CIS 6200: Emerging Theory of LLMs **Fall 2026**

Topic 1B: Hallucination, Calibration, and Missing Mass

Last updated: September 8, 2026 *Surbhi Goel*

Acknowledgements: *These notes draw on Kalai and Vempala [1], Good [2], Kleinberg and Mullainathan [5], Kalavasis, Mehrotra, and Velegkas [6], Miao and Kearns [7], Kalai's Simons Institute lecture [8], and the subsequent work of Kalai, Nachum, Vempala, and Zhang [9].*

AI Disclosure: *These notes were drafted and revised with the assistance of AI tools. I have reviewed the content and take responsibility for it. If you find an error, please let me know.*

In the previous lecture, success meant producing one new valid output. A learner could find a safe part of the language and remain there. A model that tries to reproduce a source distribution faces a stronger requirement. It must learn not only which outputs are valid, but also how often the source produces them.

Imagine a collection of university documents that mention courses. Half of the mentions refer to *Introduction to Programming*. The remaining mentions are spread across many specialized courses, most of which appear only once in the sample. A generative model that reproduces this source cannot simply repeat the common course. It must spread about half of its probability across the long tail of specialized courses. Copying titles from the sample is safe, but it misses real courses that were not observed. Placing probability on plausible unseen titles better represents the source, but some of those titles may not name real courses.

Kalai and Vempala [1] turn this conflict into a statistical question. Suppose every training document is correct, the documents are independent draws from a fixed distribution, and computation is unlimited. *Can learning the source distribution itself force a model to hallucinate?* For facts whose unseen identities cannot be inferred from structure, their answer is yes. The resulting hallucination rate is controlled by a statistic that we can estimate from the training data.

1 A Statistical Model of Factual Generation

We represent each document by the single factual claim it communicates. Different wordings of the same claim correspond to the same *factoid*.

As in the previous lecture, $\mathcal{U}$ denotes the universe of possible outputs, and the target language $L^\star \subseteq \mathcal{U}$ contains the valid ones, here the true factoids. Unlike the previous lecture, $\mathcal{U}$ is finite, although possibly enormous, because we will need to count how many candidates are true and false. Let

$$\mathcal{H} = \mathcal{U} \setminus L^\star$$

be the set of plausible but false factoids. We say that a generator *hallucinates* when it produces an output in $\mathcal{H}$.

The data. Documents need not mention every true factoid equally often. A source distribution p describes these frequencies. It assigns positive probability to every true factoid and zero probability to every false one:

$$p(x) > 0 \text{ for } x \in L^\star, \qquad p(x) = 0 \text{ for } x \in \mathcal{H}.$$

Together, $L^\star$ and p specify which factoids are true and how often documents mention them. The learner does not know either one. It sees n documents sampled independently from p. If X_i is the factoid in document i, write $\mathbf{X} = (X_1, \ldots, X_n)$ for the full sample and let

$$S = \{X_1, \ldots, X_n\}$$

be the set of distinct factoids in the training data. In the previous lecture, we used S_t for the examples observed by round t. Here all n documents arrive as one sample, so we write S.

Since p places probability only on truths, every factoid in S is true, and hence $S \subseteq L^\star$. We write $\bar{S} = \mathcal{U} \setminus S$ for the factoids that did not appear. This set contains both true factoids that the sample missed and false factoids in $\mathcal{H}$. Figure 1 shows the distinction that the learner cannot see.

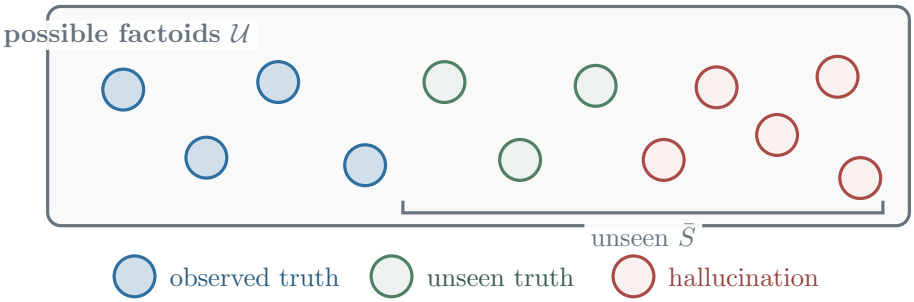

Figure 1: The learner sees the blue factoids. It does not know which of the remaining candidates are unseen truths and which are hallucinations.

The generator. After seeing the n documents, the learner constructs a distribution q over $\mathcal{U}$, where $q(x)$ is the probability that it generates factoid x.

2 A Simplified Hallucination Bound

For any set $A \subseteq \mathcal{U}$, write the total probabilities that the source and the generator assign to A as

$$p(A) = \sum_{x \in A} p(x), \qquad q(A) = \sum_{x \in A} q(x).$$

In particular, $q(\mathcal{H})$ is the probability that the generator hallucinates. As Figure 1 shows, every unseen factoid $\in \bar{S}$ is either a truth $\in L^\star \setminus S$ that the sample missed or a false factoid $\in \mathcal{H}$. Therefore

$$\underbrace{q(\mathcal{H})}_{\substack{\text{probability of a} \\ \text{hallucination}}} = \underbrace{q(\bar{S})}_{\substack{\text{probability of an} \\ \text{unseen output}}} - \underbrace{q(L^\star \setminus S)}_{\substack{\text{probability of a} \\ \text{new valid output}}}. \tag{1}$$

A generator can make $q(\bar{S})$ zero by producing only factoids in the sample. It would then avoid hallucination, but it would also ignore every true factoid that the sample missed. If the generator is to represent those missing truths, it must place some probability on $\bar{S}$. Avoiding hallucination then requires that this probability find $L^\star \setminus S$, rather than the false factoids in $\mathcal{H}$.

Define the missing mass of $\mathbf{X} = (X_1, \ldots, X_n)$ by $M_p(\mathbf{X}) := p(\mathcal{U} \setminus \{X_1, \ldots, X_n\})$. Since the source places no probability on false factoids,

$$\underbrace{M_p(\mathbf{X})}_{\text{missing mass}} = p(\bar{S}) = p(L^\star \setminus S).$$

We first prove the bound under two simplifying assumptions.

1. **Breadth.** The generator produces an unseen factoid with the same probability as a fresh source document does:
$$q(\bar{S}) = p(\bar{S}) = M_p(\mathbf{X}).$$
This fixes only the total probability outside the sample. It does not require $q = p$, nor does it tell the generator which unseen factoids are true.

2. **Symmetric uncertainty.** Before the documents are drawn, we choose a fixed number of true factoids uniformly from $\mathcal{U}$, then let the source p be uniform over them. The names of unseen factoids therefore carry no clues about which ones are true.

Now condition on the observed sample. Every possible truth set of the chosen size that contains S makes this sample equally likely. The unseen truths are therefore uniformly distributed among the candidates in $\bar{S}$. Exactly $|L^\star \setminus S|$ of the $|\bar{S}|$ unseen candidates are true, so this common probability must be

$$\mathbb{P}(x \in L^\star \mid \mathbf{X}) = \frac{|L^\star \setminus S|}{|\bar{S}|} \qquad \text{for every } x \in \bar{S}.$$

Now fix the sample and the resulting generator q. The only remaining randomness is which source p was chosen. A candidate x contributes $q(x)$ to the probability of a new valid output if it is true, and zero otherwise. Adding these expected contributions gives

$$\mathbb{E}\big[q(L^\star \setminus S) \mid \mathbf{X}\big] = \sum_{x \in \bar{S}} q(x)\mathbb{P}(x \in L^\star \mid \mathbf{X}) = \frac{|L^\star \setminus S|}{|\bar{S}|} q(\bar{S}).$$

Redistributing probability among the unseen candidates does not change this expectation because all of them have the same posterior probability of being true. Substituting this and $q(\bar{S}) = M_p(\mathbf{X})$ into Equation (1) proves the following.

Proposition 2.1 (Hallucination under breadth and symmetry)**.** *Under breadth and symmetric uncertainty, for every sample* $\mathbf{X}$*,*

$$\mathbb{E}[q(\mathcal{H}) \mid \mathbf{X}] = \left(1 - \frac{|L^\star \setminus S|}{|\bar{S}|}\right) M_p(\mathbf{X}) \;\geq\; \left(1 - \frac{|L^\star|}{|\mathcal{H}|}\right) M_p(\mathbf{X}).$$

The fraction $|L^\star \setminus S|/|\bar{S}|$ is the proportion of unseen candidates that are true. The inequality follows because every hallucination is unseen, so $|\bar{S}| \geq |\mathcal{H}|$, and $|L^\star \setminus S| \leq |L^\star|$. The bound is therefore most useful when true factoids are rare among the plausible ones. Kalai and Vempala call this *sparsity* and write it as $|L^\star| \leq \rho|\mathcal{H}|$ for a small constant ρ [1]. In the course example, there are far more plausible course titles than real courses. Under sparsity, the expected hallucination probability is at least $(1 - \rho)$ times the missing mass.

Let's return to the course example. Let the set of plausible course titles be $\mathcal{U} = \{A, B, C, D\}$. Choose the two real courses, $L^\star$, uniformly from the six pairs in $\mathcal{U}$, and let the source p assign probability $1/2$ to each. Thus $p(x)$ is the probability that a fresh document mentions course x.

Suppose the sample contains only A. Then $S = \{A\}$ and $\bar{S} = \{B, C, D\}$, and each of B, C, D is equally likely to be the other real course. Hence $|L^\star \setminus S|/|\bar{S}| = 1/3$.

After seeing the sample, the learner produces q, where $q(x)$ is the probability that the model outputs course x. The breadth condition requires

$$q(B) + q(C) + q(D) = q(\bar{S}) = p(\bar{S}) = \frac{1}{2}.$$

Thus the expected probability of a new valid course is $(1/3)(1/2) = 1/6$, and the expected hallucination probability is $1/2 - 1/6 = 1/3$.

2.1 How the Full Result Relaxes the Assumptions

Proposition 2.1 assumed breadth and symmetric uncertainty. Kalai and Vempala prove a version of the bound under weaker assumptions [1]. We describe what replaces each one.

Semantic calibration. Breadth required the generator to match the source on one particular set, the unseen factoids $\bar{S}$. Kalai and Vempala instead require agreement on sets defined by the generator itself. For each probability value z, let

$$B_z = \{x \in \mathcal{U} : q(x) = z\}$$

be the set of factoids to which the generator assigns probability z. The generator is *semantically calibrated* if

$$p(B_z) = q(B_z) \qquad \text{for every nonempty } B_z.$$

Since $q(B_z) = z|B_z|$, the condition says that among the factoids the generator assigns probability z, the source's average probability is also z. This is the usual meaning of calibration. We do not require $p(x) = q(x)$ for any individual factoid. Figure 2 shows a calibrated generator in which no individual factoid has $p(x) = q(x)$.

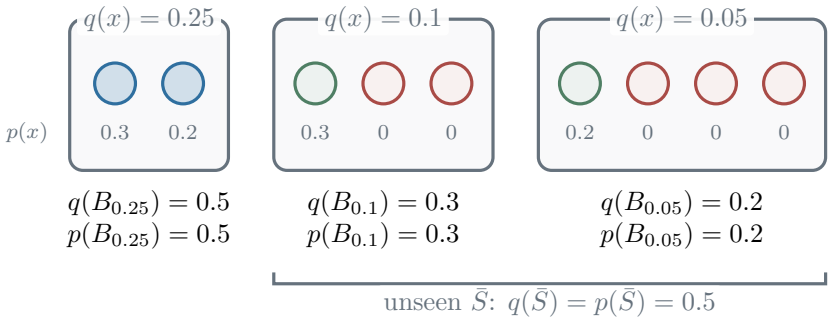

Figure 2: A semantically calibrated generator. Factoids are grouped by the probability $q(x)$ the generator assigns them. Within each group the source probabilities $p(x)$ differ from $q(x)$, and hallucinations have $p(x) = 0$, but each group's total probability agrees under p and q. The unseen factoids form a union of groups, so $q(\bar{S}) = p(\bar{S})$. Colors are as in Figure 1.

The groups B_z partition $\mathcal{U}$. Summing the calibration equalities over any collection of groups shows that p and q agree on every set that is a union of groups, that is, every set defined by a condition on $q(x)$ alone. Calibration gives breadth in the simple case where no group contains both an observed and an unseen factoid, because $\bar{S}$ is then a union of groups and $q(\bar{S}) = p(\bar{S})$. The

full theorem allows approximate calibration and groups that mix observed and unseen factoids. The total disagreement $\sum_z |p(B_z) - q(B_z)|$ appears as a subtracted term in the bound. Thus a generator can hallucinate less by placing more probability on observed truths, but it then becomes less calibrated.

Regularity. We used symmetry once, to conclude that every unseen candidate has the same conditional probability of being true. This made the expected probability on unseen truths equal to $q(\bar{S})$ times the fraction of unseen candidates that are true, no matter how the generator distributes $q(\bar{S})$. *Regularity* replaces the equality by an inequality. After the sample is observed, no unseen candidate may be more than r times as likely to be true as the average unseen candidate. The expected probability on unseen truths is then at most r times the value under symmetry, and the bound loses the same factor. Symmetry is the case $r = 1$.

Regularity restricts which factoids are true, not how often the source mentions them. In particular, it does not require p to be uniform. For example, fix any long-tailed list of probabilities and assign it to the true factoids in a random order. The true factoids then have very different probabilities, but the name of an unseen factoid still says nothing about whether it is true.

3 Estimating Missing Mass with Good–Turing

The hallucination bound above is proportional to the missing mass $M_p(\mathbf{X})$. When $M_p(\mathbf{X})$ is large, breadth requires the generator to place substantial probability on unseen factoids; when it is small, this pressure is weak. We therefore need to estimate $M_p(\mathbf{X})$ from the sample. At first this seems impossible: the quantity depends precisely on factoids that the sample does not contain.

Good–Turing estimation uses factoids that appear exactly once as evidence of unseen source probability. If course mentions repeatedly name the same few courses, the missing mass is likely small; if many course titles appear only once, it is likely larger. Good introduced this idea for species and vocabulary, while crediting Turing for much of its development [2].

Definition 3.1 (Monofact rate)**.** Let $N_1(\mathbf{X})$ be the number of factoids that appear exactly once in the sample. The *monofact rate* is the fraction of document positions occupied by these singleton factoids:

$$\widehat{\mathrm{MF}}(\mathbf{X}) = \frac{N_1(\mathbf{X})}{n}.$$

We first show that the monofact rate and the missing mass have nearly the same expectation. Throughout, p is fixed and the expectation is over the sample $\mathbf{X} \sim p^{\otimes n}$.

Proposition 3.2 (Expected monofact rate)**.**

$$0 \;\le\; \mathbb{E}[\widehat{\mathrm{MF}}(\mathbf{X})] - \mathbb{E}[M_p(\mathbf{X})] \;\le\; \frac{1}{n}.$$

Proof. Fix a factoid x. It is missing from the sample when all n documents miss it. This happens with probability $(1-p(x))^n$, and when it happens, x contributes $p(x)$ to the missing mass. Summing over x,

$$\mathbb{E}[M_p(\mathbf{X})] = \sum_{x \in \mathcal{U}} p(x)\,(1-p(x))^n.$$

The same factoid appears exactly once when one of the n documents contains it and the other $n-1$ miss it. This happens with probability $n\,p(x)(1-p(x))^{n-1}$, and when it happens, x contributes $1/n$

to the monofact rate. Summing over x,

$$\mathbb{E}[\widehat{\mathrm{MF}}(\mathbf{X})] = \sum_{x \in \mathcal{U}} \frac{1}{n} \cdot n\, p(x)(1-p(x))^{n-1} = \sum_{x \in \mathcal{U}} p(x)\,(1-p(x))^{n-1}.$$

The two sums differ only in the exponent. Since $(1-p(x))^{n-1} \geq (1-p(x))^n$, the difference is nonnegative, and

$$\mathbb{E}[\widehat{\mathrm{MF}}(\mathbf{X})] - \mathbb{E}[M_p(\mathbf{X})] = \sum_{x \in \mathcal{U}} p(x)^2 (1-p(x))^{n-1} \leq \frac{1}{n} \sum_{x \in \mathcal{U}} p(x) = \frac{1}{n},$$

where we used $t(1-t)^{n-1} \leq 1/n$ for $t \in [0,1]$. □

The proof explains why singletons carry information about unseen factoids. A factoid appears exactly once when one document produces it and the other $n-1$ documents miss it. The expected monofact rate is therefore the expected missing mass after $n-1$ draws, and one additional draw changes the missing mass by at most $1/n$ in expectation. Concentration bounds show that the realized monofact rate and the realized missing mass are also close with high probability. We need only the direction that lower-bounds the missing mass.

Theorem 3.3 (Good–Turing missing-mass estimate [3, 4])**.** *For every distribution p, every $n \geq 1$, and every $\delta \in (0, 1/3]$, if $\mathbf{X} \sim p^{\otimes n}$, then with probability at least $1-\delta$ over $\mathbf{X}$,*

$$M_p(\mathbf{X}) \geq \widehat{\mathrm{MF}}(\mathbf{X}) - \sqrt{\frac{6\ln(2/\delta)}{n}}.$$

The guarantee is distribution-free. It assumes neither a known number of true factoids nor any particular shape for their probabilities. If 1,000 factoid occurrences contain 180 singletons, then $\widehat{\mathrm{MF}}(\mathbf{X}) = 0.18$. We estimate that the next source draw has roughly an 18% chance of being a factoid absent from the sample. It does not say that 18% of all true factoids remain undiscovered.

Returning to the simplified bound, the theorem lets us replace the unknown missing mass by a quantity computed from the sample. With probability at least $1-\delta$,

$$\mathbb{E}[q(\mathcal{H}) \mid \mathbf{X}] \geq \left(1 - \frac{|L^\star \setminus S|}{|\bar{S}|}\right)\left(\widehat{\mathrm{MF}}(\mathbf{X}) - \sqrt{\frac{6\ln(2/\delta)}{n}}\right).$$

The first factor is the fraction of unseen candidates that are false. The second is a lower estimate of how much source probability lies outside the sample.

4 Breadth, Validity, and the Tradeoff

In the previous lecture, the learner of Kleinberg and Mullainathan produced valid outputs by staying inside a safe subset of the language [5]. It was never required to cover the rest. Kalavasis, Mehrotra, and Velegkas call the additional requirement that a generator cover the full range of valid outputs *breadth* [6]. In our setting, breadth means that the generator represents the factoids the source can produce, including those that did not appear in the sample. Our simplified proof used the equality $q(\bar{S}) = p(\bar{S})$, and semantic calibration is one way to obtain it.

The lower bound has two hypotheses: the generator has breadth, and it cannot tell which unseen factoids are true. We examine each in turn. We first show that the second hypothesis can be stated without symmetry. Any generator that cannot distinguish the unseen truths in $L^\star \setminus S$ from the

hallucinations in $\mathcal{H}$ must hallucinate, whatever form its uncertainty takes. We then look at what happens when a generator gives up breadth and places its probability on the truths it has already seen. Finally, we consider a response that the model has had no way to give so far: declining to answer.

4.1 Hallucination When Validity Is Hard to Recognize

Kalai, Nachum, Vempala, and Zhang relate the hallucination rate of any generator to a classification problem [9]. We call the problem *Is-It-Valid*. With probability $1/2$, we draw x from the source p and label it valid. With probability $1/2$, we draw x uniformly from $\mathcal{H}$ and label it invalid. Valid examples appear with their source frequencies. Each invalid example has probability $1/|\mathcal{H}|$.

A generator q defines a natural classifier for this problem. It accepts x when q assigns x more probability than a uniformly random hallucination receives:

$$A_q = \left\{ x \in \mathcal{U} : q(x) > \frac{1}{|\mathcal{H}|} \right\}, \qquad \bar{A}_q = \mathcal{U} \setminus A_q.$$

The classifier makes an error when it rejects a valid example or accepts an invalid one, so its error is

$$\mathrm{IIVErr}(q) = \underbrace{\frac{1}{2}\, p(\bar{A}_q)}_{\text{truth rejected}} + \underbrace{\frac{1}{2}\, \frac{|A_q \cap \mathcal{H}|}{|\mathcal{H}|}}_{\text{hallucination accepted}}. \tag{2}$$

Proposition 4.1 (Hallucination and validity testing)**.** *If q is semantically calibrated, then*

$$q(\mathcal{H}) \geq 2\,\mathrm{IIVErr}(q) - \frac{|L^\star|}{|\mathcal{H}|}. \tag{3}$$

Proof. Every accepted hallucination has $q(x) > 1/|\mathcal{H}|$, so

$$\frac{|A_q \cap \mathcal{H}|}{|\mathcal{H}|} \leq q(A_q \cap \mathcal{H}).$$

The rejected set $\bar{A}_q = \{x : q(x) \leq 1/|\mathcal{H}|\}$ is a union of groups B_z, so calibration gives $p(\bar{A}_q) = q(\bar{A}_q)$. There are at most $|L^\star|$ rejected truths, each with $q(x) \leq 1/|\mathcal{H}|$, so

$$p(\bar{A}_q) = q(\bar{A}_q \cap L^\star) + q(\bar{A}_q \cap \mathcal{H}) \leq \frac{|L^\star|}{|\mathcal{H}|} + q(\bar{A}_q \cap \mathcal{H}).$$

Adding the two bounds gives $2\,\mathrm{IIVErr}(q) \leq q(\mathcal{H}) + |L^\star|/|\mathcal{H}|$. □

Accepting a hallucination costs the generator probability directly. Rejecting a truth is cheap for a calibrated generator only because truths are sparse. Thus, if validity is hard to recognize, so that no classifier performs well on Is-It-Valid, and truths are sparse, then a calibrated generator must hallucinate. The direction matters. The proposition does not say that a generator able to recognize validity avoids hallucination. It says that one unable to recognize validity cannot avoid it.

This also explains what symmetry was doing in Proposition 2.1. Under symmetric uncertainty, every unseen candidate is equally likely to be true given the sample, so the name of a candidate carries no information about its label. Any classifier that accepts some unseen candidates and rejects others accepts, in expectation, the same fraction of unseen truths as of hallucinations. It does no better than chance on the unseen candidates. Symmetry is thus the extreme case in which

validity is impossible to recognize among unseen candidates. The proposition replaces this extreme case by a quantitative one: the harder validity is to recognize, the more a calibrated generator must hallucinate.

Without calibration, the equality $p(\bar{A}_q) = q(\bar{A}_q)$ in the proof becomes an inequality with error $|p(\bar{A}_q) - q(\bar{A}_q)|$, and this term is subtracted from the bound. A generator can therefore avoid hallucination only by disagreeing with the source on the rejected set. The breadth equality $q(\bar{S}) = p(\bar{S})$ alone does not control this disagreement.

4.2 Trading Calibration for Fewer Hallucinations

Miao and Kearns study what happens when a generator gives up breadth [7]. They hold the training data fixed, so the monofact rate does not change, and move the generator toward the observed truths. In the experiment shown in Figure 3, they first fine-tune a model normally. Late in training, they begin repeating 5% of the examples ten times. The intervention reveals no new truths. It can only shift probability toward facts already in the sample.

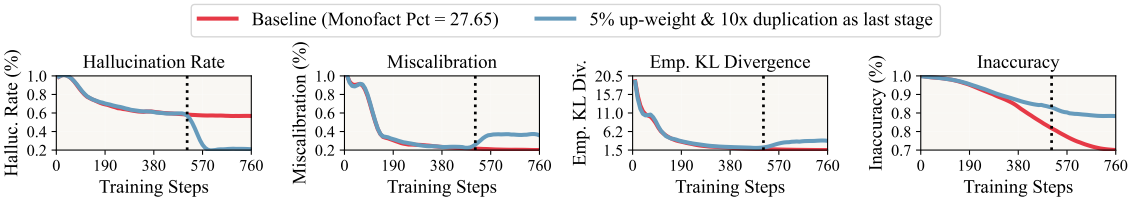

Figure 3: Late-stage upweighting in a fine-tuned T5-Small model. The dashed line marks the start of the intervention. Hallucination falls sharply, while miscalibration and empirical KL divergence rise. Prompted inaccuracy remains near its level before the intervention. Reproduced from Miao and Kearns [7], Figure 6.

After the intervention, hallucination falls, and miscalibration and the divergence from the source rise. The intervention adds no information about unseen truths, so the model hallucinates less because it has given up breadth and concentrated on familiar truths. Miao and Kearns measure miscalibration across probability bins rather than on the single set $\bar{A}_q$, but the behavior is the tradeoff described above.

4.3 Abstention and How Models Are Evaluated

Every generator in these notes had to output a factoid. A deployed model has a third option: it can say that it does not know. Abstaining produces no hallucination. It also produces no unseen truth, so it gives up breadth, but unlike the upweighting of Section 4.2, it makes the loss of coverage explicit. The model reports that it cannot tell.

Kalai, Nachum, Vempala, and Zhang argue that models rarely take this option because evaluations penalize it [9]. Their analogy is a student on a multiple-choice exam. If wrong answers cost nothing, an unsure student should always guess. If wrong answers are penalized, the student should guess only when confident enough. Most benchmarks grade like the first exam.

To make this precise, suppose the model has a candidate answer that is correct with probability c. A scoring rule gives a correct answer 1 point, an abstention 0 points, and a wrong answer $-\lambda$ points for some penalty $\lambda \geq 0$.

	correct	wrong	abstain
score	1	$-\lambda$	0

Answering gains 1 with probability c and loses λ with probability $1-c$, so its expected score is $c-(1-c)\lambda$. Abstaining scores 0. Answering is therefore the better choice exactly when

$$c > (1-c)\lambda, \qquad \text{that is,} \qquad c > \frac{\lambda}{1+\lambda}. \tag{4}$$

Most benchmarks use $\lambda = 0$: a wrong answer and an abstention both score zero. Then (4) says to answer whenever $c > 0$, however small. Such an answer is wrong with probability $1-c$, and when it is wrong, it is a hallucination in the sense of these notes. This is their explanation for why hallucination persists after post-training: the training signal rewards guessing.

The proposal is to use a positive penalty and to state it in the evaluation instructions. With $\lambda = 1$, a wrong answer costs one point, and by (4) the model should answer when $c > 1/2$, that is, when it is more likely right than wrong. With $\lambda = 9$, the model should answer only when $c > 0.9$. The instructions can equally state the confidence level $\lambda/(1+\lambda)$ at which the model should answer, since the two determine each other. Under such a rule, an optimal model answers exactly when its confidence exceeds the stated threshold. The authors call the ability to adjust this decision to the stated threshold *behavioral calibration*.

To decide whether to answer, the model must estimate c. This requires judging whether its candidate is valid, the same underlying difficulty captured by Is-It-Valid. Return to the course example of Section 2, where the sample contains only course A and each of B, C, D is equally likely to be the other real course. Asked to name a course other than A, the model can propose B, which is correct with probability $c = 1/3$, or abstain. With no penalty, it should propose B, and with probability $2/3$ the response is a hallucination. With penalty $\lambda = 1$, the break-even confidence is $1/2 > 1/3$, so it should abstain. The uncertainty is the same in both cases. The scoring rule decides whether it becomes a guess or an admission.

Abstention does not contradict the hallucination lower bound. The earlier generator had to place all its probability on factoids in $\mathcal{U}$. Adding "I do not know" changes the action space and the objective. It converts some probability of hallucination into probability of providing no answer; it does not discover the missing truths. Abstention therefore makes the tradeoff between factuality and coverage explicit.

References

[1] A. T. Kalai and S. S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 160–171, 2024.

[2] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264, 1953.

[3] D. A. McAllester and L. E. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911, 2003.

[4] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013.

[5] J. Kleinberg and S. Mullainathan. Language generation in the limit. In *Advances in Neural Information Processing Systems*, volume 37, pages 66058–66079, 2024.

[6] A. Kalavasis, A. Mehrotra, and G. Velegkas. On the limits of language generation: Trade-offs between hallucination and mode-collapse. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1732–1743, 2025.

[7] M. M. Miao and M. Kearns. Hallucination, monofacts, and miscalibration: An empirical investigation. *Proceedings of the National Academy of Sciences*, 123(8):e2533582123, 2026.

[8] A. T. Kalai. When calibration goes awry: hallucination in language models. Simons Institute talk, with slides titled *Demystifying hallucinations and other generalization issues*, September 2024.

[9] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653:1047–1051, 2026.