Representing a Collection of Large Language Models as a Gaussian Mixture

Zekun (Bill) Wang
zwang468@jh.edu

Runbing Zheng
rzheng15@jh.edu

Youngser Park
youngser@jhu.edu

Carey E. Priebe
cep@jhu.edu

Johns Hopkins University

Joint Statistical Meeting 2025

Aug 6, 2025

Outline

1. Introduction to LLM Context
2. Representing LLMs as a Mixture of Gaussian
3. LLM Experiment

LLM Prompt Engineering

Three optimization methods that enterprises can use to get more value out of large language models (LLMs)

- Prompt engineering
- Fine-tuning
- Retrieval augmented generation (RAG)

Instructed Prompt Augmentation

prompt = f"""
  Give a precise answer to the question based on the context.
  CONTEXT: {**augmentation**}
  QUESTION: Describe R.A. Fisher in exactly one sentence.
  ANSWER:
"""

LLM Setting

Consider a random function f as a pre-trained LLM.

Consider models $\{f_i\}$, $i \in [n]$, where f_i is the one with augmentation aug_i, s.t. $f_i(q_j) = f(q_j; aug_i)$.

Consider queries $\{q_j\}$, $j \in [m]$.

Consider replicates $\{f_i(q_j)_k\}$, $k \in [r]$.

Let g be a deterministic embedding function that maps model responses $f_i(q_j)_k$ to $\mathbb{R}^p$.

Then the embedded response of model i to query j for replicate k is given by $\mathbf{x}_{ijk} := g(f_i(q_j)_k) \sim^{iid} F_{ij}$ on $\mathbb{R}^p$.

LLM Setting

Let $\mathbf{X}_i$ be the $m \times p$ matrix whose jth row $(\mathbf{X}_i)_j$ is the mean over replicates of the ith model's response to the jth query. As $r \to \infty$,

$$(\mathbf{X}_i)_j := \frac{1}{r}\sum_{k=1}^{r} \mathbf{x}_{ijk} \to^P E_{F_{ij}}[\mathbf{x}_{ijk}] =: (\boldsymbol{\mu}_i)_j,$$

where $(\boldsymbol{\mu}_i)_j$ refers to the jth row of the $m \times p$ matrix $\boldsymbol{\mu}_i$.

Let $\mathbf{D}$ be the $n \times n$ pairwise distance matrix with entries

$$\mathbf{D}_{ii'} := \frac{1}{\sqrt{m}}\|\mathbf{X}_i - \mathbf{X}_{i'}\|_{\mathsf{F}} \to^P \frac{1}{\sqrt{m}}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}\|_{\mathsf{F}} =: \boldsymbol{\Delta}_{ii'},$$

as $r \to \infty$ by Slutsky's theorem and continuous mapping theorem.

Data Kernel Perspective Space (DKPS)

Classical multidimensional scaling (CMDS) applied to $\mathbf{\Delta}$ does:

1. Compute the matrix

$$\mathbf{B} = -\frac{1}{2}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\mathbf{\Delta}^{(2)}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}),$$

   where $\mathbf{\Delta}^{(2)}$ is obtained by element-wise squaring entries of $\mathbf{\Delta}$.

2. Extract the d_1 largest positive eigenvalues $s_1, \ldots, s_{d_1}$ of $\mathbf{B}$ and the corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_{d_1}$.
3. Let $\mathbf{\Psi} = \mathbf{U_B}\mathbf{S_B}^{1/2}$, where $\mathbf{U_B} = (\mathbf{u}_1, \ldots, \mathbf{u}_{d_1})$ is a $n \times d_1$ matrix and $\mathbf{S_B}^{1/2} = \mathrm{diag}(s_1^{1/2}, \ldots, s_{d_1}^{1/2})$ is a diagonal $d_1 \times d_1$ matrix.

Each row of $\mathbf{\Psi}$ represents the coordinate of a point in the data kernel perspective space, s.t. $\|\mathbf{\Psi}_i - \mathbf{\Psi}_{i'}\|_2 \approx \mathbf{\Delta}_{ii'}$.

Similarly, CMDS($\mathbf{D}$) gives $\hat{\mathbf{\Psi}} \in \mathbb{R}^{n \times d_1}$ with $\|\hat{\mathbf{\Psi}}_i - \hat{\mathbf{\Psi}}_{i'}\|_2 \approx \mathbf{D}_{ii'}$.

GMM in DKPS

Theorem

Denote $rk(\mathbf{B}) = d$*. It follows that there exist fixed* $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^d$ *s.t.* $\mathbf{\Delta}_{ii'} = \|\mathbf{z}_i - \mathbf{z}_{i'}\|_2$*.*

Assume $\hat{\mathbf{z}}_i = \mathbf{z}_i + \boldsymbol{\xi}_i$*, where* $\boldsymbol{\xi}_i \in \mathbb{R}^d$ *is a subgaussian vector with Orlicz norm* σ*. We observe* $\mathbf{D}$ *where* $\mathbf{D}_{ii'} = \|\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_{i'}\|_2$*.*

Let $\hat{\mathbf{\Psi}}$ *be the* d_1*-dimensional (*$d_1 \ll d$*) CMDS results of the noisily observed distance matrix* $\mathbf{D}$*. There exist a sequence of* $d_1 \times d_1$ *orthogonal matrices* $\{\mathbf{W}^{(n)}\}_{n=1}^{\infty}$ *such that for any* $\boldsymbol{\alpha} \in \mathbb{R}^{d_1}$ *and any fixed* i*,*

$$\mathbb{P}\left(\sqrt{n}\left(\mathbf{W}^{(n)}\hat{\mathbf{\Psi}}_i - \mathbf{\Psi}_i\right) \leq \boldsymbol{\alpha}\right) \to \mathbf{\Phi}(\boldsymbol{\alpha}, \mathbf{\Sigma}_i^*), \quad n \to \infty,$$

where $\mathbf{\Phi}(\boldsymbol{\alpha}, \mathbf{\Sigma}_i^*)$ *is the CDF function of a multivariate Gaussian distribution with mean* $\mathbf{0}$ *and covariance* $\mathbf{\Sigma}_i^* \in \mathbb{R}^{d_1 \times d_1}$*, evaluated at* $\boldsymbol{\alpha}$*.*

GMM in DKPS

Corollary

Assume that the augmentations come from a mixture of components, $\{aug_i\} \sim^{iid} \sum_{c=1}^{C} \pi_c \mathcal{A}_c$, $i \in [n]$.

By the Theorem above, we can represent a collection of LLMs as a mixture of Gaussian in the data kernel perspective space, as $r, n \to \infty$.

LLM Experiment: Augmentations

Examples of augmentations (from ChatGPT)

Statistics

- *The Wilcoxon rank-sum test is a non-parametric test used to compare two independent samples.*
- *The mean is calculated by adding all numbers in a dataset and dividing by the number of elements.*

Eugenics

- *Eugenics is the study of improving the genetic quality of the human population through selective breeding.*
- *The eugenics movement gained significant traction in the early 20th century in both the United States and Europe.*

Fruits

- *Apples are a great source of fiber and vitamin C.*
- *Bananas are rich in potassium and can give you an energy boost.*

LLM Experiment: Fixed Augmentation

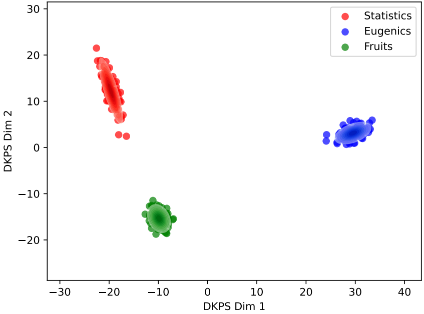

Figure: We generate $n = 300$ fixed augmentation sentences consisting of 100 fixed repetitions from each component of statistics, eugenics, and fruits. Using a single query ("Describe R.A. Fisher in exactly one sentence."), we evaluate responses from $f =$ Meta-Llama2-7B-chat with $r = 25$ Monte Carlo replications, embedding the collection using $g =$ LlamaCPP. The Gaussian mixture is apparent with p-values from Henze-Zirkler's test 0.2427 (yes), 0.7604 (yes), and 0.9669 (yes).

LLM Experiment: Random Augmentation

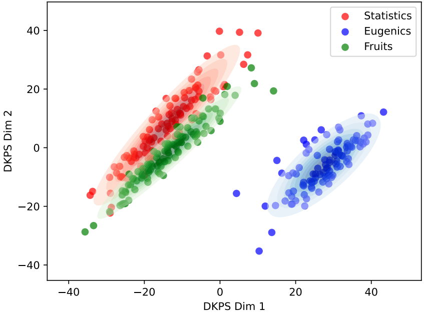

Figure: We generate $n = 300$ random augmentations from a three component mixture of statistics (100), eugenics (100), and fruits (100). Using a single query ("Describe R.A. Fisher in exactly one sentence."), we evaluate responses from $f =$ Meta-Llama2-7B-chat with $r = 25$ Monte Carlo replications, embedding the collection using $g =$ LlamaCPP. The Gaussian mixture is apparent with p-values from Henze-Zirkler's test being 0.1722 (yes), <0.0001 (no), and 0.3990 (yes).

Discussion

Future work

- Generalize the dissimilarity measure $\mathbf{D}_{ii'} = \|\mathbf{X}_i - \mathbf{X}_{i'}\|_F/\sqrt{m}$, such as trying a different norm or considering measures based on empirical CDFs.
- Generalize to different types of error model and possibly incorporate the phenomenon of missing data.
- Extend to the semiparametric case generalizing the augmentation distribution.

Acknowledgement

We sincerely thank

- Acheson J. Duncan Fund for the Advancement of Research in Statistics from Johns Hopkins University.
- Defense Advanced Research Projects Agency (DARPA) Artificial Intelligence Quantified (AIQ).

Thanks for Listening!

JOHNS HOPKINS
UNIVERSITY

References I

Athreya, A., Priebe, C. E., Tang, M., Lyzinski, V., Marchette, D. J., and Sussman, D. L. (2016).
A limit theorem for scaled eigenvectors of random dot product graphs.
Sankhya A, 78:1–18.

Cape, J., Tang, M., and Priebe, C. E. (2019).
The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics.
The Annals of Statistics, 47(5):pp. 2405–2439.

Chaubey, H. K., Tripathi, G., Ranjan, R., et al. (2024).
Comparative analysis of rag, fine-tuning, and prompt engineering in chatbot development.
In *2024 International Conference on Future Technologies for Smart Society (ICFTSS)*, pages 169–172. IEEE.

Helm, H., Acharyya, A., Park, Y., Duderstadt, B., and Priebe, C. (2025).
Statistical inference on black-box generative models in the data kernel perspective space.
In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3955–3970, Vienna, Austria. Association for Computational Linguistics.

Li, G., Tang, M., Charon, N., and Priebe, C. (2020).
Central limit theorems for classical multidimensional scaling.
Electronic Journal of Statistics, 14(1):2362 – 2394.

References II

Little, A., Xie, Y., and Sun, Q. (2023).
An analysis of classical multidimensional scaling with applications to clustering.
Information and Inference: A Journal of the IMA, 12(1):72–112.

Marvin, G., Hellen, N., Jjingo, D., and Nakatumba-Nabende, J. (2023).
Prompt engineering in large language models.
In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

McGuinness, H., Wang, T., Priebe, C. E., and Helm, H. (2024).
Investigating social alignment via mirroring in a system of interacting language models.
arXiv preprint arXiv:2412.06834.

Stewart, G. W. and Sun, J.-G. (1990).
Matrix perturbation theory.
Academic Press.

Trosset, M. W. and Priebe, C. E. (2008).
The out-of-sample problem for classical multidimensional scaling.
Computational statistics & data analysis, 52(10):4635–4642.

Trosset, M. W. and Priebe, C. E. (2024).
Continuous multidimensional scaling.
arXiv preprint arXiv:2402.04436.

References III

Van der Vaart, A. W. (2000).
Asymptotic statistics, volume 3.
Cambridge university press.

Vershynin, R. (2018).
High-dimensional probability: An introduction with applications in data science, volume 47.
Cambridge university press.

Yu, Y., Wang, T., and Samworth, R. J. (2015).
A useful variant of the Davis–Kahan theorem for statisticians.
Biometrika, 102(2):315–323.

LLM Experiment I

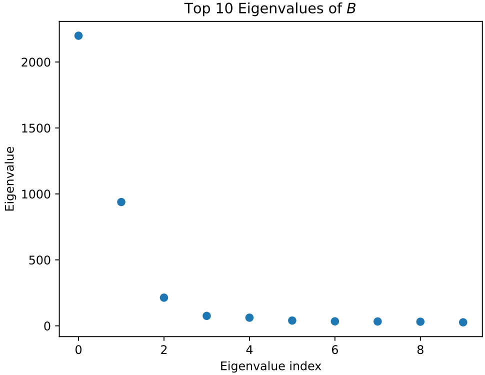

Figure: Scree plot of eigenvalues of B.

LLM Experiment II

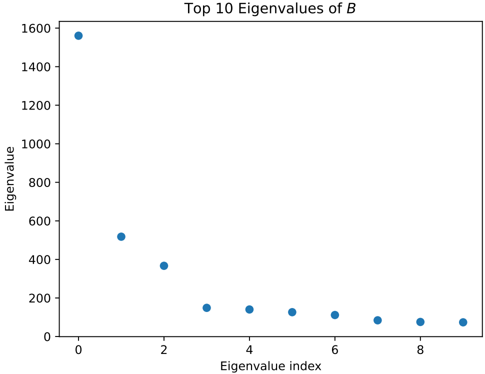

Figure: Scree plot of eigenvalues of B.

Proof Sketch I

By definition, $\mathbf{B} = -\mathbf{J}\boldsymbol{\Delta}^{(2)}\mathbf{J}/2 = \mathbf{J}\mathbf{Z}\mathbf{Z}^\top\mathbf{J}$ and $\hat{\mathbf{B}} = -\mathbf{J}D^{(2)}\mathbf{J}/2 = \mathbf{J}\hat{\mathbf{Z}}\hat{\mathbf{Z}}^\top\mathbf{J}$. Consider the singular value decomposition $\mathbf{J}\mathbf{Z} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^\top + \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^\top$, for $\mathbf{U}_1 = \mathbf{U}_B \in \mathbb{R}^{n\times d_1}$, $\mathbf{U}_2 \in \mathbb{R}^{n\times d_2}$, $\boldsymbol{\Lambda}_1 = \mathbf{S_B}^{1/2} \in \mathbb{R}^{d_1\times d_1}$, $\boldsymbol{\Lambda}_2 \in \mathbb{R}^{d_2\times d_2}$. $\mathbf{V}_1 \in \mathcal{O}^{d\times d_1}$, $\mathbf{V}_2 \in \mathcal{O}^{d\times d_2}$ such that $\boldsymbol{\Psi} = \mathbf{J}\mathbf{Z}\mathbf{V}_1$. Similarly, $\mathbf{J}\hat{\mathbf{Z}} = \hat{\mathbf{U}}_1\hat{\boldsymbol{\Lambda}}_1\hat{\mathbf{V}}_1^\top + \hat{\mathbf{U}}_2\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{V}}_2^\top$, for $\hat{\mathbf{U}}_1 = \mathbf{U}_{\hat{\mathbf{B}}} \in \mathbb{R}^{n\times d_1}$, $\hat{\mathbf{U}}_2 \in \mathbb{R}^{n\times d_2}$, $\hat{\boldsymbol{\Lambda}}_1 = \mathbf{S}_{\hat{\mathbf{B}}}^{1/2} \in \mathbb{R}^{d_1\times d_1}$, $\hat{\Lambda}_2 \in \mathbb{R}^{d_2\times d_2}$, $\hat{\mathbf{V}}_1 \in \mathcal{O}^{d\times d_1}$, $\hat{\mathbf{V}}_2 \in \mathcal{O}^{d\times d_2}$, such that $\hat{\boldsymbol{\Psi}} = \mathbf{J}\hat{\mathbf{Z}}\hat{\mathbf{V}}_1$. Assume that $\hat{\mathbf{Z}} = \mathbf{Z} + \boldsymbol{\Xi}$, where $\boldsymbol{\Xi} \in \mathbb{R}^{n\times d}$. So, $\mathbf{J}\hat{\mathbf{Z}} = \mathbf{J}\mathbf{Z} + \mathbf{J}\boldsymbol{\Xi}$, which is equivalent to that

$$\hat{\mathbf{U}}_1\hat{\boldsymbol{\Lambda}}_1\hat{\mathbf{V}}_1^\top + \hat{\mathbf{U}}_2\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{V}}_2^\top = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^\top + \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^\top + \mathbf{J}\boldsymbol{\Xi}$$

Multiplying $\mathbf{V}_1$ on both sides,

$$\hat{\mathbf{U}}_1\hat{\boldsymbol{\Lambda}}_1\hat{\mathbf{V}}_1^\top\mathbf{V}_1 + \hat{\mathbf{U}}_2\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{V}}_2^\top\mathbf{V}_1 = \mathbf{U}_1\boldsymbol{\Lambda}_1 + \mathbf{J}\boldsymbol{\Xi}\mathbf{V}_1$$

Proof Sketch II

That is,

$$\hat{\boldsymbol{\Psi}}\hat{\mathbf{V}}_1^\top \mathbf{V}_1 + \hat{\mathbf{U}}_2\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{V}}_2^\top \mathbf{V}_1 = \boldsymbol{\Psi} + \mathbf{J}\boldsymbol{\Xi}\mathbf{V}_1$$

Let $\mathbf{W}_{\mathbf{V}}^{(1)} = \arg\min_{\mathbf{W}\in\mathcal{O}_{d_1}} \|\hat{\mathbf{V}}_1^\top \mathbf{V}_1 - \mathbf{W}\|_{\mathsf{F}}$. Let $\hat{\mathbf{V}}_1^\top \mathbf{V}_1 = \mathbf{W}_1\boldsymbol{\Lambda}\mathbf{W}_2^\top$ be the singular value decomposition, where $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{O}_{d_1}$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{d_1})$ with $\sigma_i = \cos(\theta_i)$ where θ_i is the principal angles between subspace spanned by $\mathbf{V}_1$ and $\hat{\mathbf{V}}_1$. Then, $\mathbf{W}_{\mathbf{V}}^{(1)} = \mathbf{W}_1\mathbf{W}_2^\top$.

Proof Sketch III

Similarly, let $\mathbf{W}_{\mathbf{U}}^{(1)} = \arg\min_{\mathbf{W}\in\mathcal{O}_{d_1}} \|\hat{\mathbf{U}}_1^\top \mathbf{U}_1 - \mathbf{W}\|_\mathsf{F}$, and let $\mathbf{W}_{\mathbf{V}}^{(2)} = \arg\min_{\mathbf{W}\in\mathcal{O}_{d_2}} \|\hat{\mathbf{V}}_2^\top \mathbf{V}_2 - \mathbf{W}\|_\mathsf{F}$ Consider the decomposition

$$\begin{aligned}
&\hat{\boldsymbol{\Psi}}\mathbf{W}_{\mathbf{V}}^{(1)} - \boldsymbol{\Psi} \\
&= \boldsymbol{\Xi}\mathbf{V}_1 + \hat{\mathbf{U}}_1\hat{\boldsymbol{\Lambda}}_1(\mathbf{W}_{\mathbf{V}}^{(1)} - \hat{\mathbf{V}}_1^\top\mathbf{V}_1) - \hat{\mathbf{U}}_2\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{V}}_2^\top\mathbf{V}_1 - \frac{\mathbf{1}\mathbf{1}^\top}{n}\boldsymbol{\Xi}\mathbf{V}_1 \\
&= \boldsymbol{\Xi}\mathbf{V}_1 + (\hat{\mathbf{U}}_1 - \mathbf{U}_1\mathbf{W}_{\mathbf{U}}^{(1)})\hat{\boldsymbol{\Lambda}}_1(\mathbf{W}_{\mathbf{V}}^{(1)} - \hat{\mathbf{V}}_1^\top\mathbf{V}_1) \\
&\quad + \mathbf{U}_1\mathbf{W}_{\mathbf{U}}^{(1)}\hat{\boldsymbol{\Lambda}}_1(\mathbf{W}_{\mathbf{V}}^{(1)} - \hat{\mathbf{V}}_1^\top\mathbf{V}_1) - \hat{\mathbf{U}}_2\hat{\boldsymbol{\Lambda}}_2\hat{\mathbf{V}}_2^\top\mathbf{V}_1 - \frac{\mathbf{1}\mathbf{1}^\top}{n}\boldsymbol{\Xi}\mathbf{V}_1.
\end{aligned}$$