Innovation: An Almost Characterization of Hallucination

Nishant Das [1] **Piyush Srivastava** [1]

Abstract

Hallucination is a central limitation of large language models (LLMs), and substantial effort has been devoted to understanding and mitigating it. Towards this, Kalai and Vempala (STOC 2024) introduced a probabilistic framework formalizing calibration and hallucination, and showed that, with high probability, calibrated LLMs hallucinate roughly at the rate of the "missing mass", a measure of how incomplete the training data is relative to its source. This raises two fundamental questions: (i) what property of a calibrated LLM makes hallucinations unavoidable? and (ii) can hallucinations be avoided by giving up calibration? We answer these questions by introducing a simpler property we call *innovation* that measures the tendency of a model to produce outputs outside the training data. We show that innovation is implied by the condition for hallucination identified by Kalai and Vempala, and, further, that it is an almost characterization of hallucination: hallucination implies innovation, and conversely, innovation implies hallucination with high probability. We also provide lower bounds on the hallucination rate based on the "innovation rate", and by relating innovation rate back to missing mass, we obtain new hallucination rate lower bounds based on missing mass that extend the results of Kalai and Vempala.

1. Introduction

One of the principal limitations of Large Language Models (LLMs) is the phenomenon of *hallucination* (Huang et al., 2025): the model produces statements that sound plausible but that are factually incorrect or logically inconsistent. This behavior persists even in state-of-the-art systems and thus remains a central challenge in the theory and practice of machine learning. Substantial work has therefore been devoted to mitigating the effect of hallucinations in practice, and we discuss some of this later in the paper.

Another line of work however, has sought to understand hallucination abstractly as a statistical phenomenon. The approach in this line of work is to abstract away the details of the model architecture, and to view LLMs as elements of a general class of (possibly randomized) procedures. To be useful, such an abstract framework must be rich enough to model hallucination and preferably other statistical properties of LLMs as well. One can then hope to establish general relationships between such statistical properties. The advantage of such abstract frameworks is that if the framework is also realistic enough, then the relationships obtained can be expected to hold quite independently of the underlying architecture or learning procedures, and could potentially apply also to architectures and learning procedures that may be discovered in the future.

An important probabilistic framework for studying hallucination is due to Kalai & Vempala (2024). In their framework, the "world" is modeled as a *meta-distribution* D_{world} over document distributions. The "true" document distribution D_L is drawn from this meta distribution D_{world}, and the support of this "true" document distribution D_L represents the "facts". A *corpus* is generated by sampling repeatedly from the "true" document distribution. A language model (more precisely, the training process of a language model) is abstracted as an algorithm that receives such a corpus and outputs a distribution over statements. A model is said to be *calibrated* if its output distribution approximately matches the "true" document distribution under a notion of distance formalized by Kalai and Vempala (see Definition 2.7 below). A key quantity in their framework is the *missing mass*, which is the probability mass, under the "true" document distribution, of the set of statements outside the corpus. Under some natural regularity assumptions on the meta-distribution, the main result of Kalai and Vempala is that "*calibrated language models must hallucinate*," roughly at the rate of the missing mass. This *rate* result also implies an *existence* result: it shows that if a calibrated model faces sufficiently large missing mass, hallucination is inevitable.

Contributions The above results lead to a few natural questions: what is the simplest natural property of an LLM that makes hallucination unavoidable? For example, can a model avoid hallucinating by not being calibrated?

[1]School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, Maharashtra - 400 005, India. Correspondence to: Nishant Das <nishant.das@tifr.res.in>, Piyush Srivastava <piyush.srivastava@tifr.res.in>.

Preprint. March 17, 2026.

We first give qualitative answers to these questions through an intermediate property that we call *innovation* (see Definition 3.1 for a formal definition). In comparison to the sophisticated formalization of calibration developed by Kalai & Vempala (2024) (see Definition 2.7 below), this is a very simple quantity: it just measures the probability assigned by the model to statements not observed in the training data. Our main qualitative result is that this simple quantity, however, lies at the heart of the hallucination phenomenon in the Kalai-Vempala framework. At an informal high level, under the same regularity assumption as in the work of Kalai and Vempala, we refine their existence result

$$\begin{aligned}&\text{Calibration + significant missing mass}\\ &\overset{\text{w.h.p.}}{\Longrightarrow} \text{Hallucination}\end{aligned}$$

into a finer chain of implications:

$$\begin{aligned}&\text{Calibration + significant missing mass}\\ \Longrightarrow\ &\text{Innovation} \overset{\text{w.h.p}}{\Longrightarrow} \text{Hallucination} \qquad (1)\\ \Longrightarrow\ &\text{Innovation.}\end{aligned}$$

We re-emphasize two important points made in the above diagram. First, our simpler condition of innovation holds whenever the condition of calibration combined with significant missing mass used by Kalai & Vempala (2024) holds: it is therefore weaker than their condition. Second, our condition essentially characterizes hallucination: our results show that in the Kalai-Vempala framework with their regularity assumptions, *innovation and hallucination are two sides of the same coin*. Section 3 presents our main qualitative results, showing how innovation is an almost characterization of hallucination.

We then turn, in Section 4, to a more quantitative exploration of the notion of innovation: our contribution here concerns the *rate* of hallucination once innovation occurs. While Kalai and Vempala express the hallucination rate in terms of the missing mass which is a property of the "true" document distribution and the training data, we derive two lower bounds: a *Markov-style* bound and a *high-confidence* bound, in terms of the model's *innovation rate* which is a property only of the model and the training data. Under their baseline assumptions, Kalai and Vempala obtain hallucination-rate bounds that depend explicitly on the corpus size n and degrade as n grows. This leaves open the possibility that hallucination might be eliminated given a sufficiently large corpus. They remove this dependence only by imposing an additional regularity assumption (Assumption 2.5 described below). In contrast, we remove the dependence on n without this additional assumption. Our results show that the innovation rate fundamentally governs the hallucination rate and that the possibility suggested above is untrue: increasing the amount of training data alone cannot eliminate hallucination once innovation occurs.

Finally, we show how to relate innovation rate back to missing mass. Combining this relation with our innovation-based bounds yields new missing-mass lower bounds on hallucination. These results recover the qualitative conclusions of Kalai and Vempala under weaker assumptions, and continue to hold in regimes where their original bounds become vacuous.

After a brief discussion of related work, we then develop in Section 2 the elements of the Kalai-Vempala framework relevant to our work.

Related Work Hallucination has been extensively studied empirically: benchmarks such as TruthfulQA attempt to quantify factual errors in LLM outputs (Lin et al., 2022), and practical approaches for mitigation have explored methods such as retrieval augmentation, grounding, and uncertainty estimation (Lewis et al., 2020; Nakano et al., 2022; Farquhar et al., 2024). We refer to Huang et al. (2025) for a survey of causes and mitigation strategies.

On the theoretical side, Kleinberg & Mullainathan (2024) asked the following question: given only positive examples from an unknown language, and no feedback on errors, is it possible to *eventually* generate *new* strings that all belong to the target language? While classical results in the theory of language identification (Gold, 1967) show that *identifying* the unknown language is impossible under positive data alone, Kleinberg and Mullainathan demonstrate that *generation* is nonetheless possible. This framework has been extended by various subsequent works, e.g., by Charikar & Pabbaraju (2025) (who also consider modifications of the model in which hallucination becomes inevitable), Raman et al. (2025), Raman & Raman (2025), Kleinberg & Wei (2025), and Kalavasis et al. (2025). Computability theoretic formalizations of hallucination have also been analyzed (Xu et al., 2025; Suzuki et al., 2025).

Our work builds upon the more statistical formalization of hallucination by Kalai & Vempala (2024), who showed that, in their framework, calibration implies hallucination. Wu et al. (2025) develop a learning theoretic framework in which they demonstrate the inevitability of hallucination without requiring calibration as a precondition. Kalai et al. (2025) further generalize the original Kalai-Vempala framework to take into account phenomena such as pretraining. In contrast to these works, the question we ask in this paper is: what is the simplest condition that already implies hallucination in the original Kalai-Vempala framework? As discussed above, and justified formally in the rest of this paper, our results show that the simple property of *innovation* provides an answer to this question.

2. The Kalai-Vempala Framework

In this section, we introduce the Kalai-Vempala framework. Let Ω be a finite set of *statements*. In practice, Ω could represent all possible finite sequences of words from the vocabulary of the LLM of length up to the context length of the LLM. The Kalai-Vempala framework however works in the *unprompted generation* setting where the prompt is fixed to be the empty string (equivalently, one may view the prompt as fixed throughout). In this setting, the training data consists solely of responses, and during inference the model generates a response without an explicit prompt.

Definition 2.1 (Language Model)**.** A *language model* is a mapping $\mathcal{A} : \mathcal{M}(\Omega) \to \Delta(\Omega)$, where $\mathcal{M}(\Omega)$ is the set of finite multisets over Ω. Thus, given the training data $X \in \mathcal{M}(\Omega)$, the model outputs a predictive distribution $g := \mathcal{A}(X) \in \Delta(\Omega)$.

The Kalai-Vempala framework captures LLMs in the following manner: Ω is the set of strings of length up to c (the *context length* of the LLM) over a finite set $\mathcal{V}$ (the *vocabulary* of the LLM). As mentioned above, the Kalai-Vempala framework works in the unprompted generation setting. The autoregressive generating procedure of LLMs then induces a well-defined probability distribution over Ω. Hence, LLMs fit into the above abstract notion of Language Models.

We now motivate and formalize the notions of *document distributions* and *corpuses* in the framework of Kalai and Vempala. Large language models are trained on large-scale datasets collected from a variety of sources such as Wikipedia, Reddit, news articles, books, blogs, and scientific papers. Different sources contain different statements, and the same statement may appear with different frequencies across sources; e.g., technical statements may be more common in scientific articles than in Reddit, while personal statements may appear more frequently in the latter.

Each such source (or any fixed combination of sources) defines a *document distribution*: this is a probability distribution over the set Ω of all statements, and represents which statements appear in the source and how often they occur. The support of a document distribution defines the set of *facts* in the document distribution, while the complement of the support is the set of *hallucinations*. This means that in the Kalai-Vempala framework, there is no *semantic* notion of truth: a fact is defined simply by belonging to the support of the document distribution.

A *corpus* is generated by repeatedly sampling from the document distribution. Given a corpus from a particular document distribution, a language model trained on it is said to *hallucinate* (according to that document distribution) if it places positive probability mass on the set of *hallucinations* of that document distribution. For instance, for a model trained on a corpus sampled from Wikipedia, all statements that occur in Wikipedia are taken to be factual within the Wikipedia document distribution. Thus, in the Kalai-Vempala framework, a model trained on a corpus drawn from Wikipedia is said to hallucinate if and only if it produces statements that are absent in Wikipedia.

Note that while the corpus itself is observed, the underlying document distribution that produced it is not. Indeed, many distinct document distributions may give rise to the same observed corpus. As a result, from the corpus alone, one cannot comment about how much the model hallucinates. To reason about hallucination in this setting, Kalai and Vempala introduce a *meta-distribution* D_{world} over document distributions, capturing uncertainty about which document distribution generated the observed corpus. Once a distribution over document distributions is fixed, we can reason probabilistically via the posterior over document distributions given the corpus and make probabilistic statements about hallucination across document distributions.

Remark 2.2. The unprompted generation setting and the assumption that all observed statements are true are both significant simplifications. Nevertheless, both our results and those of Kalai & Vempala (2024) already hold under these assumptions. In more realistic settings, the conclusions would at the very least remain valid, and may in fact become strictly stronger. For one approach to extending this framework to the prompted setting, see Kalai et al. (2025).

Notation We now review the formal notation, largely following Kalai & Vempala (2024), for the above notions. As before, Ω denotes a finite (but large) set of statements. For any set S, we denote by $\Delta(S)$ the set of probability distributions over S. A *document distribution* is a probability distribution over Ω (i.e. an element of $\Delta(\Omega)$). We denote by $D_{\text{world}} \in \Delta(\Delta(\Omega))$ the distribution over the document distributions. Each document distribution p in $\text{supp}(D_{\text{world}})$ determines its set of *facts* $F = F(p) := \text{supp}(p)$ and its set of *hallucinations* $H = H(p) := \Omega \setminus F$. (The symbols F and H are chosen for consistency with (Kalai & Vempala, 2024).) A corpus X is generated by first drawing p^* from D_{world} and then $X = (x_1, \ldots, x_n) \sim (p^*)^{\times n}$, where n is the size of the corpus. Recall that p^* is unknown (to the model trainer). $O := \text{set}(X)$ is the set of *observed statements* (differing from X only in de-duplication of statements that appear multiple times in X) and $U := \Omega \setminus O$ are the *unobserved statements*. $p(U)$ thus denotes the *missing mass* for a document distribution p and corpus X. A Language Model $\mathcal{A}$ trained on a corpus X outputs a predictive distribution $g = \mathcal{A}(X) \in \Delta(\Omega)$ (again, the symbol g is chosen for consistency with Kalai & Vempala (2024)). Hence, g represents a trained language model. Our central quantity of interest is $g(H)$ which denotes the *rate of hallucination*.

2.1. Regularity Assumptions

We now present the regularity assumptions imposed on the meta-distribution D_{world} in the Kalai-Vempala framework. The first of these captures the idea that semantic truth is rare relative to the space of all well-formed statements. Kalai & Vempala (2024) formalize this by assuming that each document distribution $p \sim D_{\text{world}}$ contains only finitely many facts, and that number is small relative to the full statement space Ω.

Assumption 2.3 (K-sparsity: Assumption 1 in Kalai & Vempala (2024))**.** There exists K such that every document distribution $p \in \text{supp}(D_{\text{world}})$ satisfies $|F(p)| \leq K$, with $K/|\Omega| \ll 1$.

We note that Kalai & Vempala (2024) use a slightly different vut equivalent parameterization: they say that the world is s-sparse if $|F(p)| \leq \exp(-s)|H(p)|$ for every p in its support. The two parameterizations are equivalent via $K = |\Omega|/(1+\exp(s))$. Further, this assumption also implies that $K/|U| \leq e^{-s}$ for every corpus generated from a document distribution sampling from D_{world}.

The second regularity assumption in the Kalai-Vempala framework formalizes the idea that outside the observed corpus the model has no reliable signal to different facts ("true" statements) from hallucinations ("false" statements). This goes back to the idea that in this framework, there is no *semantic* notion of truth.

Assumption 2.4 (Definition 3: Regular Facts in Kalai & Vempala (2024))**.** For every corpus X and posterior $\nu := D_{\text{world}}(\cdot \mid X)$, all unobserved statements $y, y' \in U$ satisfy

$$\Pr_{p \sim D_{\text{world}}}[y \in F \mid X] = \Pr_{p \sim D_{\text{world}}}[y' \in F \mid X],$$

where F denotes $\text{supp}(p)$. Thus, conditioned on the observed data, all unseen statements are equally likely to be factual in the true document distribution.

While our results do not rely on the following assumption, for completeness and to facilitate discussions, we describe an additional assumption under which Kalai and Vempala were able to strengthen their results.

Assumption 2.5 (Definition 4: Regular Probabilities in Kalai & Vempala (2024))**.** For every corpus X and posterior $\nu := D_{\text{world}}(\cdot \mid X)$, all unobserved statements $y, y' \in U$ satisfy

$$\mathbb{E}_{p \sim D_{\text{world}}}[p(y) \mid X] = \mathbb{E}_{p \sim D_{\text{world}}}[p(y') \mid X].$$

(Mis)calibration We now describe the final ingredient of the Kalai-Vempala framework: a formalization of the notion of calibration. A *partition* of Ω is a collection $\Pi = \{B_1, \ldots, B_m\}$ of disjoint nonempty subsets of Ω whose union is Ω. Each element of a partition is called a *cell*. We write $\mathcal{P}(\Omega)$ for the set of all such partitions.

Definition 2.6 (Definition 1: Calibration and Coarsening in Kalai & Vempala (2024))**.** Let $p \in \Delta(\Omega)$ and $\Pi \in \mathcal{P}(\Omega)$. Let B_y denote the unique cell containing the statement $y \in \Omega$. The Π-*coarsening* of p is the distribution $p^\Pi \in \Delta(\Omega)$ defined by $p^\Pi(y) := \frac{p(B_y)}{|B_y|}$. Thus p^Π averages p within each cell of Π and redistributes that mass uniformly inside the cell. Given $p, g \in \Delta(\Omega)$, we say that g is *calibrated* to p if there exists $\Pi \in \mathcal{P}(\Omega)$ such that $g = p^\Pi$.

Notice that g is calibrated to p *if and only if* g is equal to the coarsening of p induced by the partition induced by the level sets of g. This leads to a natural definition for Miscalibration.

Definition 2.7 (Miscalibration: eq. (2) in Kalai & Vempala (2024))**.** Given $g \in \Delta(\Omega)$, let $\mathcal{B}_g$ be the partition of Ω into level sets of g: $\mathcal{B}_g := \{B \subseteq \Omega : \exists \alpha \text{ with } B = \{y \in \Omega : g(y) = \alpha\},\ B \neq \emptyset\}$. The *miscalibration* of g relative to p is defined as

$$\text{Mis}(g, p) := \left\| g - p^{\mathcal{B}_g} \right\|_{\text{TV}} = \max_{S \subseteq \Omega} |g(S) - p^{\mathcal{B}_g}(S)|.$$

Hence, $\text{Mis}(g, p) = 0$ if and only if g is calibrated to p.

Calibration Implies Hallucination We are now ready to state the main results of Kalai and Vempala. The main result in their paper is their Theorem 1; however, this is a technical result that is not directly interpretable. Kalai and Vempala therefore present more interpretable consequences of this main theorem as corollaries, which we state below.

The first of these corollaries (Corollary 2 in their paper) can be directly compared to our results to be stated later, since it operates under the same assumptions on the meta distribution as our results.[1]

Corollary 2.8 (Corollary 2 in Kalai & Vempala (2024))**.** *Let D_{world} be a K-Sparse (Assumption 2.3) world satisfying Regular Facts (Assumption 2.4). Let $p^* \sim D_{\text{world}}$ be the true document distribution, $X \sim (p^*)^{\times n}$ the training data, and $g = \mathcal{A}(X)$ be the predictive distribution of a language model $\mathcal{A}$. Then for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ conditioned on the input corpus X, we have*

$$g(H) \ \geq \ p(U) \ - \ \text{Mis}\big(g, p\big) \ - \ \frac{K(n+1)}{\delta\,|U|}.$$

Note the dependence on n, the size of the corpus: the bound becomes vacuous once n becomes greater than $U/|K| - 1$. This might lead one to hope that perhaps hallucinations can be avoided by collecting a large corpus. By assuming

[1] Note that the statements of Corollaries 1 and 2 of Kalai & Vempala (2024) have an additional error term since those bounds are stated in terms of an estimator for $p(U)$ rather than in terms of $p(U)$ directly. However, the bounds without that error term presented here are implied by their proof.

the Regular Probabilities assumption (Assumption 2.5 cited above), Kalai and Vempala arrive at the following corollary which removes the dependence on n.

Corollary 2.9 (Corollary 1 in Kalai & Vempala (2024))**.** *Let D_{world} be a K-Sparse (Assumption 2.3) distribution with Regular Facts (Assumption 2.4) and Regular Probabilities (Assumption 2.5). Let $p^* \sim D_{\text{world}}$ be the true document distribution, $X \sim (p^*)^{\times n}$ the training data, and $g = \mathcal{A}(X)$ be the predictive distribution of a language model $\mathcal{A}$. Then for any $\delta \in (0, 1]$, with probability at least $1-\delta$ conditioned on the input corpus X, we have*

$$g(H) \ \geq \ p(U) \ - \ \text{Mis}\big(g, p\big) \ - \ \frac{2K}{\delta\,|U|}.$$

A Remark on Conditioning Note that the probability bounds in the the above two results are with respect to the conditional distribution $D_{\text{world}}(\cdot \mid X)$ of the document distribution p, conditioned on the observed training corpus X. Our results (e.g. Theorems 3.3, 4.1 and 4.2 and their corollaries) also derive probability bounds under the same conditioned distribution.

Relaxing the Regular Facts Assumption Kalai & Vempala (2024) also consider a weaker version of the Regular Facts assumption (Assumption 2.4), called the r-Regular Facts condition, under which the posterior probability that an unseen statement is factual is allowed to deviate from the posterior mean value $\mathbb{E}_{\nu}[|F \cap U|]/|U|$ by a factor of at most r. Both their results as well as ours can be extended to use this weakened setting (of course, at a cost depending upon the parameter r): we discuss this in Appendix A.

The above two results can thus be summarized informally as follows. In a K-sparse meta-distribution with Regular Facts, any language model that is calibrated ($\text{Mis}(g, p) = 0$) must hallucinate with the hallucination *rate* roughly equal to the *missing mass* $p(U)$, up to an error term of order $Kn/|U|$. If we additionally assume that the meta-distribution has Regular Probabilities, the error term collapses to $O(K/|U|)$.

Implications of the Kalai-Vempala Bounds These results (Corollary 2.9 and Corollary 2.8) yield two distinct implications about the behavior of calibrated language models.

- *Existence:* If the model is calibrated and the missing mass $p(U)$ exceeds the error term (of order $K/|U|$, or $Kn/|U|$ in the absence of Regular Probabilities), then hallucination is unavoidable.
- *Rate:* Under calibration, the hallucination rate $g(H)$ is approximately the missing mass $p(U)$.

This distinction between existence and rate will be important in what follows. In Section 3, we strengthen upon the existence result and *almost characterize* the existence of hallucination. In Section 4, we derive new lower bounds on the hallucination rate which complement the Kalai-Vempala bounds.

3. Innovation and Hallucination

We now describe the main conceptual contribution of this paper: the notion of *innovation*. Informally, a model innovates if it assigns nonzero probability to events or outcomes that it has never directly observed, i.e., it "extends" beyond its training support.

Definition 3.1 (**Innovation**)**.** Let $\mathcal{A}$ be a language model and $g := \mathcal{A}(X)$ its predictive distribution over Ω given input $X \subseteq \Omega$. Let $U := \Omega \setminus \text{set}(X)$ denote the set of unseen outcomes. We say that $\mathcal{A}$ *innovates* on X if

$$g(U) = \mathcal{A}(X)(U) > 0.$$

That is, the model assigns nonzero probability mass to outcomes not present in the corpus.

In this section, we focus on establishing the following qualitative results:

1. We prove a two-way connection between innovation and hallucination. It is easy to see that every hallucinating model necessarily innovates; we show in the other direction that every model that innovates must hallucinate with high probability. Thus, innovation provides an almost characterization of the *existence* of hallucination.
2. Kalai and Vempala showed that calibration together with sufficiently large missing mass forces hallucination. We show that calibration together with *any* positive missing mass implies innovation. This shows that the existence-level implication of Kalai-Vempala is a special case of our innovation-based characterization.

3.1. Proof of Characterization

We begin with the easy direction: hallucination is only possible if the model innovates.

Observation 3.2 (Hallucination implies Innovation)**.** *Let $\mathcal{A}$ be a Language Model and $g := \mathcal{A}(X)$ its predictive distribution over Ω given input $X \subseteq \Omega$. If $g(H) > 0$ for some world p, then $g(U) > 0$.*

The converse direction is subtler, because H, the set of hallucinations, itself is unknown. We show however that, under the K-Sparsity and Regular Facts assumptions, if the model innovates, then with high posterior probability it must hallucinate.

Theorem 3.3 (Innovation implies hallucination with high probability)**.** *Let D_{world} be a meta-distribution satisfying*

K-Sparsity (Assumption 2.3) and Regular Facts (Assumption 2.4). Let $\mathcal{A}$ be a Language Model and $g := \mathcal{A}(X)$ its predictive distribution over Ω given input $X \subseteq \Omega$. If $g(U) > 0$ then

$$\Pr_{p \sim D_{\text{world}}} \left[g(H) > 0 \;\middle|\; X \right] \;\geq\; 1 - \frac{K}{|U|}.$$

Proof. Since $g(U) > 0$, pick $y^\star \in U$ with $g(y^\star) > 0$. For any document distribution p, we have $H = \Omega \setminus F$, so $y^\star \in H$ if and only if $y^\star \notin F$. By Regular Facts, all $y \in U$ have the same probability $q := \Pr[y \in F \mid X]$. Hence

$$\mathbb{E}\big[\, |F \cap U| \mid X \big] = \sum_{y \in U} \Pr[y \in F \mid X] = |U|\, q. \tag{2}$$

By K-sparsity, $|F| \leq K$ for every document distribution, so $|F \cap U| \leq K$. Taking conditional expectation on both sides gives $|U|\, q \leq K$, so that $q \leq K/|U|$. Applying this to $y^\star$,

$$\Pr[y^\star \in H \mid X] = 1 - \Pr[y^\star \in F \mid X] \geq 1 - \tfrac{K}{|U|}.$$

Whenever $y^\star \in H$, we have $g(H) \geq g(y^\star) > 0$. Therefore

$$\Pr[g(H) > 0 \mid X] \geq \Pr[y^\star \in H \mid X] \geq 1 - \tfrac{K}{|U|}. \quad \square$$

Observation 3.2 and Theorem 3.3 together show that, in the Kalai-Vempala formalism, innovation is an almost characterization of hallucination. To compare this with the results of Kalai and Vempala, we now turn to the relationship between innovation and their notion of calibration.

3.2. Calibration, Missing Mass, and Innovation

Kalai and Vempala, at the existence level, show that if there is significant missing mass, any calibrated model must hallucinate. We now show that calibration together with any missing mass already implies innovation. The intuition is straightforward: calibration forces the model to assign positive probability to every factual statement. As soon as there is missing mass, at least one such statement lies outside the observed corpus, and the model must assign it positive probability. Hence the model necessarily innovates. We now formalize this.

Proposition 3.4 (Calibration and missing mass imply innovation)**.** *Let $p \in \Delta(\Omega)$ be the true document distribution, $X \sim p^{\times n}$ the training data, and $U := \Omega \setminus O$ the set of unobserved statements, where $O = \mathrm{set}(X)$. Let $\mathcal{A}$ be a language model and $g := \mathcal{A}(X)$ be a predictive distribution that is calibrated to p, i.e., there exists a partition Π of Ω with $g = p^{\Pi}$. If there is any missing mass, i.e. $p(U) > 0$, then $\mathcal{A}$ innovates on X.*

Combining Proposition 3.4 with our earlier result that innovation forces hallucination with high posterior probability under K-Sparsity and Regular Facts (Theorem 3.3), we obtain the refined logical chain described in eq. (1) in the introduction. This shows that, at the level of existence, our result strictly generalizes the Kalai-Vempala result.

A natural next question is how large the hallucination rate $g(H)$ must be once innovation occurs. In the next section we show that the *innovation rate* $g(U)$ governs the *rate* at which a model that innovates must hallucinate: a fixed fraction of $g(U)$ must fall on false statements with high probability.

4. Lower Bounds on Hallucination Rate via Innovation Rate

Having characterized the existence of hallucination, we now ask a quantitative question: *how much* must a model hallucinate once it innovates? In this section we derive lower bounds on the hallucination rate $g(H)$ in terms of the *innovation rate* $g(U)$. Then, we relate innovation rate to missing mass to obtain bounds that complement those of Kalai and Vempala.

4.1. A Markov-style Lower Bound on the Hallucination Rate

We now derive a lower bound on $g(H)$ in terms of the innovation mass $g(U)$ invoking a Markov-style argument.

Theorem 4.1 (Markov-style bound)**.** *Let D_{world} be a meta-distribution satisfying K-Sparsity and Regular Facts. Let $\mathcal{A}$ be a Language Model and $g := \mathcal{A}(X)$ its predictive distribution over Ω given input $X \in \mathcal{M}(\Omega)$. Then for any $\delta \in (K/|U|, 1)$,*

$$\Pr_{p \sim D_{\text{world}}} \left[g(H) \;\geq\; g(U) \left(1 - \frac{K}{\delta |U|} \right) \;\middle|\; X \right] \;\geq\; 1 - \delta.$$

Proof. We begin the proof by lower bounding the expected hallucination rate. Write $g(F \cap U) = \sum_{y \in U} g(y) \mathbf{1}[y \in F]$. By the Regular Facts assumption, there exists a q such that $q = \Pr[y \in F \mid X]$ for every $y \in U$. As in the proof of Theorem 3.3, K-Sparsity implies $q \leq K/|U|$. Taking conditional expectations, we thus get $\mathbb{E}[g(F \cap U) \mid X] = \sum_{y \in U} g(y) \Pr[y \in F \mid X] = \left(\sum_{y \in U} g(y)\right) \cdot q = q g(U) \leq g(U)\, K/|U|$. Since $g(H) = g(U) - g(F \cap U)$,

$$\mathbb{E}[g(H) \mid X] \geq g(U)\big(1 - K/|U|\big). \tag{3}$$

Let $t \in (0, 1)$ be a parameter to be chosen later. We now invoke a Markov-style argument. First, we split the expected hallucination rate $\mathbb{E}[g(H) \mid X]$ into two terms $\mathbb{E}[g(H)\mathbf{1}[g(H) \geq t g(U)] \mid X]$ and $\mathbb{E}[g(H)\mathbf{1}[g(H) < t g(U)] \mid X]$. Because $g(H) \leq g(U)$, the first term is upper bounded by $g(U)\mathbb{E}[\mathbf{1}[g(H) \geq t g(U)] \mid X]$ and the second term is upper bounded by $tg(U)\mathbb{E}[\mathbf{1}[g(H) < tg(U)] \mid X]$. Hence, we have

$$\begin{aligned}\mathbb{E}[g(H) \mid X] \leq& \alpha\, g(U) + (1-\alpha)\, t\, g(U) \\ =& g(U)\,(t + (1-t)\alpha),\end{aligned}$$

where $\alpha := \Pr[g(H) \geq t\, g(U) \mid X]$. Combining with the lower bound on $\mathbb{E}[g(H) \mid X]$ in eq. (3) gives $\alpha \geq \frac{1-K/|U|-t}{1-t}$. Substituting $t = 1 - \dfrac{K}{\delta|U|}$ in the above and using $K/|U| < \delta < 1$ now gives

$$\Pr\left[g(H) \geq g(U) \cdot \left(1 - \frac{K}{\delta|U|}\right) \mid X\right] = \alpha \geq 1 - \delta. \quad \square$$

4.2. A High-Confidence bound on Hallucination Rate

The Markov-style lower bound in Theorem 4.1 becomes vacuous as δ approaches $K/|U|$, since the multiplicative factor $1 - K/(\delta|U|)$ then tends to zero. Similarly, the Kalai-Vempala bounds in Corollary 2.9 and Corollary 2.8 becomes vacuous whenever $\delta < 2K/|U|$, because $p(U) \leq 1$ and $\|g - p^{\Pi}\|_{\mathrm{TV}} \geq 0$ imply that the right-hand side of those bounds is non-positive. To go beyond this limitation, we now present a different argument that yields a nontrivial lower bound on hallucination rate with probability at least $1 - K/|U|$.

The intuition is simple. Any model that innovates must inevitably assign probability to hallucination. If it spreads its probability broadly, sparsity defeats it: almost all unseen statements are hallucinations. If it concentrates sharply, regularity defeats it: having no basis to distinguish the unseen truths from falsehoods, the spike almost surely lands in the wrong place. Informally, every model must either scatter or spike and both strategies fail. We now formalize this.

Theorem 4.2 (High-Confidence bound)**.** *Let D_{world} be a meta-distribution satisfying K-Sparsity and Regular Facts. Let $g = \mathcal{A}(X)$ be any language model. Then*

$$\Pr_{p \sim D_{\mathrm{world}}}\left[g(H) \;\geq\; \frac{g(U)}{K+1} \;\middle|\; X\right] \;\geq\; 1 - \frac{K}{|U|}.$$

Proof. Let $m := \max_{y \in U} g(y)$ denote the largest mass assigned by g to any unseen statement, and fix $y^\star \in U$ satisfying $g(y^\star) = m$. Note that both m and $y^\star$ are random variables, but they are completely determined given the training corpus X (formally, they are measurable with respect to the σ-field generated by the random variable X). Note also that applying the same line of reasoning as in the proof of Theorem 3.3 on $y^\star$ gives $\Pr[y^\star \in H \mid X] \geq 1 - \frac{K}{|U|}$. We now analyze two cases.

Case 1: $m \leq g(U)/(K+1)$**.** In this case, using $|F \cap U| \leq K$ and K-Sparsity, we get $g(F \cap U) = \sum_{y \in F \cap U} g(y) \leq K \cdot m \leq K \cdot \frac{g(U)}{K+1}$, so that $g(H) = g(U) - g(F \cap U) \geq g(U) - K \cdot \frac{g(U)}{K+1} = \frac{g(U)}{K+1}$.

Case 2: $m > g(U)/(K+1)$**.** In this case $g(y^\star) > \frac{g(U)}{K+1}$, so that the event $y^* \in H$ implies the event $g(H) \geq g(U)/(K+1)$.

Let $\mathcal{E}$ denote the event $g(H) \geq g(U)/K+1$, and let $\mathcal{E}_1$ and $\mathcal{E}_2$ denote the events corresponding to the two cases: note that $\mathcal{E}_1$ and $\mathcal{E}_2$ are determined by X (since m is). Writing the above conclusions in terms of indicator random variables thus gives $\Pr[\mathcal{E}|X] = I_{\mathcal{E}_1} + I_{\mathcal{E}_2} \Pr[y^\star \in H|X]$. Since $\Pr[y^\star \in H \mid X] \geq 1 - \frac{K}{|U|}$, the claim follows by noting that $I_{\mathcal{E}_1} + I_{\mathcal{E}_2} = 1$. $\square$

We thus see that the only way to avoid hallucination entirely in this framework is to choose a model that never innovates, i.e., a model with $g(U) = 0$. Any attempt to generalize beyond the observed support inevitably incurs hallucination at a rate proportional to the model's innovation rate.

The Regular Probabilities Assumption We emphasize that both Theorem 4.1 and Theorem 4.2 rely only on K-Sparsity and Regular Facts, and do *not* assume Regular Probabilities (Assumption 2.5). In contrast, the Kalai–Vempala bound for meta-distributions satisfying K-Sparsity and Regular Facts alone (Corollary 2.8) incurs an error term of order $Kn/|U|$. As the corpus size grows (whenever $n \geq |U|/K$), this bound becomes vacuous, which might suggest that with sufficiently large training data one could drive the hallucination rate to zero under these assumptions. Corollary 1 of Kalai & Vempala (2024) (Corollary 2.9 above) shows that this intuition is false, at least under an additional Regular Probabilities assumption. Our results show that this intuition is false even without this assumption: even if the corpus size n is large, a model with a positive innovation rate $g(U)$ must hallucinate at a comparable rate.

4.3. From Innovation Rate to Missing Mass

We now show how to translate bounds expressed in terms of *innovation rate* $g(U)$, which is a property of the model, into bounds expressed in terms of the document distribution's *missing mass* $p(U)$, which is a property of the document distribution and the corpus X. This crucial translation also allows us to extend our results in a way that complements the bounds of Kalai and Vempala.

Proposition 4.3 (Missing mass lower-bounds innovation rate)**.** *Assume K-Sparsity. Let $p \sim D_{\mathrm{world}}$, $X \sim {p^*}^{\times n}$, and let $O := \mathrm{set}(X)$, $U := \Omega \setminus O$. Let Π be a partition of Ω and let p^{Π} denote the Π-coarsening of p as defined in Definition 2.6. Let $\mathcal{A}$ be a language model and $g := \mathcal{A}(X)$ be its predictive distribution. Then, we have*

$$g(U) \;\geq\; \frac{p(U)}{K+1} \;-\; \|g - p^{\Pi}\|_{\mathrm{TV}}.$$

Proposition 4.3 is the link we need. Combining it with Theorems 4.1 and 4.2, we get the following corollaries:

Corollary 4.4 (A missing-mass bound from the Markov-style bound)**.** *Under the assumptions of Theorem 4.1, for any $\delta \in (K/|U|, 1)$, with probability at least $1-\delta$ conditioned on the corpus X,*

$$g(H) \geq \frac{p(U)}{K+1} - \frac{1}{\delta|U|} - \|g - p^{\Pi}\|_{\mathrm{TV}}.$$

For $\Pi = \mathcal{B}_g$ (level sets of g), this becomes

$$\Pr\left[g(H) \geq \frac{p(U)}{K+1} - \frac{1}{\delta|U|} - \mathrm{Mis}(g,p) \;\middle|\; X\right] \;\geq\; 1-\delta.$$

Corollary 4.5 (A missing-mass bound from the high-confidence bound)**.** *Under the assumptions of Theorem 4.2, with probability at least $1 - \dfrac{K}{|U|}$ conditioned on the corpus X,*

$$g(H) \;\geq\; \frac{p(U)}{(K+1)^2} - \frac{\|g - p^{\Pi}\|_{\mathrm{TV}}}{K+1}.$$

For $\Pi = \mathcal{B}_g$ (level sets of g), this becomes

$$\Pr\left[g(H) \geq \frac{p(U)}{(K+1)^2} - \frac{\mathrm{Mis}(g,p)}{K+1} \;\middle|\; X\right] \;\geq\; 1 - \frac{K}{|U|}.$$

Comparison with the Kalai-Vempala Bounds These corollaries complement the Kalai-Vempala bounds in a precise sense. First, Corollary 4.5 yields a missing-mass lower bound on the hallucination rate in a regime for the error probability δ where both the Kalai-Vempala bounds and the Markov-style bounds collapse: In particular, the Kalai-Vempala bounds (in both Corollaries 2.8 and 2.9) become vacuous for $\delta < 2K/|U|$, while Corollary 4.5 continues to provide a nontrivial guarantee with error probability at most $\delta = K/|U|$.

Second, Corollary 4.4 gives a missing-mass lower bound in the high error probability regime under K-Sparsity and Regular Facts alone. Compared to the corresponding bound obtained by Kalai and Vempala (Corollary 2.8), this bound is weaker by a factor of $1/(K+1)$ in its dependence on the missing mass. However, compared to the Kalai-Vempala bounds, it has the advantage of having no dependence on the corpus size n. Consequently, it ensures a hallucination rate on the order of $p(U)/(K+1)$ even when n is comparable to $|U|/K$, a regime in which the Kalai-Vempala bound under K-Sparsity and Regular Facts (Corollary 2.8) alone is vacuous. Thus, compared to the Kalai-Vempala bound that additionally assumes Regular Probabilities (Corollary 2.9) in order to remove the dependence on corpus size, our Corollary 4.4 can be viewed as a qualitative recovery of their conclusion under strictly weaker assumptions. While the quantitative dependence on the missing mass is looser by a factor of $1/(K+1)$, the interpretation remains the same: any calibrated language model must hallucinate at a rate controlled by the missing mass.

5. Discussion

Lower Bounding the Hallucination Rate of LLMs A potential use of quantities such as calibration and innovation, and their connection to hallucination proved by Kleinberg & Mullainathan (2024) and in this paper, is that they might allow estimation of lower bounds on the hallucination rate of a model. For example, Kalai and Vempala provide a method to estimate the missing mass using the *Good-Turing Estimator* (Good, 1953), and show that the missing mass can be estimated from a proxy quantity which they call the "monofact rate" of the training data. If real systems satisfy the assumptions of the Kalai-Vempala framework, then by measuring the monofact rate of the training data, and then using the results of Kalai & Vempala (2024) (Corollaries 2.8 and 2.9 quoted above), one can therefore give a high probability lower bound on the hallucination rate of LLMs *if an LLM is calibrated in the sense of Definition 2.7*. However, whether an LLM is calibrated or not or how miscalibrated it is, is not something that is easy to estimate. Hence, the results of Kalai and Vempala, while theoretically insightful, cannot be directly used to give lower bounds on the hallucination rate of LLMs. We note, however, that the innovation based lower bound in Theorem 4.1 does not suffer from this problem, since the innovation rate can be estimated directly given access to a trained model and the training data. (The bound given by Theorem 4.1 formally depends upon $K/|U|$, which cannot, in principle, be estimated just from the training data and the model's output: this is not a problem under the sparsity assumption, which provides that $K/|U|$ must be very close to 0.)

Feedback and Hallucination One way to interpret our results is that hallucination can be seen as a consequence of innovation in the absence of semantic feedback. In practical settings, where model outputs can be evaluated and models can be retrained or adjusted based on this feedback, such feedback mechanisms may allow hallucination to be mitigated even while innovation persists.

6. Conclusion

The outstanding question in the theoretical modeling of hallucination is to design theoretical frameworks that can capture as many aspects of the phenomenon as possible. One way to test existing frameworks, such as the framework of Kalai & Vempala (2024) studied in the paper, is to probe the framework by asking what are the qualitatively strongest implications that are implied by it. Equivalently, the question is to identify the qualitatively *weakest* conditions under which the framework implies hallucination. Our work contributes to this line of work by identifying our notion of innovation as the weakest possible condition that implies hallucination in the Kalai-Vempala framework.

Acknowledgements

We gratefully acknowledge support from the Department of Atomic Energy, Government of India [project numbers RTI4001 and RTI4014]; by the Infosys-Chandrasekharan virtual center for Random Geometry at the Tata Institute of Fundamental Research; by the Science and Engineering Research Board [grant number MATRICS MTR/2023/001547]; and by a Google India Research Award. The contents of this paper do not necessarily reflect the views of the funding agencies listed above

References

Charikar, M. and Pabbaraju, C. Exploring facets of language generation in the limit. In *Proceedings of 38th Conference on Learning Theory (COLT)*, volume 291 of *PMLR*, pp. 854–887, July 2025.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, June 2024. doi: 10.1038/s41586-024-07421-0.

Gold, E. M. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967. doi: 10.1016/S0019-9958(67)91165-5.

Good, I. J. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4): 237–264, December 1953. doi: 10.1093/biomet/40.3-4.237.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, January 2025. doi: 10.1145/3703155.

Kalai, A. T. and Vempala, S. S. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 160–171. ACM, 2024. doi: 10.1145/3618260.3649777.

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. Why language models hallucinate, 2025. URL `https://arxiv.org/abs/2509.04664`.

Kalavasis, A., Mehrotra, A., and Velegkas, G. 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 (STOC)*, pp. 1732–1743. ACM, 2025. doi: 10.1145/3717823.3718108.

Kleinberg, J. and Mullainathan, S. Language generation in the limit. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 66058–66079. Curran Associates, Inc., 2024. doi: 10.52202/079017-2111.

Kleinberg, J. and Wei, F. Density measures for language generation, 2025. URL `https://arxiv.org/abs/2504.14370`.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, May 2022. URL `https://aclanthology.org/2022.acl-long.229/`.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. WebGPT: Browser-assisted question-answering with human feedback, June 2022. URL `https://arxiv.org/abs/2112.09332v3`.

Raman, A. and Raman, V. Generation from noisy examples. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 51079–51093, October 2025. URL `https://proceedings.mlr.press/v267/raman25a.html`.

Raman, V., Li, J., and Tewari, A. Generation through the lens of learning theory. In *Proceedings of 38th Conference on Learning Theory (COLT)*, volume 291 of *PMLR*, pp. 4740–4776, July 2025. URL `https://proceedings.mlr.press/v291/raman25a.html`.

Suzuki, A., He, Y., Tian, F., and Wang, Z. Hallucinations are inevitable but can be made statistically negligible. The ”innate” inevitability of hallucinations cannot explain practical LLM issues, May 2025. URL `https://arxiv.org/abs/2502.12187v2`.

Wu, C., Grama, A., and Szpankowski, W. No free lunch: Fundamental limits of learning non-hallucinating generative models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=OwNoTs2r8e`.

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models, 2025. URL https://arxiv.org/abs/2401.11817v2.

A. Relaxing the Regular Facts Assumption

A central assumption in the Kalai–Vempala framework is *Regular Facts* (Assumption 2.4), under which, conditioned on the observed corpus, all unobserved statements are equally likely to be factual. Kalai and Vempala (2024) introduce a natural relaxation of this assumption, allowing limited non-uniformity. This relaxation is technically straightforward and we describe its effect on our bounds in this appendix. We show that the only effect of relaxing Regular Facts to r-Regular Facts is a degradation of constants and failure probabilities proportionate to the relaxation.

Assumption A.1 (Definition 3: r-Regular Facts in Kalai & Vempala (2024))**.** We say that a meta-distribution D_{world} satisfies r*-Regular Facts* if for every corpus X and all unobserved statements $y \in U := \Omega \setminus \text{set}(X)$,

$$\Pr_{p\sim D_{\text{world}}}[y \in F \mid X] \;\le\; r \cdot \frac{\mathbb{E}_{p\sim D_{\text{world}}}[|F \cap U| \mid X]}{|U|}.$$

The case $r = 1$ recovers Regular Facts (Assumption 2.4). Intuitively, r-Regular Facts allows posterior factual probabilities of unseen statements to vary by at most a multiplicative factor of r from their mean.

A.1. Markov-Style Bound under r-Regular Facts

We now state the extension of the Markov-style hallucination bound.

Theorem A.2 (Markov-style bound under r-Regular Facts)**.** *Assume K-Sparsity and r-Regular Facts (Assumptions 2.3 and A.1). Let $g = \mathcal{A}(X)$ be the predictive distribution of a language model. Then for any $\delta \in (rK/|U|, 1)$,*

$$\Pr_{p\sim D_{\text{world}}}\left[g(H) \;\ge\; g(U)\Big(1 - \frac{rK}{\delta|U|}\Big) \;\Big|\; X\right] \;\ge\; 1 - \delta.$$

Proof. As in the proof of Theorem 4.1, we begin by bounding the expected factual mass assigned to unseen statements. By linearity of expectation,

$$\mathbb{E}[g(F \cap U) \mid X] = \sum_{y\in U} g(y) \Pr[y \in F \mid X].$$

By r-Regular Facts and K-Sparsity,

$$\Pr[y \in F \mid X] \;\le\; r \cdot \frac{\mathbb{E}[|F \cap U| \mid X]}{|U|} \;\le\; r \cdot \frac{K}{|U|}.$$

Substituting yields

$$\mathbb{E}[g(F \cap U) \mid X] \;\le\; r \cdot \frac{K}{|U|} \sum_{y\in U} g(y) = r \cdot g(U) \frac{K}{|U|}.$$

Since $g(H) = g(U) - g(F \cap U)$, it follows that

$$\mathbb{E}[g(H) \mid X] \;\ge\; g(U)\Big(1 - \frac{rK}{|U|}\Big).$$

The remainder of the proof proceeds exactly as in Theorem 4.1, yielding the stated probability bound. □

A.2. High-Confidence Bound under r-Regular Facts

We next describe the effect of r-Regular Facts on the high-confidence bound.

Theorem A.3 (High-confidence bound under r-Regular Facts)**.** *Assume K-Sparsity and r-Regular Facts (Assumptions 2.3 and A.1). Let $g = \mathcal{A}(X)$ be the predictive distribution of a language model. Then*

$$\Pr_{p\sim D_{\text{world}}}\left[g(H) \;\ge\; \frac{g(U)}{K+1} \;\Big|\; X\right] \;\ge\; 1 - \frac{rK}{|U|}.$$

Proof. The proof follows that of Theorem 4.2, except that the bound $\Pr[y^\star \in F \mid X] \le K/|U|$ is replaced under r-Regular Facts by $\Pr[y^\star \in F \mid X] \le rK/|U|$. The remainder of the argument is unchanged. □

A.3. Consequences for Missing-Mass Bounds

Since the relationship between innovation rate and missing mass (Proposition 4.3) does not rely on Regular Facts, all missing-mass corollaries in Section 4 extend immediately to r-Regular Facts by combining Theorems A.2 and A.3 with the same translation arguments. The effect of r-Regular Facts is limited to a multiplicative degradation of constants or, in the high-confidence case, a degradation of the confidence level.

B. Proofs omitted from the main paper

Proof of Observation 3.2. By definition, $H = \Omega \setminus F$ and $F \subseteq \Omega$. Since $O \subseteq F$, every hallucination is unobserved, i.e. $H \subseteq U$, so that $g(U) \geq g(H)$. Thus, $g(H) > 0$ implies $g(U) > 0$. □

Proof of Proposition 3.4. By calibration, there exists a partition Π of Ω such that $g = p^{\Pi}$. Now suppose that $p(U) > 0$. Then there exists some $y^\star \in U$ with $p(y^\star) > 0$. Let $B^\star$ be the cell of Π containing $y^\star$. Since $p(y^\star) > 0$, we have $p(B^\star) \geq p(y^\star) > 0$, and therefore $g(y^\star) = \frac{p(B^\star)}{|B^\star|} > 0$. Because $y^\star \in U$, this implies $g(U) \ \geq \ g(y^\star) \ > \ 0$. □

B.1. Proof of Proposition 4.3

To prove Proposition 4.3, we begin with a simple but key observation that lower-bounds the amount of probability mass that any coarsening can remove from the set U.

Lemma B.1 (Coarsening preserves a $1/(K+1)$-fraction of the missing mass)**.** *Let $p \in \Delta(\Omega)$ such that $|\operatorname{supp}(p)| \leq K$. Let $X \sim p^{\times n}$, and $O := \operatorname{set}(X)$, $U := \Omega \setminus O$. Let Π be a partition of Ω and let p^{Π} denotes the Π-coarsening of p as defined in Definition 2.6. Then, we have*

$$p^{\Pi}(U) \ \geq \ \frac{p(U)}{K+1}.$$

Proof of Lemma B.1. Fix a partition Π of Ω. First, let us show that the inequality holds for all the cells in the partition Π. Fix any $B \in \Pi$. We want to show the following

$$p^{\Pi}(U \cap B) \ \geq \ \frac{p(U \cap B)}{K+1},$$

If $U \cap B = \emptyset$, then $p^{\Pi}(U \cap B) = p(U \cap B) = 0$ and the above inequality holds trivially. Hence, we only have to argue for the case $U \cap B \neq \emptyset$. We argue as follows.

$$\begin{aligned}
p^{\Pi}(U \cap B) &= |U \cap B| \cdot \frac{p(B)}{|U \cap B| + |O \cap B|} && (1)\\
&\geq p(U \cap B) \cdot \frac{|U \cap B|}{|U \cap B| + |O \cap B|} && (2)\\
&\geq p(U \cap B) \cdot \frac{|U \cap B|}{|U \cap B| + K} && (3)\\
&= p(U \cap B) \cdot \frac{1}{1 + \frac{K}{|U \cap B|}} \\
&\geq p(U \cap B) \cdot \frac{1}{1 + K} && (4)
\end{aligned}$$

where, (1) follows from the coarsening definition, (2) follows from $U \cap B \subseteq B$, (3) uses $|O \cap B| \leq |O| \leq |F| \leq K$, and (4) follows from $U \cap B \neq \emptyset \implies |U \cap B| \geq 1 \implies \frac{K}{|U \cap B|} \leq K$.

Hence, for all the cells of the partition, the inequality holds. Summing over all the cells finishes the proof. □

Using this lemma, we prove Proposition 4.3 which relates innovation rate and missing mass.

Proof of Proposition 4.3. Total variation satisfies

$$\|g - p^{\Pi}\|_{\mathrm{TV}} \geq p^{\Pi}(U) - g(U).$$

Rearranging gives $g(U) \geq p^{\Pi}(U) - \|g - p^{\Pi}\|_{\mathrm{TV}}$. Applying Lemma B.1 completes the proof. $\square$

B.2. Proofs of Corollaries 4.4 and 4.5

Proof of Corollary 4.4. Theorem 4.1 implies that with probability at least $1 - \delta$ conditioned on the corpus X, it holds that

$$g(H) \geq g(U)\left(1 - \frac{K}{\delta|U|}\right). \tag{4}$$

Substituting the bound $g(U) \geq \frac{p(U)}{K+1} - \|g - p^{\Pi}\|_{\mathrm{TV}}$ from Proposition 4.3 in eq. (4) gives

$$\begin{aligned}
g(H) &\geq \left(\frac{p(U)}{K+1} - \|g - p^{\Pi}\|_{\mathrm{TV}}\right)\left(1 - \frac{K}{\delta|U|}\right) &(5)\\
&\geq \frac{p(U)}{K+1} - \frac{K}{(K+1)\delta|U|} - \|g - p^{\Pi}\|_{\mathrm{TV}} &(6)\\
&\geq \frac{p(U)}{K+1} - \frac{1}{\delta|U|} - \|g - p^{\Pi}\|_{\mathrm{TV}}, &(7)
\end{aligned}$$

where eq. (6) uses the conditions $p(U) \leq 1$ and $\delta > K/|U|$. $\square$

Proof of Corollary 4.5. Theorem 4.2 implies that with probability at least $1 - K/|U|$ conditioned on the corpus X, it holds that

$$g(H) \geq \frac{g(U)}{K+1}. \tag{8}$$

Substituting the bound $g(U) \geq \frac{p(U)}{K+1} - \|g - p^{\Pi}\|_{\mathrm{TV}}$ from Proposition 4.3 in eq. (8) gives the claim. $\square$