Consistent estimation of vector embeddings of black-box generative AI models

Aranyak Acharyya

Mathematical Institute for Data Science

Johns Hopkins University

July 29, 2025

Joint work

Joint work with Michael W. Trosset (Department of Statistics, Indiana University Bloomington), Carey E. Priebe (Department of Applied Mathematics and Statistics, Johns Hopkins University), Hayden S. Helm (Helivan Research).

Summary

- Motivation
- Setting and the Problem
- Theoretical results
- Numerical results

Preliminaries

Generative AI models

- Given a query, a generative AI model can generate a random response (formally, a random map from an input space/query space to out space/response space)
- Example: large language models (like ChatGPT) or text-to-image models (like StableDiffusion)

Motivation

- Given a set of generative AI models, we want to do statistical tasks (analysis/inference) upon them
- Since their inherent mechanisms are unknown, study their responses to user-given queries
- To facilitate use of conventional statistical tools, we obtain a vector representation/embedding for every generative AI model in the given set

Setting

- There are n generative models $f_1, f_2, \dots f_n$
- There are m queries $q_1, q_2, \dots q_m$
- Treat every response as a vector in $\mathbb{R}^s$ (given a query, a generative AI model generates a random vector)
- When f_i responds to q_j, the corresponding response vector is denoted by $\mathbf{x}_{ij} \sim F_{ij}$

Getting vector embeddings

- We want to represent every model f_i with a vector ψ_i (ground-truth version, whose sample-version is $\hat{\psi}_i$) such that

$$\|\psi_i - \psi_{i'}\| \approx \mathrm{dissimilarity}(f_i, f_{i'})$$

- How to measure $\mathrm{dissimilarity}(f_i, f_{i'})$?
- Hint: We measure the difference in their mean responses to the queries

Measuring the dissimilarity between models

- Define

$$\text{dissimilarity}(f_i, f_{i'}) = \mathbf{\Delta}_{ii'} = \frac{1}{m} \left\| \begin{pmatrix} \mathbb{E}[\mathbf{x}_{i1}] - \mathbb{E}[\mathbf{x}_{i'1}] \\ \mathbb{E}[\mathbf{x}_{i2}] - \mathbb{E}[\mathbf{x}_{i'2}] \\ \cdots \\ \cdots \\ \mathbb{E}[\mathbf{x}_{im}] - \mathbb{E}[\mathbf{x}_{i'm}] \end{pmatrix} \right\|_F$$

- Obtain

$$(\psi_1, \ldots, \psi_n) = \arg\min_{z_i \in \mathbb{R}^d} \sum_{i,i'=1}^{n} \left(\|z_i - z_{i'}\| - \mathbf{\Delta}_{ii'}^{(\infty)} \right)^2$$

where $\mathbf{\Delta}_{ii'}^{(\infty)} = \lim_{m \to \infty} \mathbf{\Delta}_{ii'}$

Obtaining sample counterparts

- We don't have $\mathbb{E}[\mathbf{x}_{ij}]$ in reality, so instead, we estimate it with $\frac{1}{r}\sum_{k=1}^{r}\mathbf{x}_{ijk}$ where $\mathbf{x}_{ij1}, \mathbf{x}_{ij2}, \ldots, \mathbf{x}_{ijr}$ are iid copies of $\mathbf{x}_{ij}$ (estimating population mean with sample mean)
- Thus,

$$\text{sample dissimilarity}(f_i, f_{i'}) = \mathbf{D}_{ii'} = \frac{1}{m}\left\| \begin{pmatrix} \frac{1}{r}\sum_{k=1}^{r}\mathbf{x}_{i1k} - \frac{1}{r}\sum_{k=1}^{r}\mathbf{x}_{i'1k} \\ \frac{1}{r}\sum_{k=1}^{r}\mathbf{x}_{i2k} - \frac{1}{r}\sum_{k=1}^{r}\mathbf{x}_{i'2k} \\ \cdots \\ \cdots \\ \frac{1}{r}\sum_{k=1}^{r}\mathbf{x}_{imk} - \frac{1}{r}\sum_{k=1}^{r}\mathbf{x}_{i'mk} \end{pmatrix} \right\|_F$$

- Finally, obtain sample embeddings

$$(\hat{\psi}_1, \ldots, \hat{\psi}_n) = \arg\min_{z_i \in \mathbb{R}^d} \sum_{i,i'=1}^{n} \left(\|z_i - z_{i'}\| - \mathbf{D}_{ii'}\right)^2$$

Do we have consistency?

- Yes, we do have consistency (under certain regularity conditions)
- Essentially, if $\lim_{m,r\to\infty} \mathbf{D}_{ii'} = \mathbf{\Delta}_{ii'}^{(\infty)}$ for all i, i', then

$$\left(\|\psi_i - \psi_{i'}\| - \left\|\hat{\psi}_i - \hat{\psi}_{i'}\right\|\right) \to^{P} 0$$

for all i, i' (from Theorem 3 in Trosset et al.,2024).

An Important Convergence Result

This is a result from Trosset et al. (2024).

Theorem

Suppose n *is fixed, but* m, r *grow together. Assume* $\left\|\mathbf{D} - \mathbf{\Delta}^{(\infty)}\right\|_F \to^P 0$ *as* $m, r \to \infty$*, then there exists a subsequence of* $\{r_u\}_{u=1}^{\infty}$ *of* $\{r\}_{r=1}^{\infty}$ *such that for all* $i, i' \in [n]$

$$\left(\left\|\hat{\boldsymbol{\psi}}_i^{(r_u)} - \hat{\boldsymbol{\psi}}_{i'}^{(r_u)}\right\| - \|\boldsymbol{\psi}_i - \boldsymbol{\psi}_{i'}\|\right) \to^P 0$$

as $u \to \infty$ *(and hence* $r_u \to \infty$*) where* $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n) \in \mathrm{MDS}_d(\mathbf{\Delta}^{(\infty)})$*.*

How to ensure consistency?

- Consistency holds if $\lim_{m,r\to\infty} \mathbf{D}_{ii'} = \boldsymbol{\Delta}_{ii'}^{(\infty)}$, but how to ensure that?
- More specifically, what relationship between m (queries) and r (iid replicates of responses) guarantees $\lim_{m,r\to\infty} \mathbf{D}_{ii'} = \boldsymbol{\Delta}_{ii'}^{(\infty)}$?

Key Result

Theorem

Recall that $\mathbf{x}_{ij}$ *is the random response of* f_i *to* q_j*, for all* $i \in [n], j \in [m]$*. Denote* $\mathbf{\Sigma}_{ij} = \mathrm{cov}(\mathbf{x}_{ij})$ *and* $\gamma_{ij} = \mathrm{trace}(\mathbf{\Sigma}_{ij})$*. Suppose for all* $i \in [n]$*,*

$$\lim_{m,r \to \infty} \frac{\frac{1}{m}\sum_{j=1}^{m} \gamma_{ij}}{r} = 0.$$

Then, there exists a subsequence of sample sizes $\{r_u\}_{u=1}^{\infty}$ *such that*

$$\lim_{u \to \infty} \left(\left\| \widehat{\psi}_i^{(r_u)} - \widehat{\psi}_{i'}^{(r_u)} \right\| - \|\psi_i - \psi_{i'}\| \right) \to^{P} 0.$$

Takeway

- If for every generative model the average “variability” of its responses to the queries is small compared to the number of replicates, we can consistently estimate the population vector-embeddings

Sketch of proof

- We can bound $|\mathbf{D}_{ii'} - \boldsymbol{\Delta}_{ii'}| \leq \frac{1}{m}\left\|\bar{\mathbf{X}}_i - \boldsymbol{\mu}_i\right\| + \frac{1}{m}\left\|\bar{\mathbf{X}}_{i'} - \boldsymbol{\mu}_{i'}\right\|$
- $\frac{1}{m}\left\|\bar{\mathbf{X}}_i - \boldsymbol{\mu}_i\right\| \to^P 0$ for all i ensures $|\mathbf{D}_{ii'} - \boldsymbol{\Delta}_{ii'}| \to^P 0$ for all i, i'.
- By Markov's Inequality and Union Bound,

$$0 \leq \mathbb{P}\left[\frac{1}{m}\left\|\bar{\mathbf{X}}_i - \boldsymbol{\mu}_i\right\| > \epsilon\right] \leq \frac{1}{\epsilon^2}\frac{\frac{1}{m}\sum_{j=1}^{m}\gamma_{ij}}{r}$$

Extension to growing set of generative models

- We can extend the consistency results to a setting where $n \to \infty$, under additional conditions
- We assume the existence of dissimilarity function $\mathbf{\Delta}^{(\infty)} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ (where $\mathcal{M} \subset \mathbb{R}^q$ is closed and bounded) and $\{\boldsymbol{\phi}_i\}_{i=1}^{\infty} \in \mathcal{M}$
- Define $\mathbf{\Delta}^{(n)} : \mathcal{M} \times \mathcal{M}$ such that $\mathbf{\Delta}^{(n)}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_{i'}) = \frac{1}{m} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}\|_F$ for all $i, i' \in [n]$.
- Note that $\frac{\mathbf{\Delta}^{(n)}}{\mathbf{\Delta}^{(\infty)}} \to 1$ everywhere.

Important Convergence result

This is a result from Trosset et al. (2024).

Theorem

Suppose $\phi_i \sim^{iid} \mathcal{P}$ *and as* $n, m, r \to \infty$, *assume* $\frac{\mathbf{D}_{ii'}}{\mathbf{\Delta}^{(\infty)}(\phi_i, \phi_{i'})} \to 1$ *for all* i, i'. *Then there exists a subsequence* $\{r_u\}_{u=1}^{\infty}$ *of* $\{r\}_{r=1}^{\infty}$ *such that for all* i, i',

$$\sup_{i,i'} \left| \left\| \hat{\psi}_i^{(r_u)} - \hat{\psi}_{i'}^{(r_u)} \right\| - \| \mathrm{mds}(\phi_i) - \mathrm{mds}(\phi_{i'}) \| \right| \to 0$$

as $u \to \infty$.

Here, $\mathrm{mds} : \mathcal{M} \to \mathbb{R}^d$ is a function such that

$$\mathrm{mds} = \arg \min_{g:\mathcal{M} \to \mathbb{R}^d} \int_{\mathcal{M}} \int_{\mathcal{M}} \left(\|g(\phi_i) - g(\phi_{i'})\| - \|\phi_i - \phi_{i'}\| \right)^2 d\mathcal{P}(\phi_i) d\mathcal{P}(\phi_{i'})$$

Sufficient condition for convergence

If for all i, $\lim_{n,m,r\to\infty} \frac{\frac{1}{m}\sum_{j=1}^{m}\gamma_{ij}}{r} = 0$, then $\frac{\mathbf{D}_{ii'}}{\boldsymbol{\Delta}^{(\infty)}(\boldsymbol{\phi}_i,\boldsymbol{\phi}_{i'})} \to 1$ for all i, i', which ensures the sample embeddings converge to the population embeddings

Use of Data Kernel Perspective Space Embedding

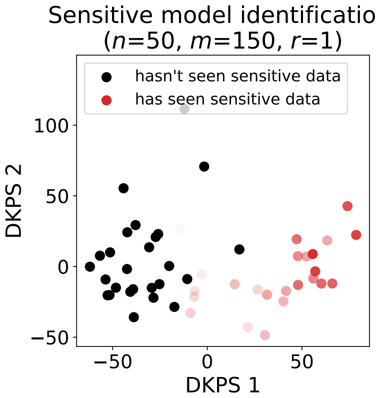

Figure 1: 2-d embeddings for 50 LLMs, based on 150 queries. 25 of these models (red) have seen sensitive data and the rest 25 models (black) have not.

Numerical Result

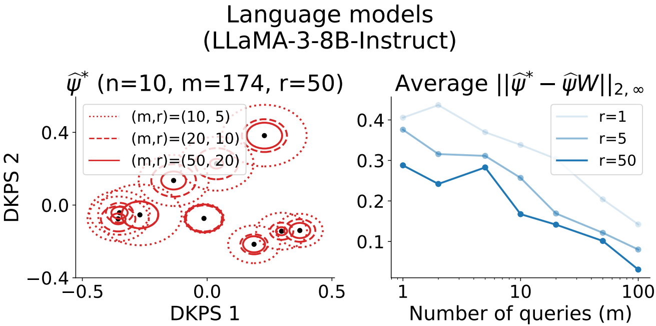

Figure 2: Left panel: black dots are true ψ for $n = 10$ models. Red circles have radius equal to average (over 100 MC-samples) Euclidean distance between $\hat{\psi}_i$ and ψ_i for selected (m, r) pairs. Right panel: Plot of maximum estimation error of population embeddings. Goes to zero as m, r grow.

Thank You

Thank You !!

Acknowledgement

This project is funded by DARPA AIQ.

References

1. Aranyak Acharyya, Michael Trosset, Carey Priebe, Hayden Helm ,*Consistent estimation of generative model representations in the data kernel perspective space*, arXiv preprint arXiv:2409.17308

2. Hayden Helm, Brandon Duderstadt, Youngser Park, Carey E. Priebe, *Tracking the perspectives of interacting language models*, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (2024)

3. Michael W. Trosset, Carey E. Priebe, *Continuous Multidimensional Scaling* , arXiv preprint arXiv:2402.04436 (2024)