^{N} H(U_i \mid T_i).$$

Using $H(U_i) = h(\rho)$ and $H(K) = \log \binom{N}{n}$, we obtain

$$\begin{aligned} \sum_{i=1}^{N} I(U_i; T_i) &= N h(\rho) - \sum_{i=1}^{N} H(U_i \mid T_i) \\ &\le N h(\rho) - H(K \mid T^N) \\ &= I(T^N; K) + N h(\rho) - \log \binom{N}{n}. \end{aligned}$$

Since $N\rho = n$, the assumed information budget yields

$$\frac{J_\rho(P, Q)}{\rho} \le \frac{I(T^N; K)}{n} + \frac{N h(\rho) - \log \binom{N}{n}}{n} \le R + \frac{N h(\rho) - \log \binom{N}{n}}{n}. \tag{12}$$

The correction accounts for the dependence among the fixed-cardinality membership indicators. As $n \to \infty$ and $\rho \to 0$, we also have $N - n \to \infty$, and Stirling's formula gives

$$\log \binom{N}{n} = N h(\rho) - \frac{1}{2} \log\big(2\pi n(1-\rho)\big) + O\left(\frac{1}{n} + \frac{1}{N-n}\right).$$

Consequently,

$$\frac{N h(\rho) - \log \binom{N}{n}}{n} = \frac{\log\big(2\pi n(1-\rho)\big)}{2n} + O\left(\frac{1}{n^2} + \frac{1}{n(N-n)}\right) = o(1),$$

so Eq. (12) gives $J_\rho(P, Q)/\rho \le R + o(1)$.

Thresholding is a post-processing of the score, so the data-processing inequality gives

$$\frac{J_\rho(P, Q)}{\rho} \ge \frac{J_\rho\big(\mathrm{Bern}(\alpha_\tau), \mathrm{Bern}(\beta_\tau)\big)}{\rho}.$$

Combining this with the preceding bound and dropping the nonnegative second term in the definition of J_ρ yields

$$\mathrm{D}_{\mathrm{KL}}\left(\mathrm{Bern}(\alpha_\tau) \| \mathrm{Bern}\big(\rho\alpha_\tau + (1-\rho)\beta_\tau\big)\right) \le \frac{J_\rho\big(\mathrm{Bern}(\alpha_\tau), \mathrm{Bern}(\beta_\tau)\big)}{\rho} \le R + o(1).$$

For any limiting recall–false-positive pair $(\alpha_\tau, \beta_\tau)$, lower semicontinuity of KL divergence therefore gives $\mathrm{D}_{\mathrm{KL}}\left(\mathrm{Bern}(\alpha_\tau) \| \mathrm{Bern}(\beta_\tau)\right) \le R$. This argument also allows the score laws and threshold to vary along the sparse sequence. For $\alpha_\tau > 0$, the exponential lower bound follows from

$$\mathrm{D}_{\mathrm{KL}}\left(\mathrm{Bern}(\alpha_\tau) \| \mathrm{Bern}(\beta_\tau)\right) = -h(\alpha_\tau) - \alpha_\tau \log \beta_\tau - (1-\alpha_\tau) \log(1-\beta_\tau) \ge -h(\alpha_\tau) - \alpha_\tau \log \beta_\tau.$$

Rearranging gives $\beta_\tau \ge \exp\big(-(R + h(\alpha_\tau))/\alpha_\tau\big)$. For $\alpha_\tau = 1$, $h(1) = 0$ gives $\beta_\tau \ge e^{-R}$. □

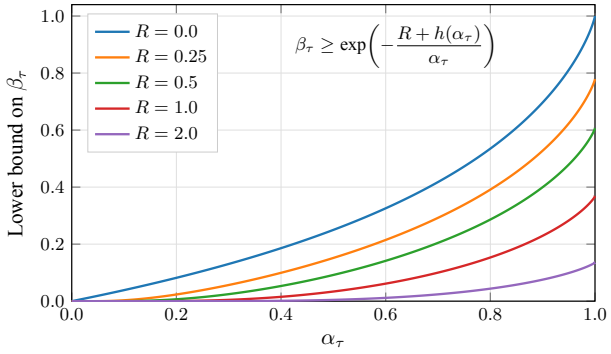


Figure 7: Lower bounds on the false-positive hallucination rate versus fact recall in the sparse random-fact setting, under a per-fact information budget R.

Lemma 4.1. *In the rules-and-facts setting, we have*

$$I(W;S) = \underbrace{I(W;\Theta)}_{\textit{Learning}} + \underbrace{I(W;B \mid \Theta)}_{\textit{Atomic fact memory}} + \underbrace{I(W;K \mid \widetilde{Z})}_{\textit{Excess memorization}}. \tag{4}$$

Proof. Since S is a deterministic function of (B, Θ, K) and $(B, \Theta, K) \to S \to W$ forms a Markov chain, the Markov property and the chain rule give

$$I(W;S) = I(W;B,\Theta,K) = I(W;\Theta) + I(W;B \mid \Theta) + I(W;K \mid B,\Theta).$$

Moreover, K is independent of (B, Θ), and the conditional distribution of W depends on (B, Θ, K) only through $\widetilde{Z}_K$. Consequently,

$$(W,K) \perp\!\!\!\perp (B,\Theta) \mid \widetilde{Z}.$$

The first equality below holds because $\widetilde{Z}$ is a deterministic function of (B, Θ); the second follows from the conditional independence above:

$$I(W;K \mid B,\Theta) = I(W;K \mid B,\Theta,\widetilde{Z}) = I(W;K \mid \widetilde{Z}).$$

Substituting this identity proves the decomposition. □

Theorem 4.1. *Suppose that $I(W;S) \leq C$ and that $\Theta \sim \mathrm{Unif}([N_\Theta])$, where $N_\Theta \geq 2$. Assume that there exists a decoder $\widehat{\Theta}(W)$ satisfying $\mathbb{P}(\widehat{\Theta} \neq \Theta) \leq \varepsilon_\Theta$. Define*

$$C_{\mathrm{rule}} := \log N_\Theta - h(\varepsilon_\Theta) - \varepsilon_\Theta \log(N_\Theta - 1), \quad \textit{and } C_{\mathrm{gen}} := \frac{2nm\max\{n,m\}}{N^2}|\mathcal{E}(\mathcal{A})|^2.$$

Then $I(W;\Theta) \geq C_{\mathrm{rule}}$ and $I(W;K \mid \widetilde{Z}) \geq C_{\mathrm{gen}}$.

Moreover, for a threshold decision on atomic claims, let α_τ denote fact recall and β_τ the false-positive hallucination rate. In the sparse atomic-fact limit $\rho_1 = \frac{n_1}{N_1} \to 0$, perfect atomic recall, $\alpha_\tau = 1$, implies

$$\beta_\tau \geq \exp\left(-\frac{C - C_{\mathrm{rule}} - C_{\mathrm{gen}}}{n_1}\right).$$

Proof. We first lower-bound the information used for rule recovery and sample selection, then relate the remaining atomic-fact budget to the false-positive rate under perfect recall. By Fano's inequality,

$$\begin{aligned} I(W;\Theta) &= H(\Theta) - H(\Theta \mid W) \\ &\geq \log N_\Theta - h(\varepsilon_\Theta) - \varepsilon_\Theta \log(N_\Theta - 1) = C_{\mathrm{rule}}. \end{aligned} \tag{13}$$

Let

$$\mathrm{eCMI}^{\ell}_{\rho}(\mathcal{A}) = I(T^N;K \mid \widetilde{Z})$$

denote the evaluated selector-density CMI associated with the bounded loss. The global generalization bound gives

$$|\mathcal{E}(\mathcal{A})| \le \min\left\{\frac{N}{m}\sqrt{\frac{\mathrm{eCMI}_\rho^\ell(\mathcal{A})}{2n}}, \frac{N}{n}\sqrt{\frac{\mathrm{eCMI}_\rho^\ell(\mathcal{A})}{2m}}\right\}.$$

Inverting both inequalities yields

$$\begin{aligned}\mathrm{eCMI}_\rho^\ell(\mathcal{A}) &\ge \max\left\{\frac{2nm^2}{N^2}, \frac{2mn^2}{N^2}\right\}|\mathcal{E}(\mathcal{A})|^2 \\ &= \frac{2nm\max\{n,m\}}{N^2}|\mathcal{E}(\mathcal{A})|^2 = C_{\mathrm{gen}}. \end{aligned} \tag{14}$$

Since T^N is obtained from $(W, \widetilde{Z})$, the data processing inequality gives

$$I(W; K \mid \widetilde{Z}) \ge I(T^N; K \mid \widetilde{Z}) \ge C_{\mathrm{gen}}.$$

By Lemma 4.1,

$$I(W; S) = I(W; \Theta) + I(W; B \mid \Theta) + I(W; K \mid \widetilde{Z}).$$

Combining this decomposition with $I(W; S) \le C$, Eq. (13), and Eq. (14), we obtain

$$I(W; B \mid \Theta) \le C - C_{\mathrm{rule}} - C_{\mathrm{gen}}. \tag{15}$$

It remains to lower-bound the information required for perfect atomic recall. Let K_1 be the atomic-fact selector, so that

$$K_1 \sim \mathrm{Unif}\left\{A \subseteq [N_1] : |A| = n_1\right\}, \qquad \rho_1 = \frac{n_1}{N_1}.$$

Because the atomic candidate pool is fixed and contains distinct claims, B and K_1 are in one-to-one correspondence. Hence

$$I(W; B \mid \Theta) = I(W; K_1 \mid \Theta). \tag{16}$$

For a threshold τ and the model's factuality score T_j of atomic claim b_j, define the set of accepted atomic claims by

$$\widehat{K}_1 \coloneqq \{j \in [N_1] : T_j \ge \tau\}, \qquad L \coloneqq |\widehat{K}_1|.$$

The fact recall satisfies

$$\alpha_\tau = \frac{1}{n_1}\mathbb{E}\Big[|K_1 \cap \widehat{K}_1|\Big].$$

Therefore, under perfect atomic recall, $\alpha_\tau = 1$,

$$\mathbb{E}\Big[|K_1 \setminus \widehat{K}_1|\Big] = n_1(1 - \alpha_\tau) = 0.$$

Since $|K_1 \setminus \widehat{K}_1|$ is nonnegative, it follows that

$$K_1 \subseteq \widehat{K}_1 \qquad \text{almost surely.} \tag{17}$$

This is the step at which perfect recall is essential.

The false-positive hallucination rate is

$$\beta_\tau = \frac{1}{N_1 - n_1}\mathbb{E}\Big[|\widehat{K}_1 \setminus K_1|\Big].$$

By Eq. (17),

$$L = n_1 + |\widehat{K}_1 \setminus K_1| \qquad \text{almost surely,}$$

and consequently

$$\mathbb{E}[L] = n_1 + (N_1 - n_1)\beta_\tau. \tag{18}$$

We now bound $H(K_1 \mid W, \Theta)$ by counting the n_1-subsets compatible with the accepted set. Since K_1 is independent of Θ and is uniform over all n_1-subsets of $[N_1]$,

$$H(K_1 \mid \Theta) = \log\binom{N_1}{n_1}.$$

Moreover, conditional on (W, Θ), the accepted set $\widehat{K}_1$ is known. By Eq. (17), K_1 must be an n_1-subset of $\widehat{K}_1$. Thus,

$$H(K_1 \mid W, \Theta) \le \mathbb{E}\left[\log \binom{L}{n_1}\right].$$

Using Eq. (16), we obtain

$$\begin{aligned} I(W; B \mid \Theta) &\ge \log \binom{N_1}{n_1} - \mathbb{E}\left[\log \binom{L}{n_1}\right] \\ &= \mathbb{E}\left[\log \frac{\binom{N_1}{n_1}}{\binom{L}{n_1}}\right]. \end{aligned} \tag{19}$$

For every $n_1 \le L \le N_1$,

$$\frac{\binom{N_1}{n_1}}{\binom{L}{n_1}} = \prod_{j=0}^{n_1-1} \frac{N_1 - j}{L - j} \ge \left(\frac{N_1}{L}\right)^{n_1}. \tag{20}$$

Therefore,

$$\begin{aligned} I(W; B \mid \Theta) &\ge n_1 \, \mathbb{E}\left[\log \frac{N_1}{L}\right] \\ &\ge n_1 \log \frac{N_1}{\mathbb{E}[L]} \\ &= n_1 \log \frac{1}{\rho_1 + (1-\rho_1)\beta_\tau}, \end{aligned} \tag{21}$$

where the second inequality follows from Jensen's inequality and the last equality follows from Eq. (18).

Combining Eq. (15) and Eq. (21) gives

$$n_1 \log \frac{1}{\rho_1 + (1-\rho_1)\beta_\tau} \le C - C_{\mathrm{rule}} - C_{\mathrm{gen}}.$$

Equivalently,

$$\rho_1 + (1-\rho_1)\beta_\tau \ge \exp\left(-\frac{C - C_{\mathrm{rule}} - C_{\mathrm{gen}}}{n_1}\right),$$

and hence the following finite-density bound holds:

$$\beta_\tau \ge \max\left\{0, \; \frac{\exp\left(-\frac{C - C_{\mathrm{rule}} - C_{\mathrm{gen}}}{n_1}\right) - \rho_1}{1 - \rho_1}\right\}. \tag{22}$$

Finally, taking the sparse atomic-fact limit $\rho_1 \to 0$ in Eq. (22) yields

$$\beta_\tau \ge \exp\left(-\frac{C - C_{\mathrm{rule}} - C_{\mathrm{gen}}}{n_1}\right),$$

which proves the result. □

Theorem 4.2. *Fix $\theta > 0$ and condition on $\Theta = \theta$. Let $\sigma_{\mathrm{c}} := \sqrt{\sigma_B^2 + \sigma_\theta^2}$. Suppose that $n_1, n_2, n \to \infty$, $a_N \to a \in [0,1]$, $N\delta_N \to 0$, and that $W_N - \widehat{W}_N \xrightarrow{\mathbb{P}} 0$. Then $W_N \xrightarrow{\mathbb{P}} a\theta$. Moreover, for every fixed $r \ge 0$,*

$$\alpha_N^{\mathrm{ato}}(r) \xrightarrow{\mathbb{P}} \alpha^{\mathrm{ato}}(r; a) := \Phi\left(\frac{r - a\theta}{\sigma_B}\right) + \Phi\left(\frac{r + a\theta}{\sigma_B}\right) - 1, \tag{5}$$

$$\alpha_N^{\mathrm{comp}}(r) \xrightarrow{\mathbb{P}} \alpha^{\mathrm{comp}}(r; a) := \Phi\left(\frac{r - (1-a)\theta}{\sigma_{\mathrm{c}}}\right) + \Phi\left(\frac{r + (1-a)\theta}{\sigma_{\mathrm{c}}}\right) - 1, \tag{6}$$

$$\beta_N(r) \xrightarrow{\mathbb{P}} \beta(r) := 2\Phi(r) - 1. \tag{7}$$

Proof of Theorem 4.2. Throughout the proof, all probabilities and expectations are conditional on $\Theta = \theta > 0$, unless specified otherwise. In particular, σ_c is fixed. We assume $\sigma_B > 0$ and use the independence of K from (Θ, B, X).

Convergence of the learned parameter. Let

$$\overline{Z}_N \coloneqq \frac{1}{N}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\widetilde{Z}_{ij}, \qquad \overline{Z}_{K,N} \coloneqq \frac{1}{n}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K_{ij}\widetilde{Z}_{ij}$$

be the averages of the full and selected composite pools. Since $\widetilde{Z}_{ij} = B_i + X_j$, the laws of large numbers give

$$\overline{Z}_N = \frac{1}{n_1}\sum_{i=1}^{n_1}B_i + \frac{1}{n_2}\sum_{j=1}^{n_2}X_j \xrightarrow{\mathbb{P}} \theta.$$

Conditional on (B, X), the selected entries form a simple random sample without replacement from the N composite values. Hence, for $N > 1$,

$$\mathbb{E}[\overline{Z}_{K,N} \mid B, X] = \overline{Z}_N, \qquad \mathrm{Var}(\overline{Z}_{K,N} \mid B, X) = \frac{1 - n/N}{n(N-1)}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\widetilde{Z}_{ij} - \overline{Z}_N)^2.$$

The empirical variance of the full Cartesian pool is the sum of the empirical variances of the atomic values and the rule instances: the cross term vanishes when summed over all (i, j). Taking expectations,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i,j}(\widetilde{Z}_{ij} - \overline{Z}_N)^2\right] = \left(1 - \frac{1}{n_1}\right)\sigma_B^2 + \left(1 - \frac{1}{n_2}\right)\sigma_\theta^2 \le \sigma_c^2.$$

Consequently,

$$\mathbb{E}[(\overline{Z}_{K,N} - \overline{Z}_N)^2] \le \frac{N\sigma_c^2}{n(N-1)} \longrightarrow 0.$$

Thus $\overline{Z}_{K,N} \xrightarrow{\mathbb{P}} \theta$. It follows that

$$\widehat{W}_N = (1 - a_N)\frac{1}{n_1}\sum_{i=1}^{n_1}B_i + a_N\overline{Z}_{K,N} \xrightarrow{\mathbb{P}} a\theta.$$

The assumption $W_N - \widehat{W}_N \xrightarrow{\mathbb{P}} 0$ now yields $W_N \xrightarrow{\mathbb{P}} a\theta$. No independence or distributional assumption on $W_N - \widehat{W}_N$ is needed.

Convergence of the two recalls. Define the empirical distribution functions of the atomic values, rule instances, and full composite pool, respectively, by

$$F_N^{\mathrm{ato}}(x) \coloneqq \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{1}\{B_i \le x\}, \qquad F_{X,N}(x) \coloneqq \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{1}\{X_j \le x\},$$

$$F_N^{\mathrm{comp}}(x) \coloneqq \frac{1}{N}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbf{1}\{\widetilde{Z}_{ij} \le x\}.$$

The Glivenko–Cantelli theorem applied separately to the two marginal samples gives

$$\sup_x \left|F_N^{\mathrm{ato}}(x) - \Phi(x/\sigma_B)\right| \xrightarrow{\mathbb{P}} 0, \qquad \sup_x |F_{X,N}(x) - \Phi((x-\theta)/\sigma_\theta)| \xrightarrow{\mathbb{P}} 0.$$

These conclusions hold jointly by the union bound and do not require independence between the two samples.

The Cartesian structure of the composite pool gives

$$F_N^{\mathrm{comp}}(x) = \frac{1}{n_2}\sum_{j=1}^{n_2}F_N^{\mathrm{ato}}(x - X_j).$$

Replacing F_N^{ato} by its Gaussian limit in this average incurs an error bounded by $\sup_y |F_N^{\mathrm{ato}}(y) - \Phi(y/\sigma_B)|$. For the resulting average, the identity

$$\frac{1}{n_2}\sum_{j=1}^{n_2}\Phi\left(\frac{x-X_j}{\sigma_B}\right) = \int_{\mathbb{R}} F_{X,N}(x-b)\frac{1}{\sigma_B}\phi(b/\sigma_B)\,db$$

and the Gaussian convolution formula yield

$$\sup_x \left|F_N^{\mathrm{comp}}(x) - \Phi\left(\frac{x-\theta}{\sigma_{\mathrm{c}}}\right)\right| \le \sup_x \left|F_N^{\mathrm{ato}}(x) - \Phi(x/\sigma_B)\right| + \sup_x |F_{X,N}(x) - \Phi((x-\theta)/\sigma_\theta)| \xrightarrow{\mathbb{P}} 0.$$

Thus the empirical distribution of the composite pool converges to $\mathcal{N}(\theta, \sigma_{\mathrm{c}}^2)$ even without assuming that B and X are independent given Θ.

To evaluate recall at the data-dependent center, let F_N denote either F_N^{ato} or F_N^{comp}, and let F denote its continuous Gaussian limit established above. The corresponding empirical proportion in $[W_N - r, W_N + r]$ equals $F_N(W_N + r) - F_N((W_N - r)^-)$, where $F_N(t^-)$ is the left limit at t. Continuity of F gives the pathwise bound

$$\left|F_N(W_N + r) - F_N((W_N - r)^-) - \big(F(W_N + r) - F(W_N - r)\big)\right| \le 2\sup_x |F_N(x) - F(x)|.$$

Combining uniform convergence with $W_N \xrightarrow{\mathbb{P}} a\theta$ therefore proves, for every fixed $r \ge 0$,

$$\alpha_N^{\mathrm{ato}}(r) \xrightarrow{\mathbb{P}} \Phi\left(\frac{a\theta + r}{\sigma_B}\right) - \Phi\left(\frac{a\theta - r}{\sigma_B}\right) = \alpha^{\mathrm{ato}}(r; a),$$
$$\alpha_N^{\mathrm{comp}}(r) \xrightarrow{\mathbb{P}} \Phi\left(\frac{a\theta + r - \theta}{\sigma_{\mathrm{c}}}\right) - \Phi\left(\frac{a\theta - r - \theta}{\sigma_{\mathrm{c}}}\right) = \alpha^{\mathrm{comp}}(r; a).$$

The final identities use $\Phi(-x) = 1 - \Phi(x)$ and also cover $r = 0$. No independence between W_N and the facts is used. Since recall is computed over the full composite pool, no condition on $N - n$ is needed.

Convergence of the false-positive rate. For this paragraph, probabilities involving Z refer only to the fresh draw $Z \sim \mathcal{N}(W_N, 1)$ with (B, X, W_N) held fixed, as in the definition of β_N. Define $\varepsilon_N := \mathbb{P}(Z \in \mathcal{T}_N)$. The total length of $\mathcal{T}_N$ is at most $2(n_1 + N)\delta_N$, and the density of Z is everywhere at most $1/\sqrt{2\pi}$. Since $n_1 \le N$, it follows that

$$0 \le \varepsilon_N \le \frac{2(n_1 + N)\delta_N}{\sqrt{2\pi}} \le \frac{4N\delta_N}{\sqrt{2\pi}} \longrightarrow 0.$$

Moreover, $\mathbb{P}(|Z - W_N| \le r) = 2\Phi(r) - 1 = \beta(r)$ for every realized W_N. For all sufficiently large N, the definition of β_N therefore gives

$$|\beta_N(r) - \beta(r)| = \frac{\left|\beta(r)\varepsilon_N - \mathbb{P}\big(|Z - W_N| \le r,\ Z \in \mathcal{T}_N\big)\right|}{1 - \varepsilon_N} \le \frac{\varepsilon_N}{1 - \varepsilon_N} \longrightarrow 0.$$

The bound is uniform over the realized facts and learned parameter. It proves the stated convergence in probability and completes the proof. □

Corollary 4.1. *Under the conditions of Theorem 4.2, fix a target recall $\alpha_0 \in (1/2, 1)$. For $a \in [0, 1]$, define $r_g(a) := \inf\{r \ge 0 : \alpha^g(r; a) \ge \alpha_0\}$ for $g \in \{\mathrm{ato}, \mathrm{comp}\}$, the smallest radius at which the limiting recall of population g reaches α_0. Let $r_{\alpha_0}(a) := \max\{r_{\mathrm{ato}}(a), r_{\mathrm{comp}}(a)\}$, and let $\beta_{\alpha_0}(a) := \beta(r_{\alpha_0}(a))$ be its limiting false-positive rate. Then*

$$\min_{a\in[0,1]} \beta_{\alpha_0}(a) \ge 2\Phi\left(\frac{\theta}{2} + \frac{\sigma_B + \sigma_{\mathrm{c}}}{2}\Phi^{-1}(\alpha_0)\right) - 1. \tag{8}$$

Proof of Corollary 4.1. For each $a \in [0,1]$ and $g \in \{\mathrm{ato}, \mathrm{comp}\}$, the function $r \mapsto \alpha^g(r;a)$ is continuous and strictly increasing from 0 to 1. Consequently, $r_g(a)$ is finite and is the unique radius satisfying $\alpha^g(r_g(a);a) = \alpha_0$. In particular, $r_{\alpha_0}(a)$ satisfies both recall constraints. The recall formulas are continuously differentiable and their derivatives with respect to r are positive, so the implicit function theorem also makes $r_g(a)$ continuous in a. Thus $\beta_{\alpha_0}(a)$ is continuous on $[0,1]$ and attains its minimum there.

Fix any $a \in [0,1]$ and any radius r satisfying both recall constraints. Since $\Phi(x) \le 1$, the limiting recall formulas imply

$$\alpha_0 \le \alpha^{\mathrm{ato}}(r;a) \le \Phi\left(\frac{r - a\theta}{\sigma_B}\right), \qquad \alpha_0 \le \alpha^{\mathrm{comp}}(r;a) \le \Phi\left(\frac{r - (1-a)\theta}{\sigma_{\mathrm{c}}}\right).$$

Applying the increasing function Φ^{-1} yields

$$r \ge a\theta + \sigma_B \Phi^{-1}(\alpha_0), \qquad r \ge (1-a)\theta + \sigma_{\mathrm{c}} \Phi^{-1}(\alpha_0).$$

Adding these inequalities and dividing by two eliminates a:

$$r \ge \frac{\theta}{2} + \frac{\sigma_B + \sigma_{\mathrm{c}}}{2} \Phi^{-1}(\alpha_0).$$

Apply this bound to $r = r_{\alpha_0}(a)$. Because $\beta(r) = 2\Phi(r) - 1$ is increasing, every $a \in [0,1]$ satisfies

$$\beta_{\alpha_0}(a) \ge 2\Phi\left(\frac{\theta}{2} + \frac{\sigma_B + \sigma_{\mathrm{c}}}{2} \Phi^{-1}(\alpha_0)\right) - 1.$$

Taking the minimum over a proves Eq. (8). □