Calibrated Language Models Must Hallucinate

Adam Tauman Kalai
Microsoft Research

Santosh S. Vempala
Georgia Tech

December 5, 2023

Abstract

Recent language models generate false but plausible-sounding text with surprising frequency. Such "hallucinations" are an obstacle to the usability of language-based AI systems and can harm people who rely upon their outputs. This work shows shows that there is an inherent statistical lower-bound on the rate that pretrained language models hallucinate certain types of facts, having nothing to do with the transformer LM architecture or data quality. For "arbitrary" facts whose veracity cannot be determined from the training data, we show that hallucinations must occur at a certain rate for language models that satisfy a statistical calibration condition appropriate for generative language models. Specifically, if the maximum probability of any fact is bounded, we show that the probability of generating a hallucination is close to the fraction of facts that occur exactly once in the training data (a "Good-Turing" estimate), even assuming ideal training data without errors.

One conclusion is that models pretrained to be sufficiently good *predictors* (i.e., calibrated) may require post-training to mitigate hallucinations on the type of arbitrary facts that tend to appear once in the training set. However, our analysis also suggests that there is no statistical reason that pretraining will lead to hallucination on facts that tend to appear more than once in the training data (like references to publications such as articles and books, whose hallucinations have been particularly notable and problematic) or on systematic facts (like arithmetic calculations). Therefore, different architectures and learning algorithms may mitigate these latter types of hallucinations.

1 Introduction

The surprisingly high rate at which Language Models (LMs) generate false information, such as references to non-existent article titles, has recently emerged as a critical issue. The popular term *hallucination* is defined in the Merriam-Webster (2023) dictionary as "a plausible but false or misleading response generated by an artificial intelligence algorithm." In one case, lawyers were fined $5,000 for submitting legal research containing hallucinated legal cases that they believed were correct (Shin, 2023). In healthcare, hallucinations could be life threatening to patients and physicians are concerned about malpractice cases (Mello and Guha, 2023). Furthermore, hallucinations have been widely reported on by the media (Weise and Metz, 2023), and the U.S. President recently put out an Executive Order calling for, among other things, safeguards against misleading outputs from generative AI systems (Biden, 2023). This paper presents statistical lower-bounds on the rate of hallucination for LMs that are calibrated predictors of facts. This helps illuminate the nature of hallucination. It should *not* be taken to mean that hallucination is inevitable. Rather, as we discuss it is consistent with the fact that practitioners have increasingly been augmenting "pretraining"

1

procedures by "post-training" procedures that reduce hallucination rates at the cost of reducing calibration as well.

An LM is simply a probability distribution D over sequences of *tokens*, i.e., words or other character sequences. Clearly any LM which predicts every string with positive probability (a common property of LMs) will necessarily hallucinate with positive probability. However, if this probability is small, then hallucinations will be rare. Thus it is crucial to quantify the rate of hallucinations. Every distribution D can equivalently be represented by its log-probabilities over entire sequences or conditional log-probabilities of the subsequent token given previous ones, $\log D(t_1 \ldots t_m) = \sum_{i=1}^{m} \log D(t_i \mid t_1 \ldots t_{i-1})$. This mathematically trivial equivalence[1] has a profound implication: any LM can be used either to *generate* text or *predict* the next token in naturally occurring text conditioned on the previous tokens, though prediction and generation have different desiderata. For instance, consider the sentence,

> *Alexa Wilkins had a tuna sandwich* at Salumeria for lunch last Tuesday because the reviews said that it was divine.

Sentences of this sort could be likely under a predictive LM, for example, to offer suggestions to reduce typing on phones (e.g., Tanaka-Ishii, 2007). It may be desirable to predict *sandwich* as an option of a word to type after the word *tuna*, along with other likely words such as *salad*. On the other hand, the vast majority of sentences of this sort would be false if randomly fabricated by a generative LM. This paper shows LMs with good predictive text performance should hallucinate, even under ideal conditions. Notably in the first stage of pretraining, common today, the generative LM is optimized for predictive text performance (Chowdhery et al., 2022; OpenAI, 2023). Moreover, it gives a lower-bound on the rate of hallucination, which can shed light on the different rates at which different types of facts should be hallucinated.

What is common to both potential references and the example above (which we shall refer to as 5W = Who-Ate-What-When-Where-Why factoids), above is that they are *arbitrary* in the sense that neither one can be determined systematically by rules—one cannot determine the veracity of most such facts that are not present in the training data. This in contrast to facts whose veracity can be determined systematically. We quantify how much LMs should hallucinate *even in an simplified setting* with several ideal properties. Because we are giving statistical lower-bounds, we favor simplicity over generality as the point of our lower bounds is to identify a fundamental cause of LM hallucination. Similar to classification, where one seeks a lower-bound for the difficulty of classification in noiseless settings (but noise-tolerant classifications algorithms), we seek a hallucination lower-bound that holds in the simplest setting where training data is i.i.d. without factual errors.

Calibration for generative models. Calibration is a natural requirement of a probabilistic predictor meaning that its probabilities can be interpreted as accurate confidences in its own predictions. Dawid (1982) introduced the notion with the example of a weather forecaster: among days when they predict 30% chance of rain, it rains approximately 30% of the time. Calibration metrics have been extensively studied for LMs (see, e.g., Braverman et al., 2020; Jiang et al., 2021; Zhao et al., 2021). Fig. 1 illustrates multi-class calibration for GPT-4, a large modern LM, on a

---

[1]The equivalence between next-token-prediction and document generation holds only if one disregards computational costs. It is similar to the statement that book can be output one word at a time, in order, in a single pass, even though writing a book generally involves many iterations.

multiple choice exam. Post-training alignment was applied to reduce hallucination (among other factors) but was also found to reduce calibration (OpenAI, 2023). Calibration is both *meaningful*, since a calibrated predictor's probabilities are interpretable as accurate confidences, and statistically *achievable*.[2] In contrast, the perfectly accurate predictor would also be calibrated but may be impossible to learn. However, calibration is only a minimal requirement for predictors, as not all calibrated models are useful predictors: the predictor which always outputs the annual average probability of rain is trivially calibrated.

We provide a natural generalization of calibration to generative models. Our notion differs from prior uses of calibration in LMs which were at the token-level. The problem with analyzing raw token probabilities is that there are many ways to describe any fact in natural language, and thus having calibrated token probabilities is not particularly meaningful. To illustrate, consider the old *trigram* LMs, which predict next-token probabilities based only on the previous two tokens (e.g., words). Trigram models are naturally calibrated at the token level, and yet hallucination was not a major problem for trigram models. This is because they mostly generate gibberish. Instead, our semantic-level calibration considers the probability distribution over pieces of information (facts or hallucinations) contained in the text. We say an LM is calibrated if, for any probability $z \in [0,1]$, among the pieces of information it generates with probability $\approx z$, such information occurs on average in a $\approx z$ fraction of naturally occurring language (ideally the distribution from which training data was drawn).

Why LMs hallucinate. Hallucinations have mystified LM users and some researchers alike. Section 2 surveys numerous hypotheses that have been studied for why LMs hallucinate, ranging from inaccurate or outdated training data to the next-token log-likelihood objective in training. Hallucination can also be due to an adversarial or out-of-distribution *prompt*: a textual prefix provided for the LM to complete which establishes context. For example, there may be no factual completion to a fabricated prompt such as, *The 15 Elves of San Salami are named. . . .*[3] In contrast, our work shows that even in an ideal, unchanging world with perfect training data and no prompts, one should expect hallucinations from LMs which are calibrated.

Simplified setting. We consider a stationary language distribution $D_L \in \Delta(X)$ over documents (i.e., strings of text) $x \in X$, and a learning algorithm $\mathcal{A}$ which takes training data $\mathbf{x}_{\text{train}} \in X^n$ consisting of n documents sampled independently from D_L, and outputs an LM, i.e., a distribution $D_{LM} := \mathcal{A}(\mathbf{x}_{\text{train}}) \in \Delta(X)$. For simplicity, we assume that there are only facts in the training data, and at most one per document, i.e., no training hallucinations. We focus on *arbitrary* facts such as the above examples, whose truth is usually undetermined from the training set, rather than *systematic* facts such as $572 < 120523$ predictable from a training set by learning the underlying rules governing correctness. There is no statistical reason that LMs should hallucinate on systematic facts. Additionally, mistakes on systematic facts may not be considered hallucinations at all—they are often categorized as reasoning or arithmetic errors.

We assume that each document $x \in X$ contains at most one *factoid* $f(x) \in Y$, where factoids are arbitrary pieces of information which are each either true (facts) or false (hallucinations) and

---

[2]Simple post-hoc probability modification procedures can calibrate any uncalibrated predictor and simultaneously increase its accuracy under metrics such as cross-entropy (see, e.g., Braverman et al., 2020; Błasiok et al., 2023).

[3]A completion such as . . . *actually never mind, I have no idea* is unlikely given training data that does not have such retractions.

whose truth is statistically hard to determine from the training data. We also simplify matters by considering *unconditional generation* (e.g., Tan et al., 2021) in which the LM is sampled to generate text without any prompt (equivalently, the empty-string prefix). Again, compared to our simplified setting, hallucination may be even more likely in the more realistic case where the LM is prompted with contexts that come from a different distribution than the training data.

Results. Consider n i.i.d. samples from an unknown distribution p over a large number of arbitrary factoids, such as the 5W example and references. The *missing mass*, or in our case *missing facts* $p(U)$, is the fraction of future samples from this fact distribution p that were not observed in the n training samples, where U is the subset of facts that were unobserved in training data. The *Good-Turing* estimate of the missing mass (Good, 1953) is the fraction of samples (in our case facts) that appear exactly once in the training data. In our context, we call this the *MonoFacts* estimator,

$$\widehat{MF} := \frac{\text{Number of facts appearing exactly once in training data}}{n}.$$

The Good-Turing estimator was shown to be within $|p(U) - \widehat{MF}| = \tilde{O}(\sqrt{1/n})$ of the missing mass, with high probability (McAllester and Schapire, 2000; McAllester and Ortiz, 2003), for any distribution p.

If the correctness of arbitrary factoids not contained in the training cannot be determined, the missing fact rate can provide a lower-bound on the rate of hallucination. This in turn gives us a lower-bound close to $\widehat{MF}$. In particular, under a "regularity" assumption on the factoid distribution, our simplest bound (Corollary 1) implies that, for *any* algorithm, with probability $\geq 99\%$ over training sets:

$$\text{Hallucination rate} \geq \widehat{MF} - \text{Miscalibration} - \frac{300|\text{Facts}|}{|\text{Possible hallucinations}|} - \frac{7}{\sqrt{n}},$$

where the "Hallucination rate" is the rate at which the LM generates hallucinations, the next term is "monofact" estimator of the missing facts. The subsequent term is a "miscalibration rate" which quantifies how close to calibration the distribution is. We also have a term involving the ratio of the number of arbitrary facts to similar pieces of information that are false, which is exponentially small for many types of information. The final $6/\sqrt{n}$ term is small for the large training set sizes n used in today's LM. The regularity assumption means that all unobserved factoids have, on average, equal probability of being true. More generally, the bound holds with probability $\geq 1 - \delta$ where the constant 60 can be replaced by a term which is inversely proportional to δ and proportional to a regularity term on the distribution of factoids. The regularity term measures the ratio of the most likely factoid (that was not observed in training data) to the average unobserved factoid probability. This constant is 1 for symmetric distributions and other types of simple distributions. We relax it to consider bounded regularity, which permits there to be certain negative correlations such as the fact that a person does not eat 1000 lunches in 1000 different places on the same day and allows for some factoids to have conditional probability 0, but it prohibits unobserved factoids from having very large probability.

Interpretation. Our lower-bounds have the following interpretation. First, one should identify a large set of factoids: arbitrary, plausible, regular factoids. These could be posts about 5W and plausible research article citations. Intuitively, there are exponentially more plausible factoids that

are incorrect (which we call hallucinations), than those that are facts. One can then consider what fraction of such factoids might appear exactly once in the training data. In the case of 5W, one could imagine that half of the posts occur exactly once. This would suggest that a calibrated-factoid model would hallucinate on about half of its generations on 5W factoids. On the other hand, one could imagine that there are many many fewer than n articles and since the goal of publication is advertising, each reference might be expected to appear multiple times in the training data[4] (i.e., have probability $\gg 1/n$) except for perhaps very recent publications (e.g., within the last day before they other references begin to appear). This would suggest that the missing mass for articles is low and that there is no inherent statistical necessity to hallucinate reference titles. There may be other causes of such hallucinations such as limited model capacity (even if the number of LM parameters is much greater than the number of articles, these parameters must encode many types of information beyond article titles). This also justifies the mitigation of consulting a fact database at generation time, even if that fact database is constructed solely from training data (Borgeaud et al., 2022). Corollary 4 gives a simple generalization of our analysis to multiple *types* of facts.

Conversely, one may wonder what if any facts appear only once in a large training corpus that might encompass the entire web. First, it is worth noting that significant efforts are made, in constructing LM corpora, to remove duplicates (see, e.g. Shen et al., 2023). And in other contexts, such as Zipfian and other power-law distributions, it is often observed that a constant fraction of entities appear exactly once, even as the datasets scale. While most facts that come to mind may be well-known, many public meeting notes and other posts may contain unremarkable facts that are not mentioned in other places.

Despite this tension between factuality and predictive accuracy, the parameters of both types of LMs are typically trained or "pretrained" to maximize likelihood on a corpus, or equivalently to minimize the "KL divergence", a strong statistical discrepancy metric between the LM and the distribution of data on which it was trained.

Organization. Following related work discussed in Section 2, Section 3 gives mathematical preliminaries including our definition of (mis)calibration. Section 4 defines our factoid model and Section 5 states the main results. Sections Section 6 proves the main lemma. Section 7 provides a simple upper bound showing that one cannot guarantee a hallucination rate greater than the monofact estimate for nearly calibrated estimators. Section 8 concludes and discusses limitations and future work. Appendix A in the appendix presents bounds on Good-Turing estimators from prior work. Appendix B discusses extensions to alternative models such as calibration using bins of equal width, those based on raw accuracy (cross-entropy), or those using using prompts. Appendix C gives additional proofs.

2   Related work

As discussed in the introduction, the concept of calibration was introduced by Dawid (1982) and has been extensively studied in statistics and machine learning and even specifically for LMs (Braverman et al., 2020; Zhao et al., 2021; Jiang et al., 2021) and other related fields of deep learning and generative AI (Rohrbach et al., 2018; Maynez et al., 2020). Błasiok et al. (2023) argue that

---

[4]Even papers that are never cited often appear on an authors' CVs and institutional web pages and are cross-posted and listed in multiple publication indexes since the authors and institutions often wish to publicize their output and indices such as Google/Semantic Scholar may compete for completeness.
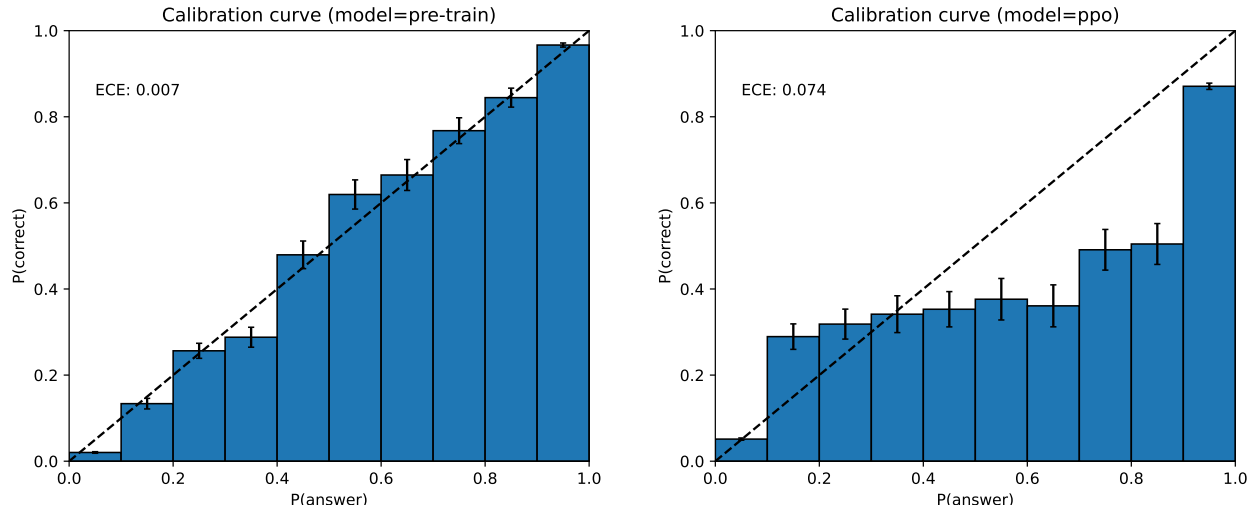
Figure 1: GPT-4 calibration curves before (left) and after (right) reinforcement learning (OpenAI, 2023, Figure 8, reprinted with permission). As suggested by our model, post-training reduces hallucination rates at the cost of increasing calibration error. Note that calibration here is on a multiple-choice test rather than generative calibration which we study.

calibration happens naturally as a byproduct of minimizing log-loss for deep neural networks, though their results are for a different architecture and different notion of calibration.

Unfortunately, there is not clear agreement on what counts as a hallucination. This is why we consider an idealized model in which there are clear-cut facts, where statements that violate these facts would generally be categorized as hallucinations by most definitions.

Open- vs. closed-domain hallucinations. Interestingly, many studies focus on hallucination with respect to a specific source document that is given to an LM, such as in translation (Xu et al., 2023) or summarization (Maynez et al., 2020). There, LMs are also found to fabricate facts not present in the source document even when instructed to use only information present in the source. This is referred to as closed-domain hallucination in contrast to our open-domain setting, where the LM generates hallucinations which are not factually grounded in its training data. There is no clear statistical argument for why closed-domain hallucinations must occur, since if one can verify such hallucinations from the source text and generation one can avoid them by filtering out generations with such mistakes. Consistent with this, (OpenAI, 2023) reports a greater reduction in closed-domain hallucinations over open-domain ones.

Honesty vs. factuality. Evans et al. (2021) points out that there is a difference between *factuality* and *truthfulness* (i.e., honesty). An LM which states something that disagrees with its training data may be said to *lie*, whereas *factuality* may be much harder to determine for a variety of reasons, including the possibility that what is considered factual may change as scientists make new discoveries and rebuke old theories. Nonetheless, in worlds where there is an absolute notion of facts, and the training data only contain facts, then any falsehood is also untruthful. Thus in ideal worlds where training data is perfectly consistent with an internally consistent ground-truth

notion of facts, our bounds on non-factuality directly imply bounds on the rate at which LMs must generate untruthful information or "lie."

Hypotheses for why LMs hallucinate. There have been many explanations of why LMs hallucinate. The primary reason proposed for hallucinations is inadequacies in training data. This can be broken into two issues. First are falsehoods contained in the training data (Ji et al., 2023; Lin et al., 2022; Dziri et al., 2022; Filippova, 2020), which Lin et al. (2022) refer to as "imitative falsehoods." Second is the temporal nature of data, i.e., the fact that training data is no longer relevant and is missing current events (Aksitov et al., 2023; Vu et al., 2023). While both of these are certainly factors in hallucination, our work shows that it is not the only cause of LM hallucinations.

An additional reason given for why LMs may hallucinate is the fact that they are trained to produce tokens one at a time may lead to hallucination (Zhang et al., 2023b) because the LM may generate a plausible-sounding sequence of tokens which is impossible to complete factually. While this may be true for transformers due to computational limitations, this is not simply a statistical byproduct of their being trained on the next-token prediction task. Since document log-likelihood is simply the sum of next-token log-likelihood, the two objectives are identical and thus from a statistical perspective this is not a factor. Mathematically, any probability distribution over documents can be represented as a conditional next-token probability distribution.

Another line of work that sheds light on the nature of hallucinations shows that LMs *know when they're hallucinating* (Kadavath et al., 2022; Azaria and Mitchell, 2023; Agrawal et al., 2023; Manakul et al., 2023). Various techniques may be used to identify LM hallucinations purely from the LM itself, either the internal weights, its token probabilities, or by querying it with additional questions. This is consistent with our work in the sense that one would expect even an ideal "super-inelligent" model should hallucinate if its goal is predictive accuracy.

A bevy of additional reasons have been proposed and studied for why LMs may hallucinate. These fall under headings such as duplicates in training data, biases, architecture, overconfidence and various types of knowledge failures, among others. A complete listing of these is beyond this scope of this work. For further details, see the recent surveys of Huang et al. (2023); Zhang et al. (2023a); Ji et al. (2023).

3 Mathematical Preliminaries

We first define preliminary notation and concepts, including statistical distance and our notion of generative calibration.

3.1 Standard definitions and notation

Let $\Delta(S)$ denote the set of probability distributions over S. For distribution $\mathcal{D} \in \Delta(S)$ and $R \subseteq S$, let $\mathcal{D}(R) := \sum_{x \in R} \mathcal{D}(x)$. The *Total Variation distance* (TV) (also called statistical distance) between distributions $\mathcal{D}, \mathcal{D}' \in \Delta(S)$ has multiple equivalent definitions (for finite or countably infinite sets S):

$$\|\mathcal{D} - \mathcal{D}'\|_{\mathrm{TV}} := \max_{R \subseteq S} |\mathcal{D}(R) - \mathcal{D}'(R)| = \frac{1}{2} \sum_{x \in S} |\mathcal{D}(x) - \mathcal{D}'(x)| = \sum_{x \in S} \left(\mathcal{D}(x) - \mathcal{D}'(x) \right)_+, \quad (1)$$

where $(z)_+ = \max(0, z)$ for $z \in \mathbb{R}$. It satisfies $\|\mathcal{D} - \mathcal{D}'\|_{\text{TV}} = \|\mathcal{D}' - \mathcal{D}\|_{\text{TV}} \in [0, 1]$. The *support* of distribution $\mathcal{D}$ is $\text{supp}(D) = \{x \in S \mid \mathcal{D}(x) > 0\}$, as long as S is finite or countably infinite. Let the set of *partitions* of set S be denoted by $\mathcal{P}(S) := \{\Pi \subseteq 2^S \mid \forall x \in S \, |\{B \in \Pi \mid x \in B\}| = 1\}$. For a function $f : X \to Y$ and set $S \subseteq X$, let $f(S) := \{f(x) \mid x \in S\}$. For a distribution $\mathcal{D} \in \Delta(X)$, let $f \circ \mathcal{D}$ denote the induced distribution of $f(x)$ for $x \sim \mathcal{D}$, i.e., $f(y) := \sum_{x:f(x)=y} \mathcal{D}(x)$. Finally, let $\mathcal{D}^{\times n}$ denote the distribution over n independent samples from $\mathcal{D}$.

3.2 Generative (mis)calibration

This section defines a notion of *miscalibration* $\text{Mis}_b(p, g) \in [0, 1]$ for a generative model g that measures how accurate its own probabilities are with respect to future examples drawn from a given pretraining distribution p, using $b \geq 1$ bins. It is a natural extension of existing notions of calibration to generative models, and can be skipped on first read for those who want to get straight to the model. Appendix B discusses the relationship between this and existing notions of calibrated classifiers. The definitions in this section apply to any distributions $p, g \in \Delta(Y)$ for any finite set Y. In other words, there is only assumed to be a generative distribution g over information $y \in Y$, and calibration is with respect to a "true" distribution p. Finiteness of Y is only assumed at this point to avoid measure-theoretic technicalities. We first define a calibrated distribution g as any *coarsening* of p.

Definition 1 (Coarsening and calibration). *For partition $\Pi \in \mathcal{P}(Y)$ and $\mathcal{D} \in \Delta(Y)$, $\mathcal{D}^\Pi \in \Delta(Y)$ is the Π-coarsening of $\mathcal{D}$ if,*

$$\forall B \in \Pi \ \forall y \in B \quad \mathcal{D}^\Pi(y) = \frac{\mathcal{D}(B)}{|B|}.$$

Distribution g is said to be calibrated *to p if $g = p^\Pi$ for some partition Π.*

Clearly $g = p$ is calibrated to p, and so is the uniform distribution $g(y) = u(y) := 1/|Y|$. To define *miscalibration*, let $B_z^g := \{y \in Y \mid g(y) = z\}$ and omit g, writing $B_z = B_z^g$ when clear from context. It is also clear that g is calibrated to p iff $g = p^{\mathcal{B}(g)}$ for partition $\mathcal{B}(g) := \{B_z \mid z \in [0, 1]\}$ and thus a natural definition of miscalibration (which is 0 iff g is calibrated to p) is:

$$\text{Mis}_\infty(g, p) := \left\|p^{\mathcal{B}(g)} - g\right\|_{\text{TV}} = \frac{1}{2} \sum_{B \in \mathcal{B}(g)} \sum_{y \in B} \left|\frac{p(B)}{B} - g(y)\right|. \tag{2}$$

The ∞ in the above notation refers to the fact that there is no limit on the number of bins. This also explains why it is called "calibration": the average probability over each bin that shares a common value of $g(y)$ is $g(y)$. This can be written as,

$$\forall z \in [0, 1] \quad \mathop{\mathbb{E}}_{y \sim g}\big[p(y) \mid g(y) = z\big] = z. \tag{3}$$

We next generalize $\mathcal{B}(g)$ to intervals. For $I \subseteq [0, 1]$, define:

$$B_I = B_I^g := \{y \in Y \mid g(y) \in I\}.$$

The definition below uses $b \geq 1$ adaptively sized bins which each have roughly equal probability mass in terms of generations $y \sim g$. Appendix B.3 gives an alternate definition using bins of equal width in terms of log-probability.

8

Definition 2 (Miscalibration). *Let $b \in \mathbb{N}$ and define the adaptive partition,*

$$\mathcal{V}_b(g) := \left\{ B_{[0,t_1]}, B_{(t_1,t_2]}, \ldots, B_{(t_{b-1},t_b]} \right\} \text{ where } t_i = \sup \left\{ z \in [0,1] \,\middle|\, \sum_{y:g(y) \leq z} g(y) \leq i/b \right\}.$$

The miscalibration *of g with respect to p is* $\text{Mis}_b(g,p) := \left\| p^{\mathcal{V}_b(g)} - g \right\|_{\text{TV}}$.

Adaptive binning has been used in supervised classification (e.g., Nixon et al., 2020). Note that $\text{Mis}_1(g,p) = \|g - u\|_{\text{TV}}$ is the total variation to the uniform distribution, which shows that $\text{Mis}_b(g,p)$ is not monotonic in b because $b = 1$ is the minimum for $g = u$ (regardless of p) while $b = \infty$ minimizes $\text{Mis}_b(p,p) = 0$ for $g = p$. Also, $\text{Mis}_b(g,p) = \text{Mis}_\infty(g,p)$ for b is large enough that $1/b \leq \min_{y \in \text{supp}(g)} g(y)$. Finally note that necessarily $t_b = 1$ in the above definition.

Advantages and limitations. An advantage of semantic-level calibration is that it naturally models the exponentially many ways there are to describe any given fact, unlike token-level calibration models. This is also a disadvantage, however, because it means that it may be intractable to measure calibration rates, and thus experimental validation may be easier in synthetic settings where facts have canonical descriptions. One nice property of the above adaptive binning is that $\text{Mis}_b(g,p) = 0$ iff g is calibrated to p, regardless of b, whereas other definitions give 0 miscalibration whenever there is a single bin. Nonetheless, Appendix B.3 shows how our analysis applies to other binning strategies.

4 The model and guarantees

Our notation is summarized in Table 1. Let Σ denote the set of tokens and Σ^* denote the set of finite token sequences, i.e., strings. Let $X \subseteq \Sigma^*$ denote the set of documents (this could be finite, e.g., if a length restriction is imposed or countably infinite if $X = \Sigma^*$).

The world is assumed to contain randomness, which is modeled by a distribution $D_{\text{world}} \in \Delta(\Delta(X))$ over language (document) distributions $D_L \sim D_{\text{world}}$. (More generally, a full world model would contain lot of other information but language distributions suffices for our purposes.) The training data consists of n documents $\mathbf{x}_{\text{train}} \sim D_L^{\times n}$ sampled independently from this distribution. It will be convenient to denote by $D_{\text{train}} \in \Delta(X^n)$ the distribution over training data, where the probability of any training document is:

$$D_{\text{train}}(\mathbf{x}_{\text{train}}) := \mathbb{E}_{D_L \sim D_{\text{world}}} \left[\prod_{i=1}^n D_L\big(x_{\text{train}}^{(i)}\big) \right].$$

This model requires a static world that does not change. Of course, real-world distributions are non-stationary and dynamic, and real-world training data contains duplicates and is not i.i.d. However, our lower bounds imply hallucination even in this ideal static setting.

4.1 Factoid assumptions

We assume there is a large but finite set Y of "factoids" by which we mean *arbitrary plausible* pieces of information, each of which could be true (facts) or false (hallucinations), as determined by

some world distribution. Their arbitrary nature means that, given the facts observed in training data, one cannot infer any likely facts among the unobserved factoids. Of course, many real-world facts are systematic, not arbitrary. For instance, mathematical inequalities such as $17 < 252395$ are systematic and should thus not be included in Y. Note that Y is not intended to capture all world facts but rather a large set of arbitrary ones. It could contain the 5W factoids (except for people who eat the same lunch every day in the same location, as their eating behaviors are too systematic). There is no statistical reason an LM must hallucinate on systematic facts.

We will make a few assumptions about factoids.

1. **One-per-doc**: First, we assume that there is at most one factoid per document by defining a surjective function $f : X \twoheadrightarrow Y$ which extracts a single factoid with each document, where $f(x) = \perp$ represents the empty fact (to allow for documents with no facts) and assume $\perp \in Y$. This makes the notation simpler as one can define the induced factoid distribution $p = f \circ D_L \in \Delta(Y)$ defined by $p(y) := \sum_{x:f(x)=y} D_L(x)$. Similarly, we can take $g = f \circ D_{LM}$ to be the induced distribution over generated factoids, where $D_{LM} \in \Delta(X)$ is the distribution over documents generated by the LM. The surjective requirement simply means that every factoid is describable by some document $x \in X$, and if this didn't hold one could always shrink Y. Our model does permit many documents to encode the same factoid, since there are typically many ways to describe any given piece of information. Again, this assumption may be generalized to a model where documents contain sets of factoids with further notation, but we show that calibrated LMs hallucinate even in a simple idealized setting with one factoid per document.

2. **Good-training-data**: Second, we assume that the set of facts $F := \mathrm{supp}(p) \cup \{\perp\}$ where $\mathrm{supp}(p) = \{y \in Y \mid p(y) > 0\} = f(\mathrm{supp}(D_L))$, i.e., the training data consists solely of facts and every fact has non-zero probability of being described under D_L. Both of these assumptions can be removed at the cost of additional notation without changing the results in the slightest—the world distribution would need to determine an arbitrary $F \subseteq Y$ and $D_L \in \Delta(Y)$. Keeping in the spirit of the ideal training data model, we choose to simplify notation and take $F := \mathrm{supp}(p) \cup \{\perp\}$. The set of hallucinations is $H := Y \setminus F$, i.e., every non-fact is defined to be a hallucination.

3. **More-false-than-true**: Third, Assumption 1 below requires that there are many more falsehoods in Y than facts, which makes sense for many types of information. For example, consider those factoids which describe a paper citation including the paper title, authors, and further details. There are vastly more plausible citations than existing ones. In this case, one may choose to include in Y a somewhat smaller sparse set, i.e., not include author middle names or other minor variations in Y which would make a reference in the "grey area" between fact and hallucination. Similarly, for 5W factoids, there are many more combinations of people who did not eat a given food on at a given time than people who did.

   Assumption 1 (s-Sparse facts). *There are many fewer facts than hallucinations. Specifically, there exists $s \in \mathbb{R}$ such that, with probability 1 over $D_L \sim D_{\mathrm{world}}$:*

   $$|F| \le e^{-s}|H|.$$

   Recall that $F := f(\mathrm{supp}(D_L)) \cup \{\perp\}$ and $H = Y \setminus F$.

10

We write sparsity as an exponential to reflect the general exponentially nature of natural language facts.

4. **(Semi)Regularity**: Finally, and most importantly, we assume that no single factoid is very likely. Specifically, perfect regularity requires that after observing the training data, all unobserved factoids are equally likely to appear as facts in the language distribution, and we also provide a relaxed r-regular notion, which we refer to as a "semi-regularity" assumption.

Definition 3 (Regular facts). *D_{world} has regular facts if for all $\mathbf{x}_{\text{train}} \in \text{supp}(D_{\text{train}})$:*

$$\forall y, y' \in U \quad \Pr[y \in F \mid \mathbf{x}_{\text{train}}] = \Pr[y' \in F \mid \mathbf{x}_{\text{train}}].$$

For $r \geq 1$, D_{world} has r-regular-facts if for all $\mathbf{x}_{\text{train}} \in \text{supp}(D_{\text{train}})$,

$$\forall y \in U \quad \Pr[y \in F \mid \mathbf{x}_{\text{train}}] \leq \frac{r}{|U|} \mathbb{E}\big[|F \cap U| \;\big|\; \mathbf{x}_{\text{train}}\big].$$

Having regular facts will suffice to prove the lower bound Corollary 2, but stronger lower bounds will be possible if we also have regular probabilities.

Definition 4 (Regular probabilities). *D_{world} has regular probabilities if for all $\mathbf{x}_{\text{train}} \in \text{supp}(D_{\text{train}})$:*

$$\forall y, y' \in U \quad \mathbb{E}[p(y) \mid \mathbf{x}_{\text{train}}] = \mathbb{E}[p(y') \mid \mathbf{x}_{\text{train}}].$$

For $r \geq 1$, D_{world} has r-regular-probabilities if for all $\mathbf{x}_{\text{train}} \in \text{supp}(D_{\text{train}})$,

$$\forall y \in U \quad \mathbb{E}[p(y) \mid \mathbf{x}_{\text{train}}] \leq \frac{r}{|U|} \mathbb{E}[p(U) \mid \mathbf{x}_{\text{train}}].$$

Finally, we say D_{world} is *regular* if it has regular facts and regular probabilities. It is not difficult to see that a distribution is regular iff it has 1-regular-facts and 1-regular-probabilities. We also note that our regularity assumptions could be modified to only holds with high-probability over $\mathbf{x}_{\text{train}}$ and not for all $\mathbf{x}_{\text{train}} \in \text{supp}(D_{\text{train}})$.

We illustrate a simple family of regular world distributions which satisfy our assumptions, followed by one which is only r-regular, does not have independence, and has anti-correlations between facts.

Regular example: Permuted power-law world. Suppose $X = Y$ and f is the identity. The world distribution D_{world} first picks $F \subseteq Y$ uniformly random over such that $|F| = N$ (where $N \leq |Y|/(1 + e^s)$ so that $|F|/|H| \leq e^{-s}$ so facts are s-sparse) and then picks p to be a power-law distribution supported on a random ordering on F. That is, it picks a random permutation $\sigma : \{1, 2, \ldots, N\} \hookrightarrow F$ and defines p so that $p(\sigma(i)) \propto i^{-k}$ for some constant $k \geq 0$. This includes the uniform distribution over F ($k = 0$) and Zipfian distributions ($k = 1$) as special cases.

Symbol	Meaning
$X \subseteq \Sigma^*$	The set of documents (strings)
Y	Set of *factoids*, arbitrary plausible pieces of information
$\perp \in Y$	Special "empty string" fact representing
$f : X \twoheadrightarrow Y$	Each document x contains a single factoid $f(x)$, and $f(X) = Y$
$D_{\text{world}} \in \Delta(\Delta(X))$	Distribution over document distributions
$D_L \in \Delta(X)$	Language distribution over documents $D_L \sim D_{\text{world}}$
$\mathbf{x}_{\text{train}} \in X^n$	Training data (i.i.d.) $\mathbf{x}_{\text{train}} = \left\langle x_{\text{train}}^{(1)}, x_{\text{train}}^{(2)}, \ldots, x_{\text{train}}^{(n)} \right\rangle \sim \mathcal{D}_L^{\times n}$
$\mathcal{A} : X^n \to \Delta(X)$	Algorithm that learns an LM from training data
$p \in \Delta(Y)$	Distribution over factoids $f(x)$ for documents $x \sim D_L$, i.e., $p := f \circ D_L$
$F \subseteq Y$	Facts $F := \text{supp}(p) \cup \{\perp\}$, non-hallucinations
$H \subseteq Y$	Hallucinations $H := Y \setminus F$
$D_{\text{train}} \in \Delta(X^n)$	Distribution over $\mathbf{x}_{\text{train}} \sim \mathcal{D}_L^{\times n}$ induced by $D_L \sim D_{\text{world}}$.
$O \subseteq Y$	Set of observed factoids $O := \left\{ f\big(x_{\text{train}}^{(i)}\big) \mid i = 1, 2, \ldots, n \right\} \cup \{\perp\} \subseteq F$
$U \subseteq Y$	Unobserved factoids $U := Y \setminus O \supseteq H$
ν	Posterior over p given training data $\mathbf{x}_{\text{train}}$
$D_{LM} \in \Delta(X)$	Distribution over documents generated by the LM, $D_{LM} := \mathcal{A}(\mathbf{x}_{\text{train}})$
$g \in \Delta(Y)$	Distribution over factoids $f(x)$ for documents $x \sim D_{LM}$, i.e., $g := f \circ D_{LM}$

Table 1: Summary of notation. Symbols below the line are all derived quantities in terms of symbols above the line.

Semi-regular example: W5 with negative correlations. The set of factoids is the product of fixed sets of people, dates, foods, and locations. D_{world} chooses the set of facts F randomly by: for each person on each date, there is a single fact which consists of that person eating a random food at a random location, and p is uniform over F. This creates *anti-correlations* between factoids because the knowledge about what and where a person ate a specific meal rules out any possible alternatives for that meal. Nonetheless, it can be seen that r-regularity holds for $r \leq n_{\text{people}} n_{\text{dates}}$.

The above model can be enriched in various ways by adding reasons and by modeling people's preferences, geographic constraints, and behaviors. However the regularity parameter r will be prohibitively large if there are predictable eaters, and indeed LMs might hallucinate less if there are large numbers of predictable eaters because an LM learning algorithm may learn these patterns and hallucinate less often.

5   Guarantees

Our results are stated in terms of missing facts, a term inspired by the missing mass in data drawn from a probability distribution (Good, 1953). The *missing facts rate* is the fraction of facts (according to the pretraining fact distribution p) that were not observed in the training data. Specifically, it is defined to be $p(U)$ where U is the set of unobserved facts in the training data. Formally, define the *observed* and *unobserved* factoids as,

$$O = O_{\mathbf{x}_{\text{train}}} := \left\{ f\big(x_{\text{train}}^{(i)}\big) \,\Big|\, i = 1, 2, \ldots, n \right\} \cup \{\perp\} \subseteq Y, \quad U = U_{\mathbf{x}_{\text{train}}} := Y \setminus O_{\mathbf{x}_{\text{train}}}, \qquad (4)$$

respectively. The *monofact estimate* of the missing fact rate is defined to be the fraction of facts that appear exactly once in the training data:

$$\widehat{MF} = \widehat{MF}_{\mathbf{x}_{\text{train}}} := \frac{\left| \left\{ y \in Y \setminus \{\perp\} \mid y = f\big(x_{\text{train}}^{(i)}\big) \text{ for exactly one } i \in \{1, 2, \dots n\} \right\} \right|}{n}. \tag{5}$$

Note that the facts in the training data are distributed according to the distribution p. Appendix A states classical results asserting that $|\widehat{MF} - p(U)| = \tilde{O}(\sqrt{1/n})$ with high probability over n samples from any distribution p.

An *algorithm* $\mathcal{A} : X^n \to \Delta(X)$ takes as input n training documents and outputs a document distribution $D_{LM} = A(\mathbf{x}_{\text{train}})$, which determines $g = f \circ D_{LM}$, i.e., $f(x)$ for $x \sim D_{LM}$. We now state our first result, which relies on a regular D_{world} defined in Definitions 3 and 4 above.

Corollary 1. *Fix any* $\delta \in [0,1], b, n \in \mathbb{N}, s \in \mathbb{R}$ *and any* s-*sparse regular* D_{world}. *Then for any algorithm* $\mathcal{A} : X^n \to \Delta(X)$, *with probability* $\geq 1 - \delta$ *over* $D_L \sim D_{\text{world}}$ *and* $\mathbf{x}_{\text{train}} \sim D_L^{\times n}$,

$$g(H) \geq \widehat{MF} - \text{Mis}_b(g, p) - \frac{3e^{-s}}{\delta} - \sqrt{\frac{6\ln(6/\delta)}{n}},$$

where $D_{LM} = A(\mathbf{x}_{\text{train}})$, $g(H)$ *is the LM hallucination rate, and* $\widehat{MF}$ *is defined in Eq. (5).*

Now, we can state a weaker guarantee for semi-regular facts alone.

Corollary 2. *Fix any* $\delta \in [0,1], b, n \in \mathbb{N}, r, s \in \mathbb{R}$ *and any* s-*sparse* D_{world} *with* r-*regular-facts. Then for any algorithm* $\mathcal{A} : X^n \to \Delta(X)$, *with probability* $\geq 1 - \delta$ *over* $D_L \sim D_{\text{world}}$ *and* $\mathbf{x}_{\text{train}} \sim D_L^{\times n}$,

$$g(H) \geq \widehat{MF} - \text{Mis}_b(g, p) - \frac{3rne^{-s}}{\delta} - \sqrt{\frac{6\ln(6/\delta)}{n}}.$$

The above is meaningful when sparsity $s \gg \log n$ is larger than the log of the number of training data. Otherwise, following bound uses semi-regularity of facts and probabilities.

Corollary 3. *Fix any* $\delta \in [0,1], b, n \in \mathbb{N}, r, s \in \mathbb{R}$ *and any* s-*sparse* D_{world} *with* r-*regular-facts and* r-*regular-probabilities. Then for any algorithm* $\mathcal{A} : X^n \to \Delta(X)$, *with probability* $\geq 1 - \delta$ *over* $D_L \sim D_{\text{world}}$ *and* $\mathbf{x}_{\text{train}} \sim D_L^{\times n}$,

$$g(H) \geq \widehat{MF} - \text{Mis}_b(g, p) - \frac{3re^{-s}}{\delta} - \sqrt{\frac{6\ln(6/\delta)}{n}}.$$

It is easy to see that Corollary 1 is a special case of this corollary for $r = 1$.

5.1 Different types of facts

Our analysis immediately generalizes to multiple distinct types of facts (e.g., article references and social media posts). Suppose there are $k > 1$ sets of factoids, $Y_1, Y_2, \dots, Y_k$ and functions $f_i : X \to Y_i$, with corresponding sets of facts and hallucinations $F_i \cup H_i = Y_i$, monofact estimates $\widehat{MF}_i \in [0,1]$ and miscalibration rates $\text{Mis}_{i,b}(g, p)$. One also would generalizes the notion of s-sparse to include the fact that, for each type of fact, $|F_i| \leq e^{-s}|H_i|$ with probability 1 and similarly generalize regular facts to hold for each type of fact. Then Corollary 1 implies:

Corollary 4. *Fix any $\delta \in [0, 1], b, k, n \in \mathbb{N}, s \in \mathbb{R}$ and any s-sparse regular D_{world}. Then for any algorithm $\mathcal{A} : X^n \to \Delta(X)$, with probability $\geq 1 - \delta$ over $D_L \sim D_{\text{world}}$ and $\mathbf{x}_{\text{train}} \sim D_L^{\times n}$,*

$$g(H_i) \geq \widehat{MF}_i - \text{Mis}_{i,b}(g, p) - \frac{3ke^{-s}}{\delta} - \sqrt{\frac{6 \ln(6k/\delta)}{n}} \ \text{for } i = 1, 2, \ldots, k.$$

The proof follows trivially from Corollary 1 using the union bound and δ/k failure probability for each type of fact.

5.2 Analysis approach

While our model supposes $D_L \sim D_{\text{world}}$ followed later by $\mathbf{x}_{\text{train}} \sim \mathcal{D}_L^{\times n}$, for analysis purposes we imagine first selecting $\mathbf{x}_{\text{train}} \sim D_{\text{train}}$ and then selecting $p \sim \nu_{\mathbf{x}_{\text{train}}}$ where $\nu = \nu_{\mathbf{x}_{\text{train}}}$ is defined to be the posterior distribution on p given $\mathbf{x}_{\text{train}}$. These two procedures result in identical joint distributions on $p, \mathbf{x}_{\text{train}}$, but the latter is easier to analyze. In particular, we show:

Theorem 1. *For any $\nu \in \Delta(\Delta(Y))$, any $O \subseteq F \subseteq Y$, any $g \in \Delta(Y)$, and any partition $\Pi \in \mathcal{P}(Y)$,*

$$\mathop{\mathbb{E}}_{p \sim \nu} \left[\left(p(U) - \left\| p^\Pi - g \right\|_{\text{TV}} - g(H) \right)_+ \right] \leq \max_{y \in U} \mathop{\text{Pr}}_{p \sim \nu} [y \in F] + |O| \max_{y \in U} \mathop{\mathbb{E}}_\nu [p(y)],$$

where $H := Y \setminus F$ and $U := Y \setminus O$.

The corollaries stated earlier follow directly from this theorem together with the definition of $\text{Mis}_b(g, p) := \left\| p^{\mathcal{V}_b(g)} - g \right\|_{\text{TV}}$, Markov's inequality for a non-negative random variable to show that the quantity in the expectation is small with high probability, which we combine with existing bounds on the Good-Turing estimator from Appendix A that show that $|\widehat{MF} - p(U)| = \tilde{O}(\sqrt{1/n})$. The quantities on the right hand side correspond to the arbitrary facts and arbitrary probability notions. The above theorem is enough to show the result for various binning strategies, such as fixed-width, which depend arbitrarily on g (but not on p).

6 Proof of Theorem 1

This section proves Theorem 1.

Lemma 2. *Let $S \subseteq Y$ and let p^Π be any coarsening of p. Then,*

$$\mathop{\mathbb{E}}_\nu \left[\left(p(S) - p^\Pi(S) \right)_+ \right] \leq |Y \setminus S| \cdot \max_{y \in S} \mathop{\mathbb{E}}_\nu [p(y)].$$

Proof. Suppose $q = p^\Pi$ for some partition $\Pi \in \mathcal{P}(Y)$. Then,

$$p(S) - q(S) = \sum_{B \in \Pi} p(S \cap B) - q(S \cap B)$$

$$= \sum_B p(S \cap B) - \frac{p(B)}{|B|}|S \cap B|$$

$$\leq \sum_B p(S \cap B) - \frac{p(S \cap B)}{|B|}|S \cap B|$$

$$= \sum_B p(S \cap B)\frac{|B| - |S \cap B|}{|B|}$$

$$= \sum_B p(S \cap B)\frac{|B \setminus S|}{|B|}$$

$$\leq \sum_B \frac{p(S \cap B)}{|S \cap B|}|B \setminus S|.$$

Since $a \leq b \Rightarrow (a)_+ \leq (b)_+$ and this last quantity is non-negative,

$$(p(S) - q(S))_+ \leq \sum_B \frac{p(S \cap B)}{|S \cap B|}|B \setminus S|$$

$$\mathbb{E}_\nu\left[(p(S) - q(S))_+\right] \leq \sum_B \mathbb{E}_\nu\left[\frac{p(S \cap B)}{|S \cap B|}\right]|B \setminus S|$$

$$\leq \sum_B |B \setminus S| \max_{y \in S} \mathbb{E}_\nu[p(y)]$$

$$= |Y \setminus S| \max_{y \in S} \mathbb{E}_\nu[p(y)].$$

$\square$

We are no ready to prove Theorem 1.

Proof of Theorem 1. Let $q = p^\Pi$. By the definition of TV,

$$g(U) \geq q(U) - \|q - g\|_{\mathrm{TV}}. \tag{6}$$

Since $Y \setminus F = H \subseteq U$, we have $H = U \setminus (F \cap U)$ and,

$$
\begin{aligned}
g(H) &= g(U) - g(F \cap U) \\
&\geq q(U) - \|q - g\|_{\mathrm{TV}} - g(F \cap U) &&\text{by Eq. (6)} \\
&= p(U) - (p(U) - q(U)) - \|q - g\|_{\mathrm{TV}} - g(F \cap U) \\
&= p(U) - \|q - g\|_{\mathrm{TV}} - \big(p(U) - q(U) + g(F \cap U)\big).
\end{aligned}
$$

Rearranging terms gives $p(U) - g(H) - \|q - g\|_{\mathrm{TV}} \leq p(U) - q(U) + g(F \cap U)$. Applying $a \leq b \Rightarrow (a)_+ \leq (b)_+$,

$$(p(U) - g(H) - \|q - g\|_{\mathrm{TV}})_+ \leq (p(U) - q(U) + g(F \cap U))_+ \leq (p(U) - q(U))_+ + g(F \cap U).$$

15

Thus, it suffices to prove:

$$\mathbb{E}_{\nu}\left[g(F \cap U) + \big(p(U) - q(U)\big)_+\right] \le \max_{y \in U} \Pr_{\nu}[y \in F] + |O| \max_{y \in U} \mathbb{E}_{\nu}[p(y)]. \tag{7}$$

To this end, linearity of expectation implies that,

$$\mathbb{E}_{\nu}[g(F \cap U)] = \sum_{y \in U} g(y) \Pr_{\nu}[y \in F] \le \max_{y \in U} \Pr_{\nu}[y \in F]. \tag{8}$$

By Lemma 2 (with $S = U$),

$$\mathbb{E}_{\nu}\left[\big(p(U) - q(U)\big)_+\right] \le |O| \max_{y \in U} \mathbb{E}_{\nu}[p(y)].$$

Combining this with Eq. (8) gives Eq. (7), as required. □

7 Upper bounds on hallucination rate

Could one prove a much better lower bound than $\widehat{MF}$? If not in general, are there better lower bounds under various assumptions on D_{world}? In this section, we argue that a significantly better lower-bound is not possible by showing that for any D_{world}, there is a algorithm that is calibrated and hallucinates at a rate near the missing facts rate, equivalent its Good-Turing estimate. The conceptually simple algorithm is neither efficient nor a good LM, but it suffices to show that a better lower-bound is not possible.

The algorithm memorizes the facts in the training data and generates random facts from the training data and random unseen factoids uniformly at random, but at different rates.

1. Inputs: $\mathbf{x}_{\text{train}} \in X^n$.

2. Let $O, U = Y \setminus O$ be the sets of observed and unobserved factoids in the training data, respectively, as defined in Eq. (4) and compute $\widehat{MF}$, the fraction of factoids that appear exactly once in the training data, as defined in Eq. (5). Let $g \in \Delta(Y)$ be defined as

$$g(y) := \begin{cases} \frac{\widehat{MF}}{|U|} & \text{if } y \in U, \\ \frac{1 - \widehat{MF}}{|O|} & \text{if } y \in O. \end{cases}$$

3. For each factoid $y \in Y$, select a corresponding document $d(y) \in X$ such that $f(d(y)) = y$. To be specific, one can take $d(y)$ to be the lexicographically first document in $\{x \in X \mid f(x) = y\}$.

4. Output $D_{LM} = d \circ g$, i.e., the distribution which samples $y \sim g$ and then outputs $d(y)$.

It is easy to see that, by design, $g = f \circ D_{LM}$.

Lemma 3. *For any $\delta, \lambda \in [0,1], D_{\text{world}}, n \ge 1$,*

$$\Pr_{\mathbf{x}_{\text{train}} \sim \mathcal{D}_{\text{train}}}\left[g(H) \le \widehat{MF} \text{ and } \mathrm{Mis}_\infty(g, p) \le 3\sqrt{\frac{\ln(4/\delta)}{n}}\right] \ge 1 - \delta,$$

where $g = f \circ D_{LM}$ for the above algorithm, $g(H)$ is its hallucination rate, and $\mathrm{Mis}_\infty(g, p)$ is defined in Eq. (2).

In other words, with high probability one can have a well-calibrated LM that hallucinates with probability close to $\widehat{MF}$.

Proof. There can either be one or two bins B_z^g based on whether or not $\frac{\widehat{MF}}{|U|} = \frac{1-\widehat{MF}}{|O|}$. If they are equal then $\text{Mis}_\infty(g, p) = 0$. In any case,

$$\text{Mis}_\infty(g, p) \leq \frac{1}{2}|p(U) - g(U)| + \frac{1}{2}|p(O) - g(O)| = |p(U) - g(U)| = |p(U) - \widehat{MF}|.$$

Since $\widehat{MF}$ is a Good-Turing estimator, by Corollary 6, with probability $\geq 1 - \delta$ the above quantity is at most $3\sqrt{\frac{\ln(4/\delta)}{n}}$. At the same time, with certainty,

$$g(H) = \frac{\widehat{MF}}{|U|}|H \cap U| \leq \widehat{MF}.$$

$\square$

8 Conclusions, limitations, and future work

When one first encounters LM hallucinations, it is perhaps surprising that a system which clearly embeds such a rich diverse array of detailed knowledge at the same time creates complete fabrications with no basis in fact or the training data. This work aims to demystify this phenomenon by showing that pretraining LMs for predictive accuracy leads to hallucination even in an ideal world where the training data is perfectly factual, there is no blur between facts and hallucinations, each document contains at most one fact, and there is not even a prompt that would encourage hallucination. Moreover, our theory explains why modern LMs hallucinate more than older LMs such as trigram models, even though both are trained on similar types of data with similar objectives.

The monofact rate may shed light on the rates at which calibrated LMs must hallucinate for different types of facts. One expects hallucination for facts that have a high monofact rate, i.e., the types of facts which often occur just once in the training data. Interestingly, this would not be common for references to books or articles, a problematic type of hallucination discussed today. Therefore, these may arise from other issues such as model capacity, when one considers the shear number of facts including references and others that an LM encounters in training. Furthermore, correcting for hallucinated references may be doable by modifying the pre-training pipeline without post-training, though this will not address other types of arbitrary facts where the monofacts are common, as in our 5W example.

There are several limitations to our work. First, we only study one statistical source of hallucination. There are many other types of hallucination and reasons LMs may hallucinate beyond pure statistics. Second, our semantic notion of calibration is different from the standard token-based ones used in classification. While natural and simple to define, our notion has the disadvantage of being computationally intractable to evaluate for many models. Third, factuality is not always clear-cut, facts are not all disjoint, and our regularity assumptions may not hold for facts that have a mild systematic component. As an example, if the training data contains the Alex Wilkins 5W fact from the introduction, then it is also follows that Alex Wilkins has eaten at Salumeria at some point, which is a different but overlapping fact. Finally, it could be the case that aspects of the real world, messy and different from our idealized setting, actually reduce the minimal hallucination

rates. For instance, it could be that having multiple facts in a document makes models less likely to hallucinate and thus our lower bounds do not apply.

In future work, it would be interesting to use the insights presented here to further reduce hallucination in LMs. An interesting question is how to convert a pretrained (calibrated) model to one that is good at factual prediction. A step in this process may be to distinguish systematic facts from arbitrary ones, which LMs may be capable of at some point in the future if not today. For example, for generation, one would not desire fabricated book titles, but one would like mathematical statements. What is the difference between fabricating a non-existent book title from generating an inequality such as $17 < 124398$, if neither has ever been written down? Humans know that the latter can be manufactured (as long as it is mathematically correct) while the former cannot, which presumably is how we avoid hallucinating. It seems conceivable that LMs can similarly represent this distinction, and the work mentioned showing that LMs "know" when they are hallucinating suggests that this may indeed be possible with today's LMs.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. Do Language Models Know When They're Hallucinating References? http://arxiv.org/abs/2305.18248 arXiv:2305.18248 [cs].

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing Attribution and Fluency Tradeoffs for Retrieval-Augmented Large Language Models. http://arxiv.org/abs/2302.05578 arXiv:2302.05578 [cs].

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When its Lying. https://doi.org/10.48550/arXiv.2304.13734 arXiv:2304.13734 [cs].

Daniel Berend and Aryeh Kontorovich. 2013. On the concentration of the missing mass. *Electronic Communications in Probability* 18, none (Jan. 2013), 1–7. https://doi.org/10.1214/ECP.v18-2359 Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Joseph R. Jr. Biden. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. http://arxiv.org/abs/2112.04426 arXiv:2112.04426 [cs].

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. 2020. Calibration, Entropy Rates, and Memory in Language Models. In *Proceedings of the 37th*

International Conference on Machine Learning. PMLR, 1089–1099. `https://proceedings.mlr.press/v119/braverman20a.html` ISSN: 2640-3498.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. 2023. Loss minimization yields multicalibration for large neural networks. `https://doi.org/10.48550/arXiv.2304.09424` arXiv:2304.09424 [cs, stat].

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311* (2022). `https://arxiv.org/abs/2204.02311` Publisher: arXiv _eprint: 2204.02311.

A. P. Dawid. 1982. The Well-Calibrated Bayesian. *J. Amer. Statist. Assoc.* 77, 379 (Sept. 1982), 605–610. `https://doi.org/10.1080/01621459.1982.10477856`

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? arXiv. `https://doi.org/10.48550/ARXIV.2204.07931` Version Number: 1.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. `https://doi.org/10.48550/arXiv.2110.06674` arXiv:2110.06674 [cs].

Vitaly Feldman. 2021. Does Learning Require Memorization? A Short Tale about a Long Tail. `http://arxiv.org/abs/1906.05271` arXiv:1906.05271 [cs, stat].

Katja Filippova. 2020. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. `http://arxiv.org/abs/2010.05873` arXiv:2010.05873 [cs].

I. J. Good. 1953. The Population Frequences of Species and the Estimation of Population Parameters. *Biometrika* 40, 3-4 (Dec. 1953), 237–264. `https://doi.org/10.1093/biomet/40.3-4.237`

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. `http://arxiv.org/abs/2311.05232` arXiv:2311.05232 [cs].

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. `https://doi.org/10.1145/3571730`

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics* 9 (Sept. 2021), 962–977. `https://doi.org/10.1162/tacl_a_00407`

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. `https://doi.org/10.48550/arXiv.2207.05221` arXiv:2207.05221 [cs].

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

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. `http://arxiv.org/abs/2303.08896` arXiv:2303.08896 [cs].

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. `https://doi.org/10.18653/v1/2020.acl-main.173`

David McAllester and Robert E Schapire. 2000. On the Convergence Rate of Good-Turing Estimators.

David A. McAllester and Luis E. Ortiz. 2003. Concentration Inequalities for the Missing Mass and for Histogram Rule Error. `https://www.semanticscholar.org/paper/Concentration-Inequalities-for-the-Missing-Mass-and-McAllester-Ortiz/d9379f1b3542f47454da4336173596b9008f642c`

Michelle M. Mello and Neel Guha. 2023. ChatGPT and Physicians' Malpractice Risk. *JAMA Health Forum* 4, 5 (May 2023), e231938. `https://doi.org/10.1001/jamahealthforum.2023.1938`

Merriam-Webster. 2023. Definition of HALLUCINATION. `https://www.merriam-webster.com/dictionary/hallucination`

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. 2020. Measuring Calibration in Deep Learning. `https://doi.org/10.48550/arXiv.1904.01685` arXiv:1904.01685 [cs, stat].

OpenAI. 2023. GPT-4 Technical Report. `http://arxiv.org/abs/2303.08774` arXiv:2303.08774 [cs].

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 4035–4045. `https://doi.org/10.18653/v1/D18-1437`

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. 2023. SlimPajama-DC: Understanding Data Combinations for LLM Training. `http://arxiv.org/abs/2309.10818` arXiv:2309.10818 [cs].

Rachel Shin. 2023. Humiliated lawyers fined \$5,000 for submitting ChatGPT hallucinations in court: 'I heard about this new site, which I falsely assumed was, like, a super search engine'. *Fortune* (June 2023). `https://fortune.com/2023/06/23/lawyers-fined-filing-chatgpt-hallucinations-in-court/`

Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P. Xing, and Zhiting Hu. 2021. Progressive Generation of Long Text with Pretrained Language Models. `http://arxiv.org/abs/2006.15720` arXiv:2006.15720 [cs].

Kumiko Tanaka-Ishii. 2007. Word-based predictive text entry using adaptive language models. *Natural Language Engineering* 13, 1 (March 2007), 51–74. `https://doi.org/10.1017/S1351324905004080` Publisher: Cambridge University Press.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. `http://arxiv.org/abs/2310.03214` arXiv:2310.03214 [cs].

Karen Weise and Cade Metz. 2023. When A.I. Chatbots Hallucinate. *The New York Times* (May 2023). `https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html`

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection. *Transactions of the Association for Computational Linguistics* 11 (June 2023), 546–564. `https://doi.org/10.1162/tacl_a_00563`

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023b. How Language Model Hallucinations Can Snowball. `http://arxiv.org/abs/2305.13534` arXiv:2305.13534 [cs].

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023a. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. `http://arxiv.org/abs/2309.01219` arXiv:2309.01219 [cs].

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 12697–12706. `https://proceedings.mlr.press/v139/zhao21c.html` ISSN: 2640-3498.

A   Good-Turing estimator bounds

The distribution bounds here are stated for a general set S, i.i.d. sample $s := (s_1, s_2, \ldots, s_n) \in S^n$ from an arbitrary distribution $\mathcal{D} \in \Delta(S)$.

Definition 5 (Missing mass). *For distribution $\mathcal{D} \in \Delta(S)$, $n \geq 1$, and sample $s \in S^n$, the missing mass is:*

$$M_{\mathcal{D}}(s) := \mathcal{D}\left(S \setminus \{s_1, s_2 \ldots, s_n\}\right).$$

Theorem 4 ((Berend and Kontorovich, 2013; McAllester and Ortiz, 2003)). *For any distribution $\mathcal{D} \in \Delta(S)$ and any $n \geq 1$, $\varepsilon \geq 0$, let $\overline{M}_{\mathcal{D}}^n := \mathbb{E}_{s \sim \mathcal{D}^n}[M_{\mathcal{D}}(s)]$. Then:*

$$\Pr_{s \sim \mathcal{D}^n}[M_{\mathcal{D}}(s) \geq \overline{M}_{\mathcal{D}}^n + \varepsilon] \leq e^{-\varepsilon^2 n} \tag{9}$$

$$\Pr_{s \sim \mathcal{D}^n}[M_{\mathcal{D}}(s) \leq \overline{M}_{\mathcal{D}}^n - \varepsilon] \leq e^{-1.92\varepsilon^2 n} \tag{10}$$

Eq. (9) is Theorem 16 of McAllester and Ortiz (2003) and Theorem 2.1 of Berend and Kontorovich (2013). Eq. (10) is Theorem 2.2 of Berend and Kontorovich (2013), though for simplicity we use the worse bound of $e^{-\varepsilon^2 n}$ which was also present in McAllester and Ortiz (2003).

Definition 6 (Good Turing estimator). *For $n \geq 1$, set S, and sample $s \in S^n$, the Good-Turing estimator (Good, 1953) is,*

$$GT(s) := \frac{1}{n}\left|\{i \in [n] \mid \forall j \neq i \ s_i \neq s_j\}\right|.$$

In words, the estimator above is defined as the fraction of elements of a sample each of which appears exactly once in the sample.

Lemma 5 ((McAllester and Ortiz, 2003)). *For any distribution $\mathcal{D} \in \Delta(S)$ and any $n \geq 1$, $\delta \in (0, 1]$, let $\overline{GT}_{\mathcal{D}}^n := \mathbb{E}_{s \sim \mathcal{D}^n}[GT_{\mathcal{D}}(s)]$. Then:*

$$\Pr_{s \sim \mathcal{D}^n}\left[GT_{\mathcal{D}}(s) \geq \overline{GT}_{\mathcal{D}}^n + \sqrt{\frac{2\ln 1/\delta}{n}}\right] \leq \delta \tag{11}$$

$$\Pr_{s \sim \mathcal{D}^n}\left[GT_{\mathcal{D}}(s) \leq \overline{GT}_{\mathcal{D}}^n - \sqrt{\frac{2\ln 1/\delta}{n}}\right] \leq \delta \tag{12}$$

Eq. (11) is Theorem 16 of McAllester and Ortiz (2003) and Eq. (12) has the identical 1-line proof using McDiarmid's inequality.

The next lemma says that the expected values of the missing mass and unique elements in training data are very close.

Lemma 6. *For any $n \geq 1$ and any $\mathcal{D} \in \Delta(S)$,*

$$\overline{M}_{\mathcal{D}}^n \leq \overline{GT}_{\mathcal{D}}^n \leq \overline{M}_{\mathcal{D}}^n + \frac{1}{n}$$

for $\overline{M}_{\mathcal{D}}^n$ as defined in Theorem 4 and $\overline{GT}_{\mathcal{D}}^n$ as defined in Lemma 5.

Corollary 5. *For any set S, distribution $\mathcal{D} \in \Delta(S)$, any $n \geq 1, \delta \in (0,1]$,*

$$\Pr_{s \sim \mathcal{D}^n} \left[|M_{\mathcal{D}}(s) - GT(s)| \leq \frac{1}{n} + 2.42\sqrt{\frac{\ln(4/\delta)}{n}} \right] \geq 1 - \delta. \tag{13}$$

$$\Pr_{s \sim \mathcal{D}^n} \left[M_{\mathcal{D}}(s) \geq GT(s) - \frac{1}{n} - 2.14\sqrt{\frac{\ln(2/\delta)}{n}} \right] \geq 1 - \delta. \tag{14}$$

Proof. Eq. (13) is established by setting $\varepsilon = \sqrt{\ln(4/\delta)/n}$ in Eqs. (9) and (10), which by the union bound implies,

$$\Pr_{s \sim \mathcal{D}^n} \left[\left| M_{\mathcal{D}}(s) - \overline{M}_{\mathcal{D}}^n \right| \geq \sqrt{\frac{\ln(4/\delta)}{n}} \right] \leq \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}.$$

Plugging $\delta' = \delta/4$ in Eqs. (11) and (12) and the union bound gives,

$$\Pr_{s \sim \mathcal{D}^n} \left[\left| GT(s) - \overline{GT}_{\mathcal{D}}^n \right| \geq \sqrt{\frac{2\ln(4/\delta)}{n}} \right] \leq \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}.$$

Combining the above two with Lemma 6, the triangle inequality, and the fact that $1 + \sqrt{2} \leq 2.41$ give Eq. (13).

Similarly, Eq. (14) follows by using $\varepsilon = \sqrt{\ln(2/\delta)/(1.92n)}$ in Eq. (10) and Lemma 6 and Eq. (11) with $\delta/2$ and summing the corresponding three inequalities to give,

$$\Pr \left[M_{\mathcal{D}}(s) \geq GT(s) - \frac{1}{n} - \sqrt{\frac{\ln(2/\delta)}{1.92n}} - \sqrt{\frac{2\ln(2/\delta)}{n}} \right] \leq \frac{\delta}{2} + \frac{\delta}{2} \leq \delta.$$

Using the fact that $\sqrt{1/1.92} + \sqrt{2} < 2.14$ completes the proof. $\qquad\square$

We now simplify the above expression.

Corollary 6. *For any set S, distribution $\mathcal{D} \in \Delta(S)$, any $n \geq 1$,*

$$\forall \delta \in (0,1] \quad \Pr_{s \sim \mathcal{D}^n} \left[|M_{\mathcal{D}}(s) - GT(s)| \leq 3\sqrt{\frac{\ln(4/\delta)}{n}} \right] \geq 1 - \delta. \tag{15}$$

$$\forall \delta \in (0, 1/3] \quad \Pr_{s \sim \mathcal{D}^n} \left[M_{\mathcal{D}}(s) \geq GT(s) - \sqrt{\frac{6\ln(2/\delta)}{n}} \right] \geq 1 - \delta. \tag{16}$$

Proof. We first show Eq. (15). Note that Eq. (15) holds trivially for $n \leq 9$ because $GT(s), M_{\mathcal{D}}(s) \in [0,1]$ and $3\sqrt{\ln(4)/9} > 1$. Thus, from Cor. 5, it suffices to verify that for $n > 9, \delta \leq 1$:

$$\frac{1}{n} + 2.42\sqrt{\frac{\ln(4/\delta)}{n}} \leq 3\sqrt{\frac{\ln(4/\delta)}{n}}$$

In other words, we need

$$0.58\sqrt{\frac{\ln(4/\delta)}{n}} \geq \frac{1}{n}.$$

23

Squaring and simplifying, this is

$$n \geq \frac{1}{(0.58)^2 \ln(4/\delta)}.$$

Since the RHS is a monotonic increasing function of δ, we can use its largest value of $\delta = 1$, and it suffices to have $n > 4.29$.

For Eq. (16), note that it holds trivially for $n \leq 6$ because $\sqrt{6\ln(2/\delta)/n} \geq \sqrt{6\ln(6)/6} > 1$. Thus, from Cor. 5, it suffices to verify that for $n > 6, \delta \leq 1/3$:

$$\frac{1}{n} + 2.14\sqrt{\frac{\ln(2/\delta)}{n}} \leq \sqrt{\frac{6\ln(2/\delta)}{n}}$$

In other words, we need

$$(\sqrt{6} - 2.14)\sqrt{\frac{\ln(2/\delta)}{n}} \geq \frac{1}{n}.$$

Squaring and simplifying, this is

$$n \geq \frac{1}{(\sqrt{6} - 2.14)^2 \ln(2/\delta)}.$$

Since the RHS is a monotonic increasing function of δ, we can use its largest value of $\delta = 1/3$, and it suffices to have $n > 5.83$. $\qquad\square$

B Generalizations and alternatives

There are alternative models one could consider. Here we discuss alternatives in terms of log-loss and prompts, as well as fixed-width (non-adaptive) binning strategies and other notions of calibration.

B.1 Hallucination with Prompts

Many uses of LMs require text to be generated conditionally based on a prefix string called a prompt. The degree of hallucination will heavily depend on the distribution of prompts and their lengths. Here we observe how prompts can be an additional source of hallucinations, but at the same time they can also make hallucinations unnecessary, depending on their distributions. The analysis we have performed earlier covers unconditional generation in which there are no prompts, which is like zero-length prompts.

First, observe that a certain distribution over prompts can make hallucination unnecessary. In particular, if the prompts themselves are very long, i.e., complete documents (ending with an end-of-document token if there is one), distributed exactly as D_{LM}, then any D_{LM} which offers empty string completions will give statistically perfect completions and never hallucinate. More generally, this (arguably good) situation occurs whenever the facts are contained in the prompts and not their completions.

Second, note that out-of-distribution (OOD) prompts can lead to significantly more hallucinations. For any LM, one can consider the single worst (adversarial) prompt a which leads to the worst hallucination rate. (In the introduction we gave the example of the prompt *The 15 Elves of San Salami are named....*) The prompt distribution could be concentrated on this single prompt in which case the hallucination rate would be terrible, assuming there is at least one adversarial prompt.

B.2 KL-divergence and log-loss

Clearly not all LMs hallucinate. Examples include an LM that avoids memorizing and regurgitating the training data or even an LM that only output the word *yay* with probability 1. Such LMs would score poorly under predictive measures such as log-likelihood on held-out data, but it is not clear that LMs that perform well under log-likelihood must hallucinate. This suggests that one may perform the analysis in terms of log-loss $\mathbb{E}_{x \sim D_{\text{train}}}[-\log D_{LM}(x)]$, since LMs are generally trained to minimize this loss. Equivalently, one can consider the KL divergence which is a non-negative quantity that measures how far the log-loss is from its minimum achievable value, the entropy of D_{train}. $\mathbb{E}_{x \sim D_{\text{train}}}[\log D_{\text{train}}(x) - \log D_{LM}(x)]$.

One advantage of calibration analysis over KL-divergence is that, from a statistical perspective ignoring computation, one can generally expect to achieve calibration (miscalibration close to 0), e.g., by generating a uniformly random factoid. In contrast, one cannot expect to achieve KL divergence close to 0. Additionally, one can take any D_{LM} which hallucinates and convert it to a model which does not without significantly increasing its KL divergence by, for example, mixing it with a model which regurgitates the training distribution. Specifically, consider a model D_{LM} which, with probability 99% outputs the word *yay* and with probability 1% outputs $x \sim D_{LM}$. This model hallucinates with probability $< 1\%$. Furthermore, the log-loss this new model at most $\log 100$ bits larger than that of D_{LM}. This difference is small difference on the scale of the entropy of documents, especially for longer documents since entropy typically scales linearly with the document length.

Nonetheless, it would be interesting to see if one can quantify hallucination rates purely on the basis of accuracy rather than calibration. A starting point may be the work of Feldman (2021) which shows that statistical models must memorize their training data for accuracy purposes.[5]

B.3 Alternative calibration definitions

In this section, we define a more standard alternative definition of calibration based on log-probability bins of equal width. Recall that $B_z := \{y \in Y \mid g(y) = z\}$ and $B_I := \{y \in Y \mid g(y) \in I\}$. We now define bins of fixed width in probability space.

Definition 7 (Binning). *For $\varepsilon \in (0,1)$, the binning $\mathcal{B}(g, \varepsilon)$ with equally spaced bins in log-probability space, is the following partition:*

$$\mathcal{B}(g, \varepsilon) := \left\{ B_{((1-\varepsilon)^{i+1}, (1-\varepsilon)^i]} \,\middle|\, i = 0, 1, 2, \ldots \right\} \cup \{B_0\}. \tag{17}$$

For $\varepsilon \in \{0, 1\}$, let $\mathcal{B}(g, 0) := \mathcal{B}(g) = \{B_z \mid z \in [0,1]\}$ and $\mathcal{B}(g, 1) := \{B_{[0,1]}\} = \{Y\}$.

Thus ε determines the bid widths on a log-scale, with small ε corresponding to narrow bins. Thus one could use $\left\| p^{\mathcal{B}(g, \varepsilon)} - g \right\|_{\text{TV}}$ as a definition of miscalibration and the corresponding corollary would follow directly from our previous analysis.

Corollary 7. *Fix any $\delta \in [0,1], n \in \mathbb{N}, s \in \mathbb{R}, \varepsilon \in (0,1)$ and any s-sparse regular D_{world}. Then for any algorithm $\mathcal{A} : X^n \to \Delta(X)$, with probability $\geq 1 - \delta$ over $D_L \sim D_{\text{world}}$ and $\mathbf{x}_{\text{train}} \sim D_L^{\times n}$,*

$$g(H) \geq \widehat{MF} - \left\| p^{\mathcal{B}(g, \varepsilon)} - g \right\|_{\text{TV}} - \frac{3e^{-s}}{\delta} - \sqrt{\frac{6 \ln(6/\delta)}{n}}.$$

[5]We acknowledge Gautam Kamath for pointing out this connection.

Proof. The proof of this Corollary follows exactly the same proofs as that of Corollary 1 except that we use $p^{\mathcal{B}(g,\varepsilon)}$ in place of $\mathcal{V}_b(g)$. $\qquad\square$

We next use an even more standard definition which is not based on statistical distance. Recall that our first definition of calibration, without limits on bins as in Eq. (2), can be written as,

$$\text{Mis}_\infty(g,p) := \left\|p^{\mathcal{B}(g)} - g\right\|_{\text{TV}} = \frac{1}{2}\sum_{B\in\mathcal{B}(g)}\sum_{y\in B}\left|\frac{p(B)}{B} - g(y)\right| = \frac{1}{2}\sum_{B\in\mathcal{B}(g)}|p(B) - g(B)|.$$

This is the most obvious definition and the question is how to generalize it to bins. The above also suggests the following alternative generative definition of calibration error.

Definition 8 (Generative calibration error). *For $\varepsilon \in [0,1]$, and distributions $p,g \in \Delta(Y)$, the ε-generative calibration error is,*

$$\text{GCE}_\varepsilon(g,p) := \frac{1}{2}\sum_{B\in\mathcal{B}(g,\varepsilon)}\left|p(B) - g(B)\right|.$$

This definition means that $\text{GCE}_0(g,p) = \text{Mis}_\infty(g,p)$. Note that these two definitions of calibration error are related by the following lemma.

Lemma 7. *Let $\varepsilon \in [0,1]$. Then,*

$$\left\|p^{\mathcal{B}(g,\varepsilon)} - g\right\|_{\text{TV}} - \varepsilon \leq \text{GCE}_\varepsilon(g,p) \leq \left\|p^{\mathcal{B}(g,\varepsilon)} - g\right\|_{\text{TV}}.$$

Before we prove Lemma 7, we observe that Corollary 7 implies the following.

Corollary 8. *Fix any $\delta \in [0,1], n \in \mathbb{N}, s \in \mathbb{R}, \varepsilon \in (0,1)$ and any s-sparse regular D_{world}. Then for any algorithm $\mathcal{A}: X^n \to \Delta(X)$, with probability $\geq 1 - \delta$ over $D_L \sim D_{\text{world}}$ and $\mathbf{x}_{\text{train}} \sim D_L^{\times n}$,*

$$g(H) \geq \widehat{MF} - \text{GCE}_\varepsilon(g,p) - \varepsilon - \frac{3e^{-s}}{\delta} - \sqrt{\frac{6\ln(6/\delta)}{n}}.$$

We now return to prove Lemma 7.

Proof of Lemma 7. Let $\Pi = \mathcal{B}(g,\varepsilon)$. Then,

$$\begin{aligned}
\text{GCE}_\varepsilon(g,p) &= \frac{1}{2}\sum_{B\in\Pi}|p(B) - g(B)| \\
&= \frac{1}{2}\sum_{B\in\Pi}\sum_{y\in B}\frac{1}{|B|}|p(B) - g(B)| \\
&= \frac{1}{2}\sum_{B\in\Pi}\sum_{y\in B}\left|\frac{p(B)}{|B|} - \frac{g(B)}{|B|}\right| \\
&= \frac{1}{2}\sum_{B\in\Pi}\sum_{y\in B}\left|p^\Pi(y) - g^\Pi(y)\right| \\
&= \frac{1}{2}\sum_{B\in\Pi}\sum_{y\in B}\left|p^\Pi(y) - g(y) + g(y) - g^\Pi(y)\right| \\
&\geq \frac{1}{2}\sum_{B\in\Pi}\sum_{y\in B}\left|p^\Pi(y) - g(y)\right| - \left|g^\Pi(y) - g(y)\right| \qquad \text{by } |a+b| \geq |b| - |a| \\
&= \left\|p^\Pi - g\right\|_{\text{TV}} - \left\|g^\Pi - g\right\|_{\text{TV}}.
\end{aligned}$$

This proves the RHS inequality of the lemma. Thus it suffices to show $\left\|g^{\Pi} - g\right\|_{\mathrm{TV}} \leq \varepsilon$. We first claim that for all y:

$$g^{\Pi}(y) - g(y) \leq \varepsilon g^{\Pi}(y). \tag{18}$$

Let $B \in \Pi$ be the bin containing $y \in B$. Now, recall that each bucket can be written as:

$$B_I^g = \left\{y \ \middle| \ g(y) \in I\right\} \text{ for some interval } I \subseteq [0, 1].$$

If $I = [0, 0]$, then Eq. (18) is trivially true because $g(y) = 0 = g^{\Pi}(y)$. Otherwise, say $I = ((1 - \varepsilon)^{(}i + 1), (1 - \varepsilon)^i]$ for some $i \geq 0$. Then, by definition of g^{Π},

$$g^{\Pi}(y) = \frac{g(B)}{|B|} \in I,$$

because the weighted average of an numbers in an interval is also contained in the interval. Since this interval has (multiplicative) width $e^{-\varepsilon}$, $g(y) \geq (1 - \varepsilon)g^{\Pi}(y)$. Equivalently,

$$g^{\Pi}(y) - g(y) \leq \varepsilon g^{\Pi}(y).$$

Thus we have established Eq. (18) which trivially implies that,

$$\forall y \in Y \ \left(g^{\Pi}(y) - g(y)\right)_+ \leq \varepsilon g^{\Pi}(y).$$

Therefore,

$$\left\|g^{\Pi} - g\right\|_{\mathrm{TV}} = \sum_{y \in Y} \left(g^{\Pi}(y) - g(y)\right)_+ \leq \varepsilon \sum_{y \in Y} g^{\Pi}(y) = \varepsilon,$$

which is all that remained to prove the lemma. □

C Proofs of remaining corollaries

To prove Corollaries 1 to 3, we use Markov's inequality on the expectation of a non-negative random variable W, which states that $\Pr[W \geq \mathbb{E}[W]/\delta] \leq \delta$. In our case, $W := (p(U) - \left\|p^{\Pi} - g\right\|_{\mathrm{TV}} - g(H))_+$ and thus using $\delta \to (2/3)\delta$ and Theorem 1 imply that for any partition Π:

$$\Pr_{p \sim \nu}\left[p(U) - \left\|p^{\Pi} - g\right\|_{\mathrm{TV}} - g(H) \geq \frac{3}{2\delta}\left(\max_{y \in U} \Pr_{p \sim \nu}[y \in F] + |O| \max_{y \in U} \mathbb{E}_\nu[p(y)]\right)\right] \leq \frac{2\delta}{3}.$$

Rearranging terms,

$$\Pr_{p \sim \nu}\left[g(H) \geq p(U) - \left\|p^{\Pi} - g\right\|_{\mathrm{TV}} - \frac{3}{2\delta}\left(\max_{y \in U} \Pr_{p \sim \nu}[y \in F] + |O| \max_{y \in U} \mathbb{E}_\nu[p(y)]\right)\right] \geq 1 - \frac{2\delta}{3}. \tag{19}$$

Also, Corollary 6 of Appendix A with $\delta \to \delta/3$ implies that for any $\delta \in (0, 1]$ and any D_L,

$$\Pr_{\mathbf{x}_{\mathrm{train}} \sim D_L^{\times n}}\left[p(U) \geq \widehat{MF} - \sqrt{\frac{6 \ln(6/\delta)}{n}}\right] \geq 1 - \frac{\delta}{3}. \tag{20}$$

It is now straightforward to prove the corollaries.

Proof of Corollary 1. For a regular D_{world}, because it has 1-regular-facts and 1-regular-probabilities, with probability 1 the posterior satisfies:

$$\max_{y \in U} \Pr_{p \sim \nu}[y \in F] + |O| \max_{y \in U} \mathbb{E}_\nu[p(y)] \leq \frac{\mathbb{E}[|F \cap U|]}{|U|} + \frac{|O|}{|U|} \mathbb{E}[p(U)] \leq 2\frac{|F|}{|U|} \leq 2e^{-s}.$$

In the above we have used the fact that $O \subseteq F$ and $U \supseteq H$. The proof follows immediately from this, Eqs. (19) and (20) and the union bound, using $\Pi = \mathcal{V}_b(g)$. $\qquad \square$

The proofs of Corollary 2 and Corollary 3 also follow directly, where in Corollary 2 we use the fact that $|O| \leq n$.