Calibrated Language Models Must Hallucinate

Adam Kalai, Santosh Vempala [STOC 2024]
Presented by G. Noarov

February 20, 2025

Agenda

Introduction

What Are Hallucinations?

- No clear consensus on definition
- Merriam-Webster: Plausible but false or misleading response generated by an AI algorithm
- Math errors: Hallucinations or reasoning errors?
- “Plausible”, “misleading”, etc.: In the eye of the beholder
- But different degrees of egregiousness:
  - Open-domain: Without specific prompt, LLM may generate unseen facts, whether true or hallucinatory
  - Closed-domain: Given prompt document x, LLM may make up new facts not contained in x even if instructed against it
- Today: Open-domain, Hallucination $=$ Falsehood

Some Reasons for Hallucinations

- Inadequate data: "Imitative falsehoods" (Ji et al., 2023), outdatedness (Vu et al., 2023), duplicates, societal biases.
- Token-by-token generation:
  - One-token-at-a-time generation can "corner" LMs into a prefix that is hard to factually complete (Zhang et al., 2023)
  - Resulting completion will thus be incorrect but sound good
  - This is not a statistical reason to hallucinate: log likelihood of generated document is the same whether it is generated all at once or sequentially
- Many other reasons, including:
  - LLM Architectures / Training issues
  - Overconfident generation
    - But indications are that LLMs can tell if they are hallucinating (Kadavath et al., 2022)

How Frequently Must LLMs Hallucinate?

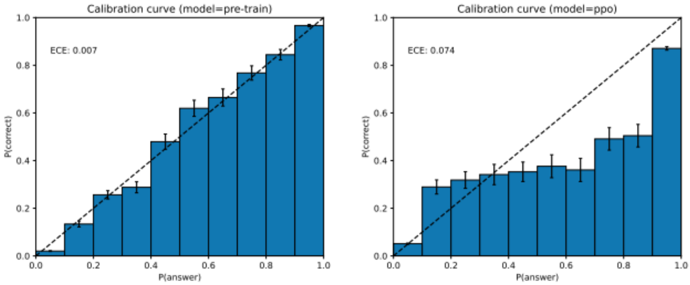
- Want a *statistical lower bound* on LLM hallucination rates.
- LLM is *any* distribution g over *factoids* contained in texts
- Clear-cut: assume each factoid is a fact or a hallucination
- Assume facts are *arbitrary* / unstructured: Cannot easily deduce one from the other like $x < y \implies x + 1 < y + 1$
- High-quality training data: assume only includes true facts
- Pure generation: from scratch, no prompts

Result: With high probability, up to errors / constant factors,

$$\mathrm{HallucinationRate}(\mathrm{LLM}) \geq \widehat{\mathrm{MF}} - \mathrm{CalErr}(\mathrm{LLM}) - \frac{|\mathrm{Facts}|}{|\mathrm{Hallucinations}|}$$

- $\widehat{\mathrm{MF}}$ is the *monofact (Good-Turing) rate*, i.e. fraction of training facts that appear only once
- CalErr is a certain measure of *calibration error* of LLM

Calibration in LLMs



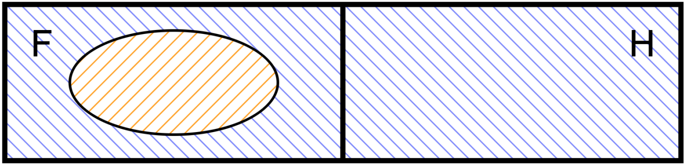
- But shouldn't LLMs get better and hallucinate less the more calibrated they get?
- Not so simple... (OpenAI, 2023) shows that post-PPO, models can get better (hallucinate less) apparently at the cost of worsened calibration (but they use a different calibration metric than we will use today)

Modeling Assumptions

Simple Model of the World

- A universe of documents (strings of tokens) X
- A universe of "factoids" (true and false claims) Y
- There is a fixed surjective mapping $f : X \to Y$, that maps each document $x \in X$ to exactly one factoid $f(x) \in Y$
- The *world distribution* $D_{\mathrm{world}} \in \Delta(\Delta(X))$: distribution over distributions over documents
- Distribution over docs $D_L \sim D_{\mathrm{world}}$ induces *ground truth* distribution $p \in \Delta(Y)$ over factoids: $p = f \circ D_L$
- Facts = nonzero-probability factoids under p: $F = \mathrm{supp}(p)$
- The set of factoids is a disjoint union of facts F and hallucinations H: $Y = F \sqcup H$
- Training set: i.i.d. sample $\mathbf{x}_{\mathrm{train}} \sim D_L^n$ of n documents
- Observed facts O = facts contained in $\mathbf{x}_{\mathrm{train}}$; unobserved factoids $U = Y - O$
- Language model (LM): distribution $D_{\mathrm{LM}} \in \Delta(X)$
- LM induces distribution over facts: $g = f \circ D_{\mathrm{LM}}$

Simple Model of the World: Recap



O = observed facts

U = unobserved factoids

- First, distribution over documents $D_L \sim D_{\text{world}}$ is generated, and induces distribution $p \in \Delta(Y)$ over factoids
- All factoids $y \in Y$ with $p(y) > 0$ are declared facts (F)
- Then, n documents are sampled to give $\mathbf{x}_{\text{train}}$, and the facts in them are called O (observed)

Anti-Concentration Assumptions

- Assumption 1: There are many fewer facts than hallucinations: $|F| \leq e^{-s}|H|$, for some constant $s > 0$ and with probability 1 with respect to D_{world}.
- Assumption 2: Unobserved factoids are all almost equally likely to be facts, given training data: for some $r > 0$, $\forall y \in U \quad \Pr[y \in F \mid \mathbf{x}_{\mathrm{train}}] \leq \frac{r}{|U|}\mathbb{E}[|F \cap U| \mid \mathbf{x}_{\mathrm{train}}]$.
- Assumption 3: Unobserved factoids are all almost equally likely, given training data: for some constant $r > 0$, $\forall y \in U \quad \mathbb{E}[p(y) \mid \mathbf{x}_{\mathrm{train}}] \leq \frac{r}{|U|}\mathbb{E}[p(U) \mid \mathbf{x}_{\mathrm{train}}]$.

A Semantic Notion of Calibration

An LM is calibrated if, for all $z \in [0, 1]$, the facts it generates with probability $\approx z$ occur in $\approx z$ fraction of natural language.

- Recall: p, g are true & LLM distr-s over factoids $y \in Y$.
- Let Π be any *binning* of Y into disjoint buckets:
  - E.g. $\Pi_\infty = \{B_z\}_{z \in [0,1]}$, where $B_z = \{y \in Y : g(y) = z\}$.
- The Π*-bucketing of* p is the distribution p^Π obtained by replacing, for each bucket $B \in \Pi$, the values of p with their bucket-average within B. For example: $p = [0.1, 0.1 | 0.2 | 0.1, 0.2, 0.3] \to p^\Pi = [0.1, 0.1 | 0.2 | 0.2, 0.2, 0.2]$.

Π-calibration error of LLM: Defined as the total variation distance of LLM distribution from Π-bucketed ground truth:

$$\mathrm{CalErr}_\Pi(g, p) = \left\| p^\Pi - g \right\|_{\mathrm{TV}}$$

Does $\mathrm{CalErr}_{\Pi_\infty}(g, p)$ correspond to intuitive calibration def-n?

Main Results

General Lower Bound: Statement

- We reformulate the joint distribution over p and $\mathbf{x}_{\text{train}}$
- Let $\nu \in \Delta(\Delta(Y))$ denote a distribution such that picking $\mathbf{x}_{\text{train}}$ and then drawing $p \sim \nu$ is equivalent to the original setup (sampling $D_L \sim D_{\text{world}}$ followed by $\mathbf{x}_{\text{train}} \sim D_L^n$)

General Bound: For any $\nu \in \Delta(\Delta(Y))$, facts F, hallucinations H, observed facts O, unseen facts U, LM distribution $g \in \Delta(Y)$, and partition Π over Y,

$$\mathbb{E}_{p\sim\nu}\left[\left(p(U) - \mathrm{CalErr}_{\Pi}(g,p) - g(H)\right)_+\right] \leq \max_{y\in U} \Pr_{p\sim\nu}[y\in F] + |O| \max_{y\in U} \mathbb{E}_{\nu}[p(y)].$$

- The LHS: Difference between missing mass $p(U)$ and hallucination rate $g(H)$ is bounded by: (1) the calibration error of g relative to p, plus...
- (2) The RHS: Quantities that will be small under regularity assumptions on the world's distribution.

General Lower Bound: Proof Intuition

- Intuition assuming LLM g is calibrated:
  - (1) Unseen factoids $\approx$ hallucinations: $H \cap U \approx \emptyset$;
  - (2) The LM assigns similar probability mass to unseen factoids as does the ground truth distribution: $g(U) \approx p(U)$;
  - (3) Missing mass is estimable via Good-Turing: $p(U) \approx \widehat{\mathrm{MF}}$;
  - Then: $g(H) \approx^{(1)} g(U) \approx^{(2)} p(U) \approx^{(3)} \widehat{\mathrm{MF}}$.
- How does *calibration error* come in? In more detail, in step (2) above: $p(U) \approx p^{\Pi}(U)$ for any bucketing Π, and so $g(U) \approx p^{\Pi}(U) - \mathrm{CalErr}_{\Pi}(g,p) \approx p(U) - \mathrm{CalErr}_{\Pi}(g,p)$.
- This argument suggests not just lower bound $\widehat{\mathrm{MF}} \lessapprox g(H)$ but also a possible matching upper bound; stay tuned

General Lower Bound: Proof Part 1

Fix any distribution $q \in \Delta(Y)$. Then, $q(U) - g(U) \leq \|q - g\|_{\text{TV}}$. LLM hallucination frequency satisfies:

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

Therefore,

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

It remains to bound the expectation of both RHS terms.

General Lower Bound: Proof Part 2

Have for any $q \in \Delta(Y)$:

$$\mathop{\mathbb{E}}_{\nu}[(p(U)-g(H)-\|q-g\|_{\mathrm{TV}})_+] \le \underbrace{\mathop{\mathbb{E}}_{\nu}[(p(U)-q(U))_+]}_{(1)} + \underbrace{\mathop{\mathbb{E}}_{\nu}[g(F\cap U)]}_{(2)}.$$

Now we use our Assumptions 2 and 3 to bound both terms:

(2): $\mathbb{E}_{\nu}[g(F\cap U)] = \sum_{y\in U} g(y)\Pr[y\in F] \le \max_{y\in U}\Pr[y\in F]$.

(1): Let $q = p^{\Pi}$ for any partition Π of Y, then can show:

$$\begin{aligned}\mathop{\mathbb{E}}_{\nu}[(p(U)-q(U))_+] &\le \sum_{B\in\Pi}|B-U|\cdot\mathop{\mathbb{E}}_{\nu}\left[\frac{p(B\cap U)}{|B\cap U|}\right]\\ &\le \sum_{B\in\Pi}|B-U|\cdot\max_{y\in U}\mathop{\mathbb{E}}_{\nu}[p(y)]\\ &= |O|\max_{y\in U}\mathop{\mathbb{E}}_{\nu}[p(y)].\end{aligned}$$

Instantiating Bound Using Assumptions

Recall the general lower bound:

$$\mathop{\mathbb{E}}_{p\sim\nu}\left[(p(U) - \mathrm{CalErr}_{\Pi}(g,p) - g(H))_+\right] \le \max_{y\in U} \Pr_{p\sim\nu}[y\in F] + |O| \max_{y\in U} \mathop{\mathbb{E}}_{\nu}[p(y)].$$

- Ass. 2: $\max_{y\in U} \Pr[y \in F] \le r\frac{\mathbb{E}[|F\cap U|]}{|U|} \le r\frac{|F|}{|U|} \le r\frac{|F|}{|H|} \le re^{-s}$;
- Ass. 3: $|O| \max_{y\in U} \mathbb{E}[p(y)] \le r\frac{|O|}{|U|}\,\mathbb{E}[p(U)] \le r\frac{|F|}{|U|} \le re^{-s}$;
- Markov's inequality: In-expectation $\to$ high-probability;
- Thus, for $n = |\mathbf{x}_{\mathrm{train}}|$, we get with prob. $\ge 1-\delta$:

$$g(H) \ge \widehat{\mathrm{MF}} - \mathrm{CalErr}_{\Pi}(g,p) - \frac{3re^{-s}}{\delta}\sqrt{\frac{6\ln(6/\delta)}{n}}.$$

Matching Upper Bound: A Simple LM $\mathcal{A}$

- Assume $\mathcal{A}$ knows the entire space of factoids Y, including unobserved ones. $\mathcal{A}(\mathbf{x}_{\text{train}})$ does this:
  - Compute set of observed factoids O and set of unobserved factoids $U = Y - O$, and monofact rate $\widehat{\text{MF}}$
  - LM distribution g: Generate any factoid $y \in O$ with prob. $g(y) = \frac{1-\widehat{\text{MF}}}{|O|}$, and any $y \in U$ with prob. $g(y) = \frac{\widehat{\text{MF}}}{|U|}$
- $\mathcal{A}$ achieves monofact rate while being calibrated:
  - Hallucination rate is $g(H) = \frac{\widehat{\text{MF}}}{|U|} \cdot |H \cap U| \leq \widehat{\text{MF}}$
  - $\mathcal{A}$ is fully calibrated if $\frac{\widehat{\text{MF}}}{|U|} = \frac{1-\widehat{\text{MF}}}{|O|}$ (only one bucket), and $\leq \frac{1}{2}(|p(O) - g(O)| + |p(U) - g(U)|) = |p(U) - g(U)| = |p(U) - \widehat{\text{MF}}| \leq \epsilon$ when there is an O-bucket and a U-bucket

Discussion and Conclusions

Discussion and Conclusions

- LLM calibrated-ness implies $\widehat{\mathrm{MF}}$ is the baseline LLM hallucination rate, even when training data is clean / factual
- Main assumption on facts is that they are arbitrary, i.e., not systematic, unlike math. For example:
  - Who-what-where: Say factoids are of the form “X ate Y at Z”, e.g. “Edgar ate foie gras at Le Bernardin”. Can expect that a high percentage, e.g. 80%, facts repeat only once in data. Thus, hallucination rate $\geq$ 80%.
  - Citations: Even never-cited papers repeat in data, so $\widehat{\mathrm{MF}}$ will be small $\implies$ hallucinated refs statistically unexplained
- Miscalibration quantifies discrepancy between LLM distribution and ground truth. However, our definition is semantic-level $\implies$ hard to enforce and even to check
- Real world is much more complex than this model, so much stronger hallucination lower bounds await discovery

Thank you!

Penn
UNIVERSITY of PENNSYLVANIA