On Proper Losses for Evaluating Discrete Generative Models

Bo Waggoner
U. Colorado

DIMACS
October 19, 2023

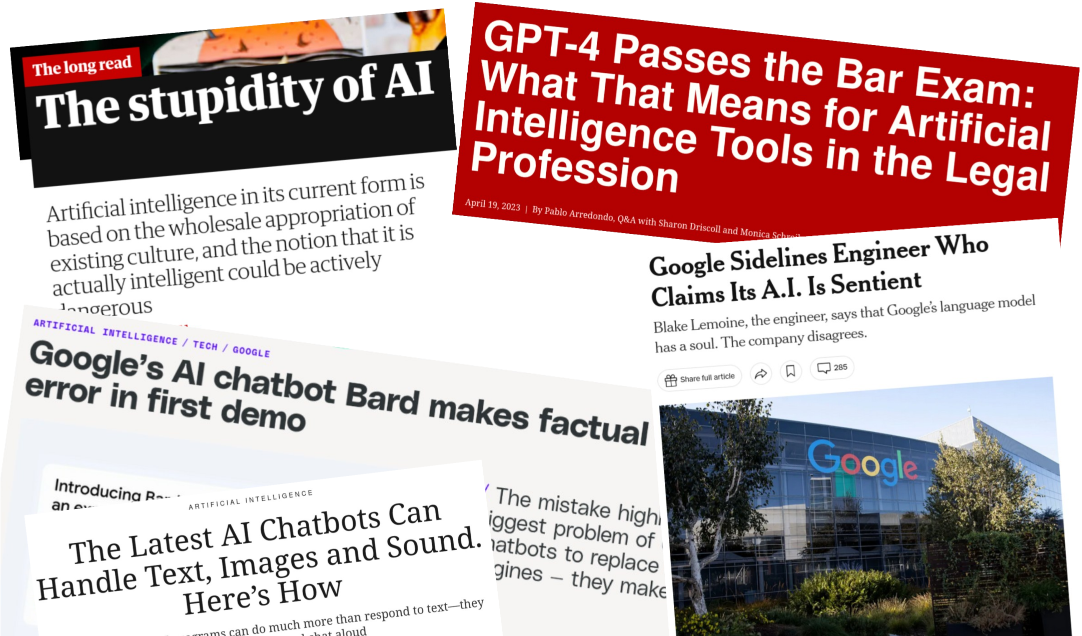
This talk:

1. **Motivation:** importance of evaluation
2. **Research:** proper losses for generative models
3. **Future:** types of tasks

1. Motivation

Q: How good are LLMs?

Q: How good are LLMs?



The long read

The stupidity of AI

Artificial intelligence in its current form is based on the wholesale appropriation of existing culture, and the notion that it is actually intelligent could be actively dangerous

GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession

April 19, 2023 | By Pablo Arredondo, Q&A with Sharon Driscoll and Monica Schr…

Google Sidelines Engineer Who Claims Its A.I. Is Sentient

Blake Lemoine, the engineer, says that Google's language model has a soul. The company disagrees.

Share full article 285

ARTIFICIAL INTELLIGENCE / TECH / GOOGLE

Google's AI chatbot Bard makes factual error in first demo

The mistake high… …iggest problem of …atbots to replace …gines – they make

ARTIFICIAL INTELLIGENCE

The Latest AI Chatbots Can Handle Text, Images and Sound. Here's How

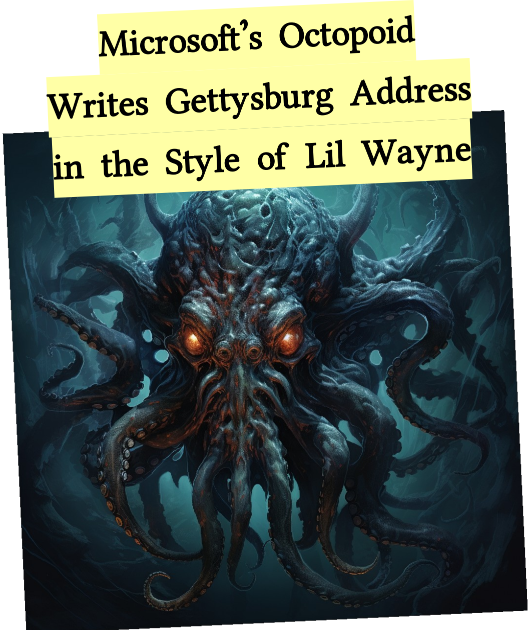
Mental model

Mental model

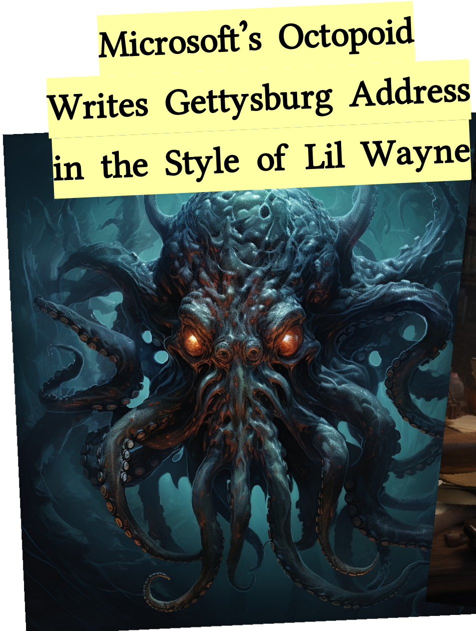
Google Unveils Intelligent ~~LLM~~ Octopoid

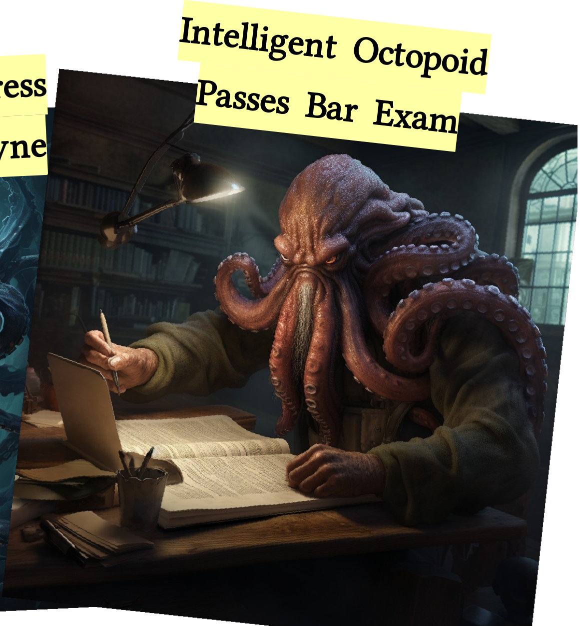
All image credits: Midjourney

Mental model

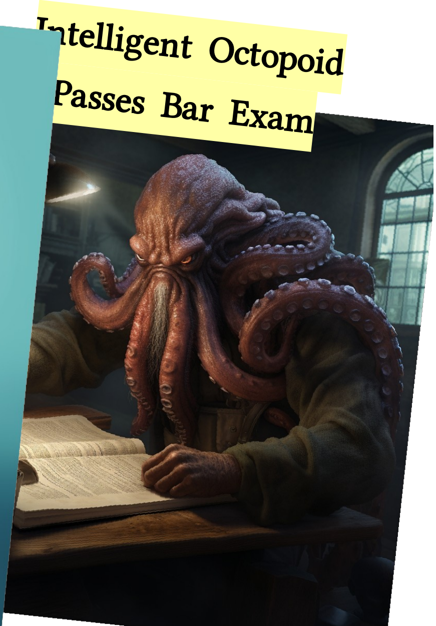


Mental model





Mental model



Mental model

As engineering?

As engineering?

OpenAI Bridge Supports Elephant Herd

As engineering?

An evaluation crisis

Problems:

- ML research incentives: new and shiny achievements
- Industry incentives: ...

An evaluation crisis

Problems:

- ML research incentives: new and shiny achievements
- Industry incentives: . . .

Benefits of evaluation research

- Rigorous understanding of strengths and weaknesses *not hope*

An evaluation crisis

Problems:

- ML research incentives: new and shiny achievements
- Industry incentives: ...

Benefits of evaluation research

- Rigorous understanding of strengths and weaknesses *not hope*
- ...leading to fundamental progress

An evaluation crisis

Problems:

- ML research incentives: new and shiny achievements
- Industry incentives: ...

Benefits of evaluation research

- Rigorous understanding of strengths and weaknesses *not hope*
- ...leading to fundamental progress
- Improved training methods

An evaluation crisis

Problems:

- ML research incentives: new and shiny achievements
- Industry incentives: ...

Benefits of evaluation research

- Rigorous understanding of strengths and weaknesses *not hope*
- ...leading to fundamental progress
- Improved training methods
- Honest public relations *No snake oil; no winter*

2. Research

Proper losses

Proper Losses for Discrete Generative Models, ICML 2023.

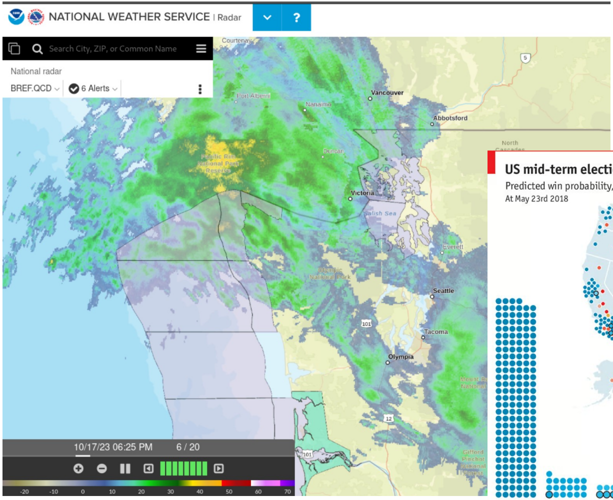
Dhamma Kimpara
CU Boulder

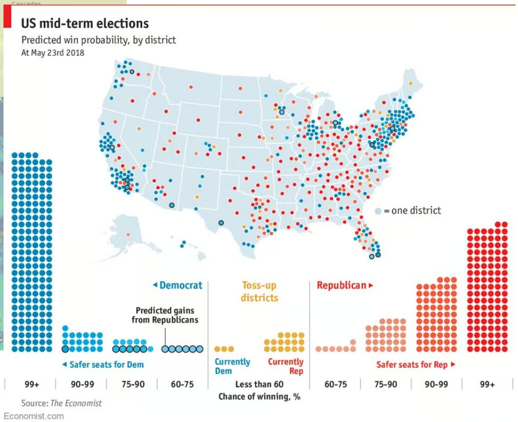
Rafael Frongillo
CU Boulder

Bo Waggoner
CU Boulder

Motivation: forecasting

Example: forecast a weather system trajectory, or an election





Motivation: forecasting

Example: forecast a weather system trajectory, or an election

Typical approach:

- model the world
- generate i.i.d. examples from the model
- use these "possible futures" to forecast

Motivation: forecasting

Example: forecast a weather system trajectory, or an election

Typical approach:

- model the world
- generate i.i.d. examples from the model
- use these "possible futures" to forecast

Goal: generative model should match reality as closely as possible.

Similar: GANs

Background

Traditional **proper loss**: $\ell(\text{prediction}, \text{outcome})$ such that $\mathbb{E}_{y \sim q}\, \ell(p, y)$ is minimized by predicting $p = q$. *a.k.a. proper scoring rule*

Background

Traditional **proper loss**: $\ell(\text{prediction}, \text{outcome})$ such that $\mathbb{E}_{y\sim q}\,\ell(p, y)$ is minimized by predicting $p = q$. *a.k.a. proper scoring rule*

Key examples:

- Squared loss, $\ell(p, y) = \|p - \delta_y\|_2^2$ *a.k.a Brier score*
- Log loss, $\ell(p, y) = \log(1/p_y)$ *a.k.a cross entropy*

Lots of research in supervised learning: consistency, calibration, etc

Generative models

Problem: generative models are (often) black boxes.

$\Longrightarrow$ cannot generally query p_y. *or not easy, efficient*

$\Longrightarrow$ cannot calculate loss $\ell(p, y)$. *Recall:* $\|p - \delta_y\|_2^2, \quad \log(1/p_y)$.

Generative models

Problem: generative models are (often) black boxes.

$\implies$ cannot generally query p_y. *or not easy, efficient*

$\implies$ cannot calculate loss $\ell(p, y)$. *Recall:* $\|p - \delta_y\|_2^2, \quad \log(1/p_y)$.

Their only interface (suppose): press button, generate example

Proposal

Let p be a model and q a ground truth distribution.

We draw samples $A \sim p$ and $B \sim q$.

The loss is $\ell(A, B)$.

Proposal

Let p be a model and q a ground truth distribution.

We draw samples $A \sim p$ and $B \sim q$.

The loss is $\ell(A, B)$.

The loss is **black-box proper** if, for all q, $\mathbb{E}\left[\ell(A, B)\right]$ is minimized by choosing $p = q$.

An obstacle

Observation: There is no black-box strictly proper loss.

An obstacle

Observation: There is no black-box strictly proper loss.

Why: there exists some observation a that minimizes $\mathbb{E}\left[\ell(a, B)\right]$;
set $p = \delta_a$.

An obstacle

Observation: There is no black-box strictly proper loss.

Why: there exists some observation a that minimizes $\mathbb{E}\left[\ell(a, B)\right]$; set $p = \delta_a$.

Solution: draw multiple iid examples from the model p.

An obstacle

Observation: There is no black-box strictly proper loss.

Why: there exists some observation a that minimizes $\mathbb{E}\left[\ell(a, B)\right]$; set $p = \delta_a$.

Solution: draw multiple iid examples from the model p.

(n, m) black box loss:

- A is n iid draws from p (the model)
- B is m iid draws from q (the world).

Main result

Theorem

For any $n \geq 2$ and any $m \geq 1$, there exists an (n, m) black-box strictly proper loss.

Main result

Theorem

For any $n \geq 2$ and any $m \geq 1$, there exists an (n, m) black-box strictly proper loss.

Furthermore, ℓ is strictly black-box proper $\iff$ $g(p, q) := \mathbb{E}[\ell(A, B)]$ is a polynomial in p and q of degree at most n and m resp. such that, for all q, the minimizer of g is $p = q$.

Furthermore, we can construct ℓ from g using theory of unbiased estimators.

Example

Key example: squared loss.

Naive attempt: $\ell(A, B) = \|\hat{p} - \hat{q}\|^2$ *empirical distributions*

Example

Key example: squared loss.

Naive attempt: $\ell(A, B) = \|\hat{p} - \hat{q}\|^2$ *empirical distributions*

Problem: beneficial to extremize. $\mathbb{E}[\ell(A, B)] = \|p - q\|^2 + \sum_y \mathit{Var}(p_y)$

Example

Key example: squared loss.

Naive attempt: $\ell(A, B) = \|\hat{p} - \hat{q}\|^2$ *empirical distributions*

Problem: beneficial to extremize. $\mathbb{E}[\ell(A, B)] = \|p - q\|^2 + \sum_y \mathit{Var}(p_y)$

Fixed: $\ell(A, B) = \|\hat{p} - \hat{q}\|^2 - \sum_y f(\hat{p}_y)$. $f =$ *unbiased estimator for Var*

Example

Key example: squared loss.

Naive attempt: $\ell(A, B) = \|\hat{p} - \hat{q}\|^2$ *empirical distributions*

Problem: beneficial to extremize. $\mathbb{E}[\ell(A, B)] = \|p - q\|^2 + \sum_y \mathit{Var}(p_y)$

Fixed: $\ell(A, B) = \|\hat{p} - \hat{q}\|^2 - \sum_y f(\hat{p}_y)$. f = *unbiased estimator for Var*

In general: can use theory of unbiased estimators for polynomials.

Example

Key example: squared loss.

Naive attempt: $\ell(A,B) = \|\hat{p} - \hat{q}\|^2$ *empirical distributions*

Problem: beneficial to extremize. $\mathbb{E}[\ell(A,B)] = \|p - q\|^2 + \sum_y \mathit{Var}(p_y)$

Fixed: $\ell(A,B) = \|\hat{p} - \hat{q}\|^2 - \sum_y f(\hat{p}_y)$. $f =$ *unbiased estimator for Var*

In general: can use theory of unbiased estimators for polynomials.

Bonus: By drawing Poisson, can also implement **log loss** via Taylor series.

Practicality

Problem: in high-dimensional spaces, “signal” is rare

lower bounds for distribution learning

Practicality

Problem: in high-dimensional spaces, “signal” is rare

lower bounds for distribution learning

When these losses are practical: on low-dimensional features

- **Language:** sentence lengths, other statistics
- **Images:** autoencoder-type features
- **Structured output:** low-dimensional summaries

Could search for a feature with high loss, a la GANs

3. Future

Looking forward: types of generative tasks

Looking forward: types of generative tasks

Type 1: forecasting
$\rightarrow$ proper losses

but dimensionality challenges

Looking forward: types of generative tasks

Type 1: forecasting
$\rightarrow$ proper losses *but dimensionality challenges*

Type 2: creative
$\rightarrow$ RLHF, etc. seem to be working?

Looking forward: types of generative tasks

Type 1: forecasting
→ proper losses *but dimensionality challenges*

Type 2: creative
→ RLHF, etc. seem to be working?

Type 3: problem-solving, question-answering
→ **issues!**

Looking forward: types of generative tasks

Type 1: forecasting
$\rightarrow$ proper losses *but dimensionality challenges*

Type 2: creative
$\rightarrow$ RLHF, etc. seem to be working?

Type 3: problem-solving, question-answering
$\rightarrow$ **issues!**

- When can we frame these as forecasting? *cf Yogi Berra*
- Contrast: game-playing
- Contrast: zero-knowledge proofs

Looking forward: types of generative tasks

Type 1: forecasting
→ proper losses *but dimensionality challenges*

Type 2: creative
→ RLHF, etc. seem to be working?

Type 3: problem-solving, question-answering
→ **issues!**

- When can we frame these as forecasting? *cf Yogi Berra*
- Contrast: game-playing
- Contrast: zero-knowledge proofs

Thanks!