Discussion of

Why Language Models Hallucinate

(Kalai, Nachum, Vempala, Zhang ’25)

Andy Haupt @ REFORM Stanford, 10/23/2025

Hallucinations

Adam Kalai's Ph.D. Thesis

Number of r's in strawberry

ChatGPT: (GPT-4o)	Adam Tauman Kalai's Ph.D. dissertation (completed in 2002 at CMU) is entitled: "Boosting, Online Algorithms, and Other Topics in Machine Learning."
DeepSeek:	"Algebraic Methods in Interactive Machine Learning"... at Harvard University in 2005.
Llama:	"Efficient Algorithms for Learning and Playing Games"... in 2007 at MIT.

Main Idea of this paper: Generative models "also" solve valid fact classification.
Hence limits of classification mean limits of generation.

- $\mathcal{X} = \mathcal{V} \cup \mathcal{E}$ strings
    - $\mathcal{V}$ valid outputs
    - $\mathcal{E}$ errors/hallucinations
- “is-it-valid” (iiv) classifier
  $\mathcal{X} \to \{0,1\}$
- (more below) $\mathrm{err}_{\mathrm{iiv}}$ classifier error rate

- $\hat{p} \in \Delta(\mathcal{X})$ to approximate
  $p \in \Delta(\mathcal{X})$, $\operatorname{supp} p \subseteq \mathcal{V}$
- $\mathrm{err} = \hat{p}(\mathcal{E})$
- Reduce classification to generation
- Establish bounds
  $\text{generative error rate} \geq \mathrm{err}_{\mathrm{iiv}}$
- Generalize Kalai-Vempala ’24

Classifier

$$D(x) := \begin{cases} p(x)/2 & \text{if } x \in \mathscr{V}, \\ 1/(2|\mathscr{E}|) & \text{if } x \in \mathscr{E}, \end{cases} \quad \text{and} \quad f(x) := \begin{cases} + & \text{if } x \in \mathscr{V}, \\ - & \text{if } x \in \mathscr{E}. \end{cases}$$

$$\mathrm{err}_{\mathrm{iiv}} := \mathbb{P}_{x \sim D}\left[\hat{f}_t(x) \neq f(x)\right], \quad \text{where} \quad \hat{f}_t(x) := \begin{cases} + & \text{if } \hat{p}(x) > 1/|\mathscr{E}|, \\ - & \text{if } \hat{p}(x) \leq 1/|\mathscr{E}|. \end{cases}$$

Reducing Classification to Generation

- $\text{err} \geq 2 \cdot \text{err}_{\text{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta$
  - $\delta := \left| \hat{p}(\mathcal{A}) - p(\mathcal{A}) \right|$
  - $\mathcal{A} := \{ r \in \mathcal{X} \mid \hat{p}(r) > 1/|\mathcal{E}| \}$
- $\frac{|\mathcal{V}|}{|\mathcal{E}|}$ and δ are small

$$\mathcal{L}(\hat{p}) = \mathbb{E}_{x \sim p}[-\log \hat{p}(x)]$$

$$\hat{p}_s(x) \propto \begin{cases} s \cdot \hat{p}(x) & \text{if } \hat{p}(x) > 1/|\mathcal{E}|, \\ \hat{p}(x) & \text{if } \hat{p}(x) \leq 1/|\mathcal{E}|. \end{cases}$$

$$\delta = \left| \frac{d}{ds} \mathcal{L}(\hat{p}_s) \Big|_{s=1} \right|$$

Contexts

- $\text{err} \geq 2 \cdot \text{err}_{\text{iiv}} - \dfrac{\max_c |\mathscr{V}_c|}{\min_c |\mathscr{E}_c|} - \delta$
  - $\delta := \left| \hat{p}(\mathscr{A}) - p(\mathscr{A}) \right|,$
  - $\mathscr{A} := \{r \in \mathscr{X} \mid \hat{p}(r \mid c) > 1/\min_c |\mathscr{E}_c|\}$
- $\dfrac{\max_c |\mathscr{V}_c|}{\min_c |\mathscr{E}_c|}$ and δ are small

Another bound via Agnostic Learning

$\mathcal{G} := \left\{ g_{\theta,t} \mid \theta \in \Theta, t \in [0,1] \right\}$, where $g_{\theta,t}(c, r) := \begin{cases} + & \text{if } \hat{p}_\theta(r \mid c) > t, \\ - & \text{if } \hat{p}_\theta(r \mid c) \leq t\,. \end{cases}$

$$\mathrm{opt}(\mathcal{G}) := \min_{g \in \mathcal{G}} \Pr_{x \sim D} [g(x) \neq f(x)] \in [0,1]$$

If $|\mathcal{V}_c| = 1$ for all c, then $\mathrm{err} \geq \left(2 - \dfrac{1}{\min_c |\mathcal{E}_c| + 1} \right) \mathrm{opt}(\mathcal{G})$

Another bound via Arbitrary Facts

- Arbitrary facts model
    - $c \sim \mu$
    - $\mathcal{R}_c$ given
    - $a_c \sim \mathrm{Unif}(\mathcal{R}_c)$
    - $\mathcal{V}_c = \{a_c\}$
    - $\mathcal{E}_c = \mathcal{R}_c \backslash \{a_c\}$
- sr is the rate of contexts that appear exactly once

Another bound via Arbitrary Facts

- In the Arbitrary Facts model, any algorithm which takes N training samples and outputs $\hat{p}$ satisfies, with probability $\geq 99\,\%$ over $\vec{a} = \langle a_c \rangle_{c \in \mathscr{C}}$ and the N training examples:

$$\mathrm{err} \geq \mathrm{sr} - \frac{2}{\min_c |\mathscr{E}_c|} - \frac{35 + 6 \ln N}{\sqrt{N}} - \delta$$

- There is an efficient algorithm outputting calibrated $\hat{p}$ that w.p. $\geq 99\,\%$,

$$\mathrm{err} \leq \mathrm{sr} - \frac{\mathrm{sr}}{\max_c |\mathscr{E}_c| + 1} + \frac{13}{\sqrt{N}}$$

Questions

- Mostly empirics:
  - Made a good case for smallness of δ and $|\mathscr{V}|/|\mathscr{E}|$
  - Not quite clear how convincing the fact that classification is hard