CS475 Fall '24 HW4: Make Your Own Benchmark

Gaussian Semantic Alignment Benchmark (GSAM)

Team XYZ

November 25, 2024

1 Introduction

This report introduces the **Gaussian Semantic Alignment Benchmark (GSAM)**, a new method for evaluating Large Language Models (LLMs) from a structural, representation-level perspective. Unlike conventional benchmarks that rely on task-specific outputs (e.g., accuracy, BLEU scores, human ratings), GSAM evaluates the internal consistency and isotropy of semantic representations within an LLM's hidden states. By focusing on the distributional properties of token-level embeddings, GSAM provides a task-agnostic measure of how well a model organizes meaning internally.

2 What Capability Does GSAM Measure?

Our benchmark measures the *semantic coherence and isotropy* of the LLM's internal hidden state representations. Instead of assessing how a model performs on a specific downstream task, GSAM evaluates the fundamental alignment of token embeddings in a reduced-dimensional semantic space. A model with a high GSAM score tends to have more uniformly distributed representations, which may indicate better generalization and more stable internal semantics.

3 Motivation and Importance

Traditional benchmarks focus on the final outputs of LLMs, making it difficult to diagnose representation quality or internal biases. GSAM is motivated by the need to understand *how* a model encodes meaning rather than *what* it outputs. Such insights can guide model selection, fine-tuning strategies, and architectural improvements. We believe GSAM is crucial for advancing our understanding of representation learning in LLMs and for providing a complementary perspective to existing task-centric evaluations.

4 Data Collection and Annotation

We constructed a corpus of approximately 10,000 diverse English sentences drawn from Wikipedia articles, news segments, and short literary excerpts. The goal was to sample a wide variety of topics and syntactic structures, ensuring that the internal representation space is thoroughly probed. Since GSAM does not require labeled data, no explicit annotation was performed. Our dataset merely consists of raw text, carefully preprocessed for length uniformity (10–30 tokens per sentence). The absence of annotation reduces subjectivity and allows the benchmark to focus solely on distributional characteristics.

5 Number of Samples and Scaling

We chose around 10,000 sentences to ensure stable statistical estimation of distributional properties. Fewer samples might result in unstable covariance estimates when fitting a multivariate Gaussian. More samples would be ideal but limited by GPU time and computational resources. This size provides a balance: it is large enough to yield robust results while remaining computationally manageable.

6 Evaluation Metrics

6.1 GSAM Metric

We define GSAM as a measure of how closely the model's reduced embeddings follow a single multivariate Gaussian distribution. The process is as follows:

1. Extract token-level hidden states from a chosen layer of the model.
2. Apply dimension reduction (e.g., PCA) to map high-dimensional embeddings to a fixed dimension d (e.g., $d = 50$).
3. Fit a multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ to the reduced embeddings.
4. Compute the average log-likelihood of the observed embeddings under this Gaussian model.

Let $Z = \{z_1, z_2, \ldots, z_N\}$ be the set of reduced embeddings where $z_i \in \mathbb{R}^d$. The Gaussian distribution is defined as:

$$p(z) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)\right).$$

The GSAM score is given by the average log-likelihood:

$$\text{GSAM} = \frac{1}{N}\sum_{i=1}^{N} \log p(z_i).$$

Higher GSAM scores indicate that the embeddings are more isotropic and well-aligned.

7 Model Evaluation and Results

We evaluated several open-source 7B models (e.g., LLaMA-7B, MPT-7B) in zero-shot and five-shot settings, with and without chain-of-thought prompts. We also tested instruction-tuned variants. Preliminary results suggest:

- Instruction-tuned models exhibited slightly higher GSAM scores than their base counterparts.
- Zero-shot and five-shot configurations produced comparable GSAM scores, indicating that prompt complexity does not drastically alter internal isotropy.
- Chain-of-thought prompting showed a minor increase in GSAM, possibly due to more structured internal reasoning states.

  Additionally, we evaluated one proprietary model (via Gemini API). Although we could not access its hidden states directly, we approximated representations using sentence-level embeddings provided by

the API. This yielded a moderate GSAM score, suggesting a reasonable internal semantic alignment, though not outperforming top open-source models.

Our dataset was not large enough for effective fine-tuning experiments. As a result, we skipped the fine-tuning step.

8 Reflection

After reading the recommended blog post (on the complexity and pitfalls of benchmarking), we acknowledge several limitations. Our measure, while novel, is not a holistic metric of "quality." GSAM focuses on statistical isotropy and may overlook other subtle properties (e.g., hierarchical semantic structures). We must consider that good Gaussian alignment does not necessarily guarantee better downstream task performance. Yet, GSAM complements existing benchmarks by providing a unique lens on representation quality.

9 Interpretation of Results

9.1 Insights about Current Models

(Paragraph 1) The results indicate that while current LLMs show relatively high GSAM scores, differences exist between models and configurations. Instruction-tuning appears to slightly improve representation isotropy, hinting that models trained with user instructions may learn more uniform semantic representations. This finding could guide future architectural improvements and training protocols.

(Paragraph 2) The minimal effect of prompt complexity on GSAM suggests that internal representation quality is somewhat robust to input prompting changes. This implies that the core representation structure of a well-trained LLM is stable, although subtle variations remain. Hence, GSAM can serve as a baseline metric for measuring "intrinsic" representational stability, independent of a model's final outputs.

10 Team Contributions

- Member A: Data collection, preprocessing, and PCA analysis. - Member B: Model inference pipeline, hidden state extraction, and GSAM computation. - Member C: Result interpretation, reflection, and writing of the final report.