

Hallucination, Calibration, and Missing Mass

Why Error-Free Training Data Do Not Guarantee Error-Free
Answers

Aaron Roth

University of Pennsylvania

Fall 2026

Overview

Why can a model trained on true statements still make false claims?

- ▶ The sample can omit facts that cannot be inferred from other observations. Abstaining avoids the errors that guessing would make.

Overview

Why can a model trained on true statements still make false claims?

- ▶ The sample can omit facts that cannot be inferred from other observations. Abstaining avoids the errors that guessing would make.
- ▶ But pretraining aims to imitate a source that gives answers, not one that says “I do not know” when the learner lacks information.

Overview

Why can a model trained on true statements still make false claims?

- ▶ The sample can omit facts that cannot be inferred from other observations. Abstaining avoids the errors that guessing would make.
- ▶ But pretraining aims to imitate a source that gives answers, not one that says “I do not know” when the learner lacks information.
- ▶ We will show that even matching the source’s probabilities in the aggregate—a calibration requirement—can force hallucinations.

Overview

Why can a model trained on true statements still make false claims?

- ▶ The sample can omit facts that cannot be inferred from other observations. Abstaining avoids the errors that guessing would make.
- ▶ But pretraining aims to imitate a source that gives answers, not one that says “I do not know” when the learner lacks information.
- ▶ We will show that even matching the source’s probabilities in the aggregate—a calibration requirement—can force hallucinations.
- ▶ The chance of being asked an unseen fact controls this tradeoff. Facts seen only once help us estimate that chance.

Error-Free Data Need Not Contain Every Fact

Consider prompts of the form

What is this person's birthday? Respond in the format MM-DD.

Error-Free Data Need Not Contain Every Fact

Consider prompts of the form

What is this person's birthday? Respond in the format MM-DD.

For a specified person, there are 365 syntactically plausible answers and only one correct answer.

Error-Free Data Need Not Contain Every Fact

Consider prompts of the form

What is this person's birthday? Respond in the format MM-DD.

For a specified person, there are 365 syntactically plausible answers and only one correct answer.

Seeing many other people's birthdays need not reveal this one. No error in the training data is required; the relevant fact may simply be absent.

Error-Free Data Need Not Contain Every Fact

Consider prompts of the form

What is this person's birthday? Respond in the format MM-DD.

For a specified person, there are 365 syntactically plausible answers and only one correct answer.

Seeing many other people's birthdays need not reveal this one. No error in the training data is required; the relevant fact may simply be absent.

The sample can be error-free without being complete.

Unseen Facts Cannot Be Inferred from Other Facts

Let $\mathcal{X}$ be a finite set of prompts with population distribution μ . Fix an integer $K \geq 1$, and give every prompt the same set of factual answers

$$\mathcal{R} = \{1, \dots, K + 1\}.$$

Unseen Facts Cannot Be Inferred from Other Facts

Let $\mathcal{X}$ be a finite set of prompts with population distribution μ . Fix an integer $K \geq 1$, and give every prompt the same set of factual answers

$$\mathcal{R} = \{1, \dots, K + 1\}.$$

For each $x \in \mathcal{X}$, one answer $a_x \in \mathcal{R}$ is correct and the other K are incorrect. Before training, nature draws these answers once:

$(a_x)_{x \in \mathcal{X}}$ independently and uniformly from $\mathcal{R}$.

Unseen Facts Cannot Be Inferred from Other Facts

Let $\mathcal{X}$ be a finite set of prompts with population distribution μ . Fix an integer $K \geq 1$, and give every prompt the same set of factual answers

$$\mathcal{R} = \{1, \dots, K + 1\}.$$

For each $x \in \mathcal{X}$, one answer $a_x \in \mathcal{R}$ is correct and the other K are incorrect. Before training, nature draws these answers once:

$$(a_x)_{x \in \mathcal{X}} \quad \text{independently and uniformly from } \mathcal{R}.$$

Thus learning $a_{x'}$ gives no information about a_x for $x \neq x'$. This models arbitrary facts, not arithmetic, logical consequences, or any fact derivable or predictable from related observations.

Missing Mass Measures How Often the Sample Has No Answer

The training sample contains $n \geq 1$ error-free examples:

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mu, \quad a_{X_1}, \dots, a_{X_n}.$$

Write $O_n = \{X_1, \dots, X_n\}$ for the prompts that appeared.

Missing Mass Measures How Often the Sample Has No Answer

The training sample contains $n \geq 1$ error-free examples:

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mu, \quad a_{X_1}, \dots, a_{X_n}.$$

Write $O_n = \{X_1, \dots, X_n\}$ for the prompts that appeared.

Definition

The *missing mass* is

$$M_n := \mu(\mathcal{X} \setminus O_n) = \mathbb{P}_{X \sim \mu}(X \notin O_n).$$

Missing Mass Measures How Often the Sample Has No Answer

The training sample contains $n \geq 1$ error-free examples:

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mu, \quad a_{X_1}, \dots, a_{X_n}.$$

Write $O_n = \{X_1, \dots, X_n\}$ for the prompts that appeared.

Definition

The *missing mass* is

$$M_n := \mu(\mathcal{X} \setminus O_n) = \mathbb{P}_{X \sim \mu}(X \notin O_n).$$

It is a population quantity: the probability that a fresh prompt asks for a fact that was absent from training.

Forced Answers Fail on the Missing Mass

Theorem

Suppose an algorithm sees only the training sample, the current test prompt, and its private randomness, and must output one answer in $\mathcal{R}$ on every prompt. Conditional on the realized training prompts,

$$\mathbb{E}[\mathbb{P}_{X \sim \mu}(\hat{a}(X) \neq a_X) \mid X_1, \dots, X_n] \geq \left(1 - \frac{1}{K+1}\right) M_n.$$

The expectation is over the random arbitrary facts and the algorithm's private randomness.

Forced Answers Fail on the Missing Mass

Theorem

Suppose an algorithm sees only the training sample, the current test prompt, and its private randomness, and must output one answer in $\mathcal{R}$ on every prompt. Conditional on the realized training prompts,

$$\mathbb{E}[\mathbb{P}_{X \sim \mu}(\hat{a}(X) \neq a_X) \mid X_1, \dots, X_n] \geq \left(1 - \frac{1}{K+1}\right) M_n.$$

The expectation is over the random arbitrary facts and the algorithm's private randomness. For birthdays, $K = 364$: on an unseen person, an answer-only system is wrong with probability $364/365$.

An Unseen Answer Is Still Uniform

Fix an unseen prompt $x \notin O_n$.

An Unseen Answer Is Still Uniform

Fix an unseen prompt $x \notin O_n$.

- ▶ The training sample contains no observation of a_x .
- ▶ The arbitrary-facts model makes a_x uniform on the $K + 1$ answers, independently of everything the algorithm has seen.

An Unseen Answer Is Still Uniform

Fix an unseen prompt $x \notin O_n$.

- ▶ The training sample contains no observation of a_x .
- ▶ The arbitrary-facts model makes a_x uniform on the $K + 1$ answers, independently of everything the algorithm has seen.

Therefore, whatever distribution the algorithm uses for its answer,

$$\mathbb{P}(\hat{a}(x) = a_x \mid \text{training sample}) = \frac{1}{K + 1}.$$

Its conditional error is $1 - 1/(K + 1)$.

An Unseen Answer Is Still Uniform

Fix an unseen prompt $x \notin O_n$.

- ▶ The training sample contains no observation of a_x .
- ▶ The arbitrary-facts model makes a_x uniform on the $K + 1$ answers, independently of everything the algorithm has seen.

Therefore, whatever distribution the algorithm uses for its answer,

$$\mathbb{P}(\widehat{a}(x) = a_x \mid \text{training sample}) = \frac{1}{K + 1}.$$

Its conditional error is $1 - 1/(K + 1)$.

A fresh prompt is unseen with probability M_n . Keeping only the error on unseen prompts proves the theorem.

Abstention Breaks the Answer-Only Lower Bound

The theorem assumes the system had to provide an answer.

Abstention Breaks the Answer-Only Lower Bound

The theorem assumes the system had to provide an answer.

Add a new output,

$$\mathcal{Y} := \mathcal{R} \cup \{\perp\}, \quad \perp = \textit{I do not know}.$$

A system can now memorize observed facts and output $\perp$ on every unseen prompt. It makes no false factual claim.

Abstention Breaks the Answer-Only Lower Bound

The theorem assumes the system had to provide an answer.
Add a new output,

$$\mathcal{Y} := \mathcal{R} \cup \{\perp\}, \quad \perp = \textit{I do not know}.$$

A system can now memorize observed facts and output $\perp$ on every unseen prompt. It makes no false factual claim.

Missing information does not by itself force
hallucination once abstention is available. We need to
ask what objective the model is trained to satisfy.

Pretraining Introduces a Different Goal

Recall our error-free training source: draw $X \sim \mu$ and report its correct answer a_X . Write D for this same source distribution. With the correct answers fixed,

$$D(r \mid x) = \mathbf{1}\{r = a_x\}, \quad D(\perp \mid x) = 0.$$

Pretraining aims to match D using a generator G learned from the sample and private randomness. On prompt x , it draws $R \sim G(\cdot \mid x)$ on $\mathcal{Y} = \mathcal{R} \cup \{\perp\}$.

Pretraining Introduces a Different Goal

Recall our error-free training source: draw $X \sim \mu$ and report its correct answer a_X . Write D for this same source distribution. With the correct answers fixed,

$$D(r \mid x) = \mathbf{1}\{r = a_x\}, \quad D(\perp \mid x) = 0.$$

Pretraining aims to match D using a generator G learned from the sample and private randomness. On prompt x , it draws $R \sim G(\cdot \mid x)$ on $\mathcal{Y} = \mathcal{R} \cup \{\perp\}$.

Its hallucination rate counts incorrect factual answers, but not abstention:

$$\text{err}_{\text{gen}}(G) := \mathbb{P}_{\substack{X \sim \mu \\ R \sim G(\cdot \mid X)}} (R \in \mathcal{R} \setminus \{a_X\}).$$

Pretraining Introduces a Different Goal

Recall our error-free training source: draw $X \sim \mu$ and report its correct answer a_X . Write D for this same source distribution. With the correct answers fixed,

$$D(r \mid x) = \mathbf{1}\{r = a_x\}, \quad D(\perp \mid x) = 0.$$

Pretraining aims to match D using a generator G learned from the sample and private randomness. On prompt x , it draws $R \sim G(\cdot \mid x)$ on $\mathcal{Y} = \mathcal{R} \cup \{\perp\}$.

Its hallucination rate counts incorrect factual answers, but not abstention:

$$\text{err}_{\text{gen}}(G) := \mathbb{P}_{\substack{X \sim \mu \\ R \sim G(\cdot \mid X)}} (R \in \mathcal{R} \setminus \{a_X\}).$$

Always abstaining makes $\text{err}_{\text{gen}}(G) = 0$, but mismatches the answer-giving population D . How much distribution matching can we retain without making false claims?

Two Ways to Handle an Unseen Birthday

Return to the birthday example: 365 possible dates, one correct date, and M_n probability of being asked about someone absent from training.

Both generators memorize the birthdays that appeared in the sample.

	unseen person	hallucination rate
Abstain	output $\perp$	0
Guess	uniformly sample a date	$\frac{364}{365} M_n$

Two Ways to Handle an Unseen Birthday

Return to the birthday example: 365 possible dates, one correct date, and M_n probability of being asked about someone absent from training.

Both generators memorize the birthdays that appeared in the sample.

	unseen person	hallucination rate
Abstain	output $\perp$	0
Guess	uniformly sample a date	$\frac{364}{365} M_n$

Neither generator has learned any missing birthday. But only the second always gives a date, as the source text does.

Two Ways to Handle an Unseen Birthday

Return to the birthday example: 365 possible dates, one correct date, and M_n probability of being asked about someone absent from training.

Both generators memorize the birthdays that appeared in the sample.

	unseen person	hallucination rate
Abstain	output $\perp$	0
Guess	uniformly sample a date	$\frac{364}{365} M_n$

Neither generator has learned any missing birthday. But only the second always gives a date, as the source text does.

In fact, the guessing generator passes a natural calibration test.

Calibration Matches Groups, Not Each Fact

For a candidate response r to prompt x , the probability $G(r \mid x)$ predicts that the source answers r . We calibrate these response probabilities, not a separate confidence score attached to an already-generated answer.

Calibration Matches Groups, Not Each Fact

For a candidate response r to prompt x , the probability $G(r \mid x)$ predicts that the source answers r . We calibrate these response probabilities, not a separate confidence score attached to an already-generated answer. Weighting prompts by μ defines the joint laws

$$G(x, r) := \mu(x)G(r \mid x), \quad D(x, r) := \mu(x)D(r \mid x).$$

Calibration Matches Groups, Not Each Fact

For a candidate response r to prompt x , the probability $G(r \mid x)$ predicts that the source answers r . We calibrate these response probabilities, not a separate confidence score attached to an already-generated answer. Weighting prompts by μ defines the joint laws

$$G(x, r) := \mu(x)G(r \mid x), \quad D(x, r) := \mu(x)D(r \mid x).$$

For each prediction value $p \in [0, 1]$, group pairs assigned that value:

$$B_p := \{(x, r) : G(r \mid x) = p\}.$$

Exact *semantic calibration* requires $G(B_p) = D(B_p)$ for every p : each group has the same total mass under the generator and source.

Uniform Guessing Can Be Semantically Calibrated

For the birthday guesser, grouping pairs by $G(r \mid x)$ gives:

prediction value	generator mass	source mass
1: observed correct answers	$1 - M_n$	$1 - M_n$
1/365: all dates on unseen prompts	M_n	M_n
0: all other pairs	0	0

Uniform Guessing Can Be Semantically Calibrated

For the birthday guesser, grouping pairs by $G(r \mid x)$ gives:

prediction value	generator mass	source mass
1: observed correct answers	$1 - M_n$	$1 - M_n$
1/365: all dates on unseen prompts	M_n	M_n
0: all other pairs	0	0

Every group matches, yet the hallucination rate is $\frac{364}{365} M_n$.

Matching a group's *total* mass does not locate the correct date inside it.

Abstention Avoids Errors but Breaks Calibration

Now consider the abstaining generator. Recall that O_n is the set of observed prompts. This generator assigns probability 1 to a_x when $x \in O_n$, and to $\perp$ when $x \notin O_n$.

Abstention Avoids Errors but Breaks Calibration

Now consider the abstaining generator. Recall that O_n is the set of observed prompts. This generator assigns probability 1 to a_x when $x \in O_n$, and to $\perp$ when $x \notin O_n$. Thus its prediction-value group is

$$\begin{aligned} B_1 &= \{(x, r) : G(r \mid x) = 1\} \\ &= \underbrace{\{(x, a_x) : x \in O_n\}}_{\text{memorized correct answers}} \cup \underbrace{\{(x, \perp) : x \notin O_n\}}_{\text{certain abstentions}}. \end{aligned}$$

Abstention Avoids Errors but Breaks Calibration

Now consider the abstaining generator. Recall that O_n is the set of observed prompts. This generator assigns probability 1 to a_x when $x \in O_n$, and to $\perp$ when $x \notin O_n$. Thus its prediction-value group is

$$\begin{aligned} B_1 &= \{(x, r) : G(r \mid x) = 1\} \\ &= \underbrace{\{(x, a_x) : x \in O_n\}}_{\text{memorized correct answers}} \cup \underbrace{\{(x, \perp) : x \notin O_n\}}_{\text{certain abstentions}}. \end{aligned}$$

The generator always outputs a pair in this set, so $G(B_1) = 1$.

Abstention Avoids Errors but Breaks Calibration

Now consider the abstaining generator. Recall that O_n is the set of observed prompts. This generator assigns probability 1 to a_x when $x \in O_n$, and to $\perp$ when $x \notin O_n$. Thus its prediction-value group is

$$\begin{aligned} B_1 &= \{(x, r) : G(r \mid x) = 1\} \\ &= \underbrace{\{(x, a_x) : x \in O_n\}}_{\text{memorized correct answers}} \cup \underbrace{\{(x, \perp) : x \notin O_n\}}_{\text{certain abstentions}}. \end{aligned}$$

The generator always outputs a pair in this set, so $G(B_1) = 1$. The source D always supplies a_x . Its output belongs to B_1 on observed prompts, but not on unseen prompts, where B_1 contains $(x, \perp)$ instead. Hence

$$D(B_1) = \mu(O_n) = 1 - M_n.$$

Abstention Avoids Errors but Breaks Calibration

Now consider the abstaining generator. Recall that O_n is the set of observed prompts. This generator assigns probability 1 to a_x when $x \in O_n$, and to $\perp$ when $x \notin O_n$. Thus its prediction-value group is

$$\begin{aligned} B_1 &= \{(x, r) : G(r \mid x) = 1\} \\ &= \underbrace{\{(x, a_x) : x \in O_n\}}_{\text{memorized correct answers}} \cup \underbrace{\{(x, \perp) : x \notin O_n\}}_{\text{certain abstentions}}. \end{aligned}$$

The generator always outputs a pair in this set, so $G(B_1) = 1$. The source D always supplies a_x . Its output belongs to B_1 on observed prompts, but not on unseen prompts, where B_1 contains $(x, \perp)$ instead. Hence

$$D(B_1) = \mu(O_n) = 1 - M_n.$$

There are no false claims, but the mass mismatch on B_1 is M_n .

A Weaker Calibration Requirement

Exact calibration matches the source's mass at every prediction value. We will show that matching just one combined group still creates a tradeoff with hallucination. Let H collect all prompt–response pairs assigned probability greater than $1/K$:

$$H := \{(x, r) \in \mathcal{X} \times \mathcal{Y} : G(r \mid x) > 1/K\}, \quad L := H^c.$$

H combines the disjoint groups B_p with $p > 1/K$. Exact calibration therefore implies $G(H) = D(H)$.

A Weaker Calibration Requirement

Exact calibration matches the source's mass at every prediction value. We will show that matching just one combined group still creates a tradeoff with hallucination. Let H collect all prompt-response pairs assigned probability greater than $1/K$:

$$H := \{(x, r) \in \mathcal{X} \times \mathcal{Y} : G(r \mid x) > 1/K\}, \quad L := H^c.$$

H combines the disjoint groups B_p with $p > 1/K$. Exact calibration therefore implies $G(H) = D(H)$. For approximate matching, measure the discrepancy directly:

$$\delta_H(G, D) := |G(H) - D(H)|.$$

A Weaker Calibration Requirement

Exact calibration matches the source's mass at every prediction value. We will show that matching just one combined group still creates a tradeoff with hallucination. Let H collect all prompt–response pairs assigned probability greater than $1/K$:

$$H := \{(x, r) \in \mathcal{X} \times \mathcal{Y} : G(r \mid x) > 1/K\}, \quad L := H^c.$$

H combines the disjoint groups B_p with $p > 1/K$. Exact calibration therefore implies $G(H) = D(H)$. For approximate matching, measure the discrepancy directly:

$$\delta_H(G, D) := |G(H) - D(H)|.$$

In the birthday example, the guesser has $\delta_H = 0$, while the abstainer has $\delta_H = M_n$.

A Weaker Calibration Requirement

Exact calibration matches the source's mass at every prediction value. We will show that matching just one combined group still creates a tradeoff with hallucination. Let H collect all prompt–response pairs assigned probability greater than $1/K$:

$$H := \{(x, r) \in \mathcal{X} \times \mathcal{Y} : G(r \mid x) > 1/K\}, \quad L := H^c.$$

H combines the disjoint groups B_p with $p > 1/K$. Exact calibration therefore implies $G(H) = D(H)$. For approximate matching, measure the discrepancy directly:

$$\delta_H(G, D) := |G(H) - D(H)|.$$

In the birthday example, the guesser has $\delta_H = 0$, while the abstainer has $\delta_H = M_n$.

This is also one outcome-indistinguishability test: does a source or generated response fall in H ?

Missing Mass Forces Hallucination or Miscalibration

Theorem

In the arbitrary-facts model, any generator learned from the training sample and private randomness satisfies

$$\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D) \mid X_1, \dots, X_n] \geq M_n - \frac{2}{K}.$$

The expectation is over the arbitrary facts and learner randomness.

Missing Mass Forces Hallucination or Miscalibration

Theorem

In the arbitrary-facts model, any generator learned from the training sample and private randomness satisfies

$$\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D) \mid X_1, \dots, X_n] \geq M_n - \frac{2}{K}.$$

The expectation is over the arbitrary facts and learner randomness. Keeping expected calibration error small forces a large expected hallucination rate whenever $M_n - 2/K$ is large, even though abstention is allowed.

Missing Mass Forces Hallucination or Miscalibration

Theorem

In the arbitrary-facts model, any generator learned from the training sample and private randomness satisfies

$$\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D) \mid X_1, \dots, X_n] \geq M_n - \frac{2}{K}.$$

The expectation is over the arbitrary facts and learner randomness. Keeping expected calibration error small forces a large expected hallucination rate whenever $M_n - 2/K$ is large, even though abstention is allowed.

Proof idea: few hallucinations and small calibration error would give an accurate test for truth. But no such test can beat chance on unseen arbitrary facts.

A Validity Test Connects Generation to Missing Information

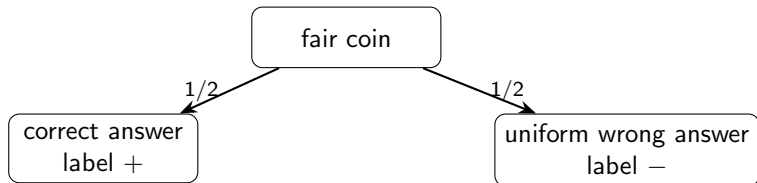
Consider an *is-it-valid* classification experiment:

1. Draw a prompt $X \sim \mu$ and flip a fair coin.
2. On the positive branch, propose $R = a_X$ with label $+$.
3. On the negative branch, propose R uniformly from the K incorrect factual answers, with label $-$.

A Validity Test Connects Generation to Missing Information

Consider an *is-it-valid* classification experiment:

1. Draw a prompt $X \sim \mu$ and flip a fair coin.
2. On the positive branch, propose $R = a_X$ with label $+$.
3. On the negative branch, propose R uniformly from the K incorrect factual answers, with label $-$.



This is a *thought experiment*, not extra training data or a procedure the learner must run. It never proposes $\perp$.

On an Unseen Prompt, Both Labels Look the Same

Fix the training sample and an unseen prompt x . Its correct answer a_x is still uniform on $\mathcal{R}$. For any proposed answer r ,

$$\mathbb{P}(R = r \mid +) = \mathbb{P}(a_x = r) = \frac{1}{K+1}.$$

On an Unseen Prompt, Both Labels Look the Same

Fix the training sample and an unseen prompt x . Its correct answer a_x is still uniform on $\mathcal{R}$. For any proposed answer r ,

$$\mathbb{P}(R = r \mid +) = \mathbb{P}(a_x = r) = \frac{1}{K+1}.$$

On the negative branch, r can be proposed for any correct answer $a \neq r$:

$$\mathbb{P}(R = r \mid -) = \sum_{a \neq r} \mathbb{P}(a_x = a) \frac{1}{K} = K \cdot \frac{1}{K+1} \cdot \frac{1}{K} = \frac{1}{K+1}.$$

On an Unseen Prompt, Both Labels Look the Same

Fix the training sample and an unseen prompt x . Its correct answer a_x is still uniform on $\mathcal{R}$. For any proposed answer r ,

$$\mathbb{P}(R = r \mid +) = \mathbb{P}(a_x = r) = \frac{1}{K+1}.$$

On the negative branch, r can be proposed for any correct answer $a \neq r$:

$$\mathbb{P}(R = r \mid -) = \sum_{a \neq r} \mathbb{P}(a_x = a) \frac{1}{K} = K \cdot \frac{1}{K+1} \cdot \frac{1}{K} = \frac{1}{K+1}.$$

The two labels give *identical distributions* of proposed answers. The label remains a fair coin after seeing the proposal, so any classifier using only the training sample and proposal has error $1/2$.

Use Generation Probabilities to Predict Validity

We now use the generator G to define a classifier for the validity-testing problem. Given a prompt x and a proposed factual answer r , predict “valid” if $G(r \mid x) > 1/K$, and “invalid” otherwise. Thus the classifier accepts pairs in H and rejects pairs in L .

Use Generation Probabilities to Predict Validity

We now use the generator G to define a classifier for the validity-testing problem. Given a prompt x and a proposed factual answer r , predict “valid” if $G(r \mid x) > 1/K$, and “invalid” otherwise. Thus the classifier accepts pairs in H and rejects pairs in L .

Let $\text{err}_{\text{iiv}}(G)$ be this classifier’s error in the balanced experiment. The unseen-prompt lower bound now applies to this rule:

$$\mathbb{E}[\text{err}_{\text{iiv}}(G) \mid X_1, \dots, X_n] \geq \frac{1}{2} M_n.$$

Use Generation Probabilities to Predict Validity

We now use the generator G to define a classifier for the validity-testing problem. Given a prompt x and a proposed factual answer r , predict “valid” if $G(r \mid x) > 1/K$, and “invalid” otherwise. Thus the classifier accepts pairs in H and rejects pairs in L .

Let $\text{err}_{\text{iiv}}(G)$ be this classifier’s error in the balanced experiment. The unseen-prompt lower bound now applies to this rule:

$$\mathbb{E}[\text{err}_{\text{iiv}}(G) \mid X_1, \dots, X_n] \geq \frac{1}{2} M_n.$$

To prove the theorem, it therefore suffices to show, for each fixed D, G ,

$$\text{err}_{\text{gen}}(G) + \delta_H(G, D) \geq 2 \text{err}_{\text{iiv}}(G) - \frac{2}{K}.$$

Accepted falsehoods give a direct bound on generated falsehoods; rejected truths give a bound through calibration.

Where Can the Generator Put Its Probability?

We compare classifier errors with generated falsehoods using this partition of prompt–response pairs:

	H : probability $> 1/K$	L : probability $\leq 1/K$
correct	accepted truth	rejected truth
incorrect	accepted falsehood	rejected falsehood
abstain	not validity-tested	not validity-tested

Let $\mathcal{E}$ be the shaded row of incorrect factual pairs. The generator's error is $\text{err}_{\text{gen}}(G) = G(\mathcal{E})$.

Where Can the Generator Put Its Probability?

We compare classifier errors with generated falsehoods using this partition of prompt–response pairs:

	H : probability $> 1/K$	L : probability $\leq 1/K$
correct	accepted truth	rejected truth
incorrect	accepted falsehood	rejected falsehood
abstain	not validity-tested	not validity-tested

Let $\mathcal{E}$ be the shaded row of incorrect factual pairs. The generator's error is $\text{err}_{\text{gen}}(G) = G(\mathcal{E})$.

The classifier errs by accepting a falsehood (a *false positive*) or rejecting a truth (a *false negative*). Their probabilities $p_{\text{FP}}, p_{\text{FN}}$ include the fair coin's choice of branch, and

$$\text{err}_{\text{iiv}}(G) = p_{\text{FP}} + p_{\text{FN}}.$$

Accepted Falsehoods Already Have Generated Mass

First consider false positives: incorrect factual pairs in $H \cap \mathcal{E}$. For one such pair (x, r) , its probability in the validity test is

$$\underbrace{\frac{1}{2}}_{\text{negative branch}} \cdot \underbrace{\mu(x)}_{\text{prompt}} \cdot \underbrace{\frac{1}{K}}_{\text{one of } K \text{ wrong answers}} = \frac{\mu(x)}{2K}.$$

Accepted Falsehoods Already Have Generated Mass

First consider false positives: incorrect factual pairs in $H \cap \mathcal{E}$. For one such pair (x, r) , its probability in the validity test is

$$\underbrace{\frac{1}{2}}_{\text{negative branch}} \cdot \underbrace{\mu(x)}_{\text{prompt}} \cdot \underbrace{\frac{1}{K}}_{\text{one of } K \text{ wrong answers}} = \frac{\mu(x)}{2K}.$$

Because the pair belongs to H , its generator probability satisfies

$$G(x, r) = \mu(x)G(r \mid x) \geq \frac{\mu(x)}{K}.$$

Accepted Falsehoods Already Have Generated Mass

First consider false positives: incorrect factual pairs in $H \cap \mathcal{E}$. For one such pair (x, r) , its probability in the validity test is

$$\underbrace{\frac{1}{2}}_{\text{negative branch}} \cdot \underbrace{\mu(x)}_{\text{prompt}} \cdot \underbrace{\frac{1}{K}}_{\text{one of } K \text{ wrong answers}} = \frac{\mu(x)}{2K}.$$

Because the pair belongs to H , its generator probability satisfies

$$G(x, r) = \mu(x)G(r \mid x) \geq \frac{\mu(x)}{K}.$$

Because of the threshold $1/K$: each accepted falsehood has at least twice as much probability under generation as under the validity test. Summing over these pairs gives

$$\boxed{G(H \cap \mathcal{E}) \geq 2p_{\text{FP}}.}$$

Rejected Truths Require Generated Mass in L

Now consider false negatives: correct answers in L . The positive branch has weight $1/2$, and D generates exactly the correct answers, so

$$p_{\text{FN}} = \frac{1}{2}D(L).$$

Rejected Truths Require Generated Mass in L

Now consider false negatives: correct answers in L . The positive branch has weight $1/2$, and D generates exactly the correct answers, so

$$p_{\text{FN}} = \frac{1}{2}D(L).$$

Because $L = H^c$ and G, D are probability distributions,

$$|G(L) - D(L)| = |G(H) - D(H)| = \delta_H(G, D).$$

Taking the lower bound and then adding $D(L)$ gives

$$\begin{aligned} G(L) - D(L) &\geq -\delta_H(G, D), \\ G(L) &\geq D(L) - \delta_H(G, D) = 2p_{\text{FN}} - \delta_H(G, D). \end{aligned}$$

Rejected Truths Require Generated Mass in L

Now consider false negatives: correct answers in L . The positive branch has weight $1/2$, and D generates exactly the correct answers, so

$$p_{\text{FN}} = \frac{1}{2}D(L).$$

Because $L = H^c$ and G, D are probability distributions,

$$|G(L) - D(L)| = |G(H) - D(H)| = \delta_H(G, D).$$

Taking the lower bound and then adding $D(L)$ gives

$$\begin{aligned} G(L) - D(L) &\geq -\delta_H(G, D), \\ G(L) &\geq D(L) - \delta_H(G, D) = 2p_{\text{FN}} - \delta_H(G, D). \end{aligned}$$

Thus a high false negative rate requires high generated probability in L , unless calibration fails.

Separating Incorrect Answers in L

We have shown that

$$G(L) \geq 2p_{\text{FN}} - \delta_H(G, D).$$

$G(L)$ is the probability of generating a pair in L , whether its response is incorrect, correct, or an abstention. To obtain the probability of generating an incorrect answer in L , subtract the probability of generating a correct answer or abstention in L .

Separating Incorrect Answers in L

We have shown that

$$G(L) \geq 2p_{\text{FN}} - \delta_H(G, D).$$

$G(L)$ is the probability of generating a pair in L , whether its response is incorrect, correct, or an abstention. To obtain the probability of generating an incorrect answer in L , subtract the probability of generating a correct answer or abstention in L . Correct answers and abstentions form $\mathcal{E}^c$, so

$$G(L \cap \mathcal{E}) = G(L) - G(L \cap \mathcal{E}^c).$$

We already have a lower bound on $G(L)$. We now need an upper bound on $G(L \cap \mathcal{E}^c)$, the probability being subtracted.

Only Two Outputs per Prompt Avoid Hallucination

For a fixed prompt x , the only pairs in $\mathcal{E}^c$ are (x, a_x) and $(x, \perp)$. If either pair lies in L , its conditional probability under $G(\cdot \mid x)$ is at most $1/K$.

Only Two Outputs per Prompt Avoid Hallucination

For a fixed prompt x , the only pairs in $\mathcal{E}^c$ are (x, a_x) and $(x, \perp)$. If either pair lies in L , its conditional probability under $G(\cdot \mid x)$ is at most $1/K$. Thus the conditional probability of a correct answer or abstention in L is at most $2/K$. Averaging over prompts gives

$$G(L \cap \mathcal{E}^c) \leq \sum_x \mu(x) \frac{2}{K} = \frac{2}{K}.$$

Only Two Outputs per Prompt Avoid Hallucination

For a fixed prompt x , the only pairs in $\mathcal{E}^c$ are (x, a_x) and $(x, \perp)$. If either pair lies in L , its conditional probability under $G(\cdot \mid x)$ is at most $1/K$. Thus the conditional probability of a correct answer or abstention in L is at most $2/K$. Averaging over prompts gives

$$G(L \cap \mathcal{E}^c) \leq \sum_x \mu(x) \frac{2}{K} = \frac{2}{K}.$$

Combining this upper bound with $G(L) \geq 2p_{\text{FN}} - \delta_H(G, D)$ gives

$$\begin{aligned} G(L \cap \mathcal{E}) &= G(L) - G(L \cap \mathcal{E}^c) \\ &\geq 2p_{\text{FN}} - \delta_H(G, D) - \frac{2}{K}. \end{aligned}$$

Adding the Two Parts Proves the Theorem

Recall $\text{err}_{\text{iiv}}(G) = p_{\text{FP}} + p_{\text{FN}}$. Split the incorrect pairs between H and L , and apply the two bounds:

$$\begin{aligned}\text{err}_{\text{gen}}(G) &= G(H \cap \mathcal{E}) + G(L \cap \mathcal{E}) \\ &\geq 2p_{\text{FP}} + 2p_{\text{FN}} - \delta_H(G, D) - \frac{2}{K} \\ &= 2\text{err}_{\text{iiv}}(G) - \delta_H(G, D) - \frac{2}{K}.\end{aligned}$$

Adding the Two Parts Proves the Theorem

Recall $\text{err}_{\text{iiv}}(G) = p_{\text{FP}} + p_{\text{FN}}$. Split the incorrect pairs between H and L , and apply the two bounds:

$$\begin{aligned}\text{err}_{\text{gen}}(G) &= G(H \cap \mathcal{E}) + G(L \cap \mathcal{E}) \\ &\geq 2p_{\text{FP}} + 2p_{\text{FN}} - \delta_H(G, D) - \frac{2}{K} \\ &= 2\text{err}_{\text{iiv}}(G) - \delta_H(G, D) - \frac{2}{K}.\end{aligned}$$

Now keep the training prompts fixed and average over the random facts and learner randomness. The expected classifier error is at least $M_n/2$, so

$$\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D) \mid X_1, \dots, X_n] \geq M_n - \frac{2}{K}. \quad \square$$

Adding the Two Parts Proves the Theorem

Recall $\text{err}_{\text{iiv}}(G) = p_{\text{FP}} + p_{\text{FN}}$. Split the incorrect pairs between H and L , and apply the two bounds:

$$\begin{aligned}\text{err}_{\text{gen}}(G) &= G(H \cap \mathcal{E}) + G(L \cap \mathcal{E}) \\ &\geq 2p_{\text{FP}} + 2p_{\text{FN}} - \delta_H(G, D) - \frac{2}{K} \\ &= 2\text{err}_{\text{iiv}}(G) - \delta_H(G, D) - \frac{2}{K}.\end{aligned}$$

Now keep the training prompts fixed and average over the random facts and learner randomness. The expected classifier error is at least $M_n/2$, so

$$\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D) \mid X_1, \dots, X_n] \geq M_n - \frac{2}{K}. \quad \square$$

With large missing mass, matching the source's probability of H forces a large expected hallucination rate.

Reweight High-Probability Responses

Why should *population* likelihood favor $G(H) = D(H)$?

Reweight High-Probability Responses

Why should *population* likelihood favor $G(H) = D(H)$?

Keep the set H defined by the original G fixed during the adjustment. On prompt x , its current probability is

$$h_x := \sum_{r:(x,r) \in H} G(r \mid x)$$

Multiply probabilities in H by $s > 0$, leaving the others unchanged. The resulting weights have total

$$\underbrace{sh_x}_{\text{weight in } H} + \underbrace{(1 - h_x)}_{\text{weight outside } H} = 1 + (s - 1)h_x.$$

Reweight High-Probability Responses

Why should *population* likelihood favor $G(H) = D(H)$?

Keep the set H defined by the original G fixed during the adjustment. On prompt x , its current probability is

$$h_x := \sum_{r:(x,r) \in H} G(r \mid x)$$

Multiply probabilities in H by $s > 0$, leaving the others unchanged. The resulting weights have total

$$\underbrace{sh_x}_{\text{weight in } H} + \underbrace{(1 - h_x)}_{\text{weight outside } H} = 1 + (s - 1)h_x.$$

Divide every response's weight by this total to obtain the new generator:

$$G_s(r \mid x) := \frac{s \mathbf{1}_{\{(x,r) \in H\}} G(r \mid x)}{1 + (s - 1)h_x}.$$

At $s = 1$ this is G ; increasing s favors responses in H .

Correcting a Mismatch Improves Population Likelihood

Measure fit by population negative log-likelihood:

$$\mathcal{L}(s) := \mathbb{E}_{(X,R) \sim D}[-\log G_s(R \mid X)].$$

Assume $\mathcal{L}(1) < \infty$. Taking negative logs of the formula for G_s and averaging gives

$$\mathcal{L}(s) = \mathcal{L}(1) - D(H) \log s + \mathbb{E}_{X \sim \mu} [\log(1 + (s - 1)h_X)].$$

Correcting a Mismatch Improves Population Likelihood

Measure fit by population negative log-likelihood:

$$\mathcal{L}(s) := \mathbb{E}_{(X,R) \sim D}[-\log G_s(R \mid X)].$$

Assume $\mathcal{L}(1) < \infty$. Taking negative logs of the formula for G_s and averaging gives

$$\mathcal{L}(s) = \mathcal{L}(1) - D(H) \log s + \mathbb{E}_{X \sim \mu}[\log(1 + (s - 1)h_X)].$$

Differentiate at $s = 1$:

$$\mathcal{L}'(1) = -D(H) + \mathbb{E}_{X \sim \mu}[h_X] = G(H) - D(H).$$

Correcting a Mismatch Improves Population Likelihood

Measure fit by population negative log-likelihood:

$$\mathcal{L}(s) := \mathbb{E}_{(X,R) \sim D}[-\log G_s(R \mid X)].$$

Assume $\mathcal{L}(1) < \infty$. Taking negative logs of the formula for G_s and averaging gives

$$\mathcal{L}(s) = \mathcal{L}(1) - D(H) \log s + \mathbb{E}_{X \sim \mu}[\log(1 + (s - 1)h_X)].$$

Differentiate at $s = 1$:

$$\mathcal{L}'(1) = -D(H) + \mathbb{E}_{X \sim \mu}[h_X] = G(H) - D(H).$$

If $G(H) > D(H)$, slightly decreasing s lowers the loss; if $G(H) < D(H)$, slightly increasing s lowers it. Thus if neither change improves the loss, $G(H) = D(H)$.

Correcting a Mismatch Improves Population Likelihood

Measure fit by population negative log-likelihood:

$$\mathcal{L}(s) := \mathbb{E}_{(X,R) \sim D}[-\log G_s(R | X)].$$

Assume $\mathcal{L}(1) < \infty$. Taking negative logs of the formula for G_s and averaging gives

$$\mathcal{L}(s) = \mathcal{L}(1) - D(H) \log s + \mathbb{E}_{X \sim \mu}[\log(1 + (s - 1)h_X)].$$

Differentiate at $s = 1$:

$$\mathcal{L}'(1) = -D(H) + \mathbb{E}_{X \sim \mu}[h_X] = G(H) - D(H).$$

If $G(H) > D(H)$, slightly decreasing s lowers the loss; if $G(H) < D(H)$, slightly increasing s lowers it. Thus if neither change improves the loss, $G(H) = D(H)$.

This conclusion requires access to this adjustment and optimization of *population* loss. Fitting the finite training sample alone does not establish it.

But How Large Is the Missing Mass?

Recall that M_n is the probability that a fresh prompt did not appear in the n training examples. We want to estimate this probability from the sample.

But How Large Is the Missing Mass?

Recall that M_n is the probability that a fresh prompt did not appear in the n training examples. We want to estimate this probability from the sample.

For each prompt x , let

$$N_x := \sum_{j=1}^n \mathbf{1}\{X_j = x\}$$

be its sample count. A prompt with $N_x = 1$ is a *singleton*.

But How Large Is the Missing Mass?

Recall that M_n is the probability that a fresh prompt did not appear in the n training examples. We want to estimate this probability from the sample.

For each prompt x , let

$$N_x := \sum_{j=1}^n \mathbf{1}\{X_j = x\}$$

be its sample count. A prompt with $N_x = 1$ is a *singleton*. The Good–Turing estimate is the fraction of the n sample positions occupied by singleton prompts:

$$\hat{M}_n := \frac{1}{n} \sum_{x \in \mathcal{X}} \mathbf{1}\{N_x = 1\}.$$

Each singleton prompt contributes one position; repeated prompts contribute none. Why should this estimate the chance that the next prompt is unseen?

A Singleton Is New When We Hold It Out

The prompt X_j is a singleton exactly when it is absent from the other $n - 1$ positions. Counting singletons by position gives

$$\hat{M}_n = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{X_j \notin \{X_i : i \neq j\}\}.$$

A Singleton Is New When We Hold It Out

The prompt X_j is a singleton exactly when it is absent from the other $n - 1$ positions. Counting singletons by position gives

$$\hat{M}_n = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{X_j \notin \{X_i : i \neq j\}\}.$$

Fix j and condition on the other $n - 1$ prompts. Since $X_j \sim \mu$ independently, its probability of being new is their missing mass.

A Singleton Is New When We Hold It Out

The prompt X_j is a singleton exactly when it is absent from the other $n - 1$ positions. Counting singletons by position gives

$$\hat{M}_n = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{X_j \notin \{X_i : i \neq j\}\}.$$

Fix j and condition on the other $n - 1$ prompts. Since $X_j \sim \mu$ independently, its probability of being new is their missing mass. Let M_{n-1} be the missing mass of $n - 1$ independent draws. Averaging over the other prompts and then over positions gives

$$\begin{aligned} \mathbb{E}[\hat{M}_n] &= \frac{1}{n} \sum_{j=1}^n \mathbb{P}(X_j \notin \{X_i : i \neq j\}) \\ &= \frac{1}{n} \sum_{j=1}^n \mathbb{E}[M_{n-1}] = \mathbb{E}[M_{n-1}]. \end{aligned}$$

A Singleton Is New When We Hold It Out

The prompt X_j is a singleton exactly when it is absent from the other $n - 1$ positions. Counting singletons by position gives

$$\hat{M}_n = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{X_j \notin \{X_i : i \neq j\}\}.$$

Fix j and condition on the other $n - 1$ prompts. Since $X_j \sim \mu$ independently, its probability of being new is their missing mass. Let M_{n-1} be the missing mass of $n - 1$ independent draws. Averaging over the other prompts and then over positions gives

$$\begin{aligned} \mathbb{E}[\hat{M}_n] &= \frac{1}{n} \sum_{j=1}^n \mathbb{P}(X_j \notin \{X_i : i \neq j\}) \\ &= \frac{1}{n} \sum_{j=1}^n \mathbb{E}[M_{n-1}] = \mathbb{E}[M_{n-1}]. \end{aligned}$$

One more example can reduce missing mass. How much?

The Estimate Is High by at Most $1/n$ on Average

We know $\mathbb{E}[\hat{M}_n] = \mathbb{E}[M_{n-1}]$. How much can removing one example increase the missing mass?

The Estimate Is High by at Most $1/n$ on Average

We know $\mathbb{E}[\hat{M}_n] = \mathbb{E}[M_{n-1}]$. How much can removing one example increase the missing mass? Fix the full sample and remove a uniformly chosen position:

- ▶ Removing a repeated prompt changes nothing: it is still observed.
- ▶ Removing a singleton x makes it unseen, increasing missing mass by $\mu(x)$.

Each singleton occupies one of the n positions. The expected increase, conditional on the full sample, is therefore

$$\frac{1}{n} \sum_{x: N_x=1} \mu(x) \leq \frac{1}{n},$$

since the probabilities of any subset of prompts sum to at most one.

The Estimate Is High by at Most $1/n$ on Average

We know $\mathbb{E}[\hat{M}_n] = \mathbb{E}[M_{n-1}]$. How much can removing one example increase the missing mass? Fix the full sample and remove a uniformly chosen position:

- ▶ Removing a repeated prompt changes nothing: it is still observed.
- ▶ Removing a singleton x makes it unseen, increasing missing mass by $\mu(x)$.

Each singleton occupies one of the n positions. The expected increase, conditional on the full sample, is therefore

$$\frac{1}{n} \sum_{x: N_x=1} \mu(x) \leq \frac{1}{n},$$

since the probabilities of any subset of prompts sum to at most one. The remaining prompts have the distribution of $n-1$ independent draws. Averaging over the original sample gives

$$0 \leq \mathbb{E}[\hat{M}_n] - \mathbb{E}[M_n] = \mathbb{E}[M_{n-1}] - \mathbb{E}[M_n] \leq \frac{1}{n}.$$

Singletons Quantify the Hallucination Tradeoff

The estimate's average overstatement is at most $1/n$, so $\mathbb{E}[M_n] \geq \mathbb{E}[\hat{M}_n] - 1/n$. Substitute this into the hallucination bound:

$$\begin{aligned}\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D)] &\geq \mathbb{E}[M_n] - \frac{2}{K} \\ &\geq \mathbb{E}[\hat{M}_n] - \frac{1}{n} - \frac{2}{K}.\end{aligned}$$

Singletons Quantify the Hallucination Tradeoff

The estimate's average overstatement is at most $1/n$, so $\mathbb{E}[M_n] \geq \mathbb{E}[\hat{M}_n] - 1/n$. Substitute this into the hallucination bound:

$$\begin{aligned}\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D)] &\geq \mathbb{E}[M_n] - \frac{2}{K} \\ &\geq \mathbb{E}[\hat{M}_n] - \frac{1}{n} - \frac{2}{K}.\end{aligned}$$

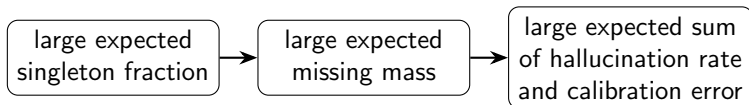
Recall that each prompt has K incorrect candidate answers. When K is large, the $2/K$ term is small.

Singletons Quantify the Hallucination Tradeoff

The estimate's average overstatement is at most $1/n$, so $\mathbb{E}[M_n] \geq \mathbb{E}[\hat{M}_n] - 1/n$. Substitute this into the hallucination bound:

$$\begin{aligned}\mathbb{E}[\text{err}_{\text{gen}}(G) + \delta_H(G, D)] &\geq \mathbb{E}[M_n] - \frac{2}{K} \\ &\geq \mathbb{E}[\hat{M}_n] - \frac{1}{n} - \frac{2}{K}.\end{aligned}$$

Recall that each prompt has K incorrect candidate answers. When K is large, the $2/K$ term is small.



Changing the Information or the Objective

The lower bound assumes incomplete information and an objective of matching a source that always answers. A deployed system can change either:

Changing the Information or the Objective

The lower bound assumes incomplete information and an objective of matching a source that always answers. A deployed system can change either:

- ▶ Retrieval or tools can supply information absent from training.
- ▶ Abstaining on unseen facts avoids guessing, but no longer matches the answer-giving source.

Summary

- ▶ Error-free examples do not supply every fact. An uninformed system can avoid false claims by abstaining.

Summary

- ▶ Error-free examples do not supply every fact. An uninformed system can avoid false claims by abstaining.
- ▶ With substantial missing information, matching even aggregate response probabilities can force hallucinations. Safe abstention departs from the answer-giving source.

Summary

- ▶ Error-free examples do not supply every fact. An uninformed system can avoid false claims by abstaining.
- ▶ With substantial missing information, matching even aggregate response probabilities can force hallucinations. Safe abstention departs from the answer-giving source.
- ▶ The chance of encountering an unseen fact controls this tradeoff. Facts observed only once help estimate that chance.

Thanks!

See you next class!

References

- ▶ Surbhi Goel. *Topic 1B: Hallucination, Calibration, and Missing Mass*. CIS 6200: Emerging Theory of LLMs, lecture notes, Fall 2026.
- ▶ Adam Tauman Kalai and Santosh S. Vempala. *Calibrated Language Models Must Hallucinate*. Proceedings of the 56th Annual ACM Symposium on Theory of Computing, 2024.
- ▶ Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. *Why Language Models Hallucinate*. arXiv:2509.04664, 2025.
- ▶ I. J. Good. *The Population Frequencies of Species and the Estimation of Population Parameters*. Biometrika, 1953.
- ▶ David McAllester and Luis Ortiz. *Concentration Inequalities for the Missing Mass and for Histogram Rule Error*. Journal of Machine Learning Research, 2003.