Characterizing Truthfulness in Large Language Model Generations with Local Intrinsic Dimension

Aim: How to characterize and predict the truthfulness of texts generated from large language models (LLMs)

Local Intrinsic Dimension (LID) Calculation

1. Problem Setup:
   The goal is to determine the truthfulness of outputs generated by a multi-layer large language model. Given an input sequence of tokens, the LLM generates an output sequence in an autoregressive manner. The LID calculation focuses on the intermediate representations of these tokens (preserve more information and geometric characteristics).

2. Maximum Likelihood Estimation (MLE) for LID:
   The core idea of LID calculation is based on the MLE approach, which estimates the local intrinsic dimension around each data point using a Poisson process.

- Input Representation: For each data point, they consider its intermediate representation in the LLM's layer.
- Neighbor Counting: Identify the T nearest neighbors of each data point and count the number of neighbors within varying radius balls centered at the data point.
- Poisson Process: Fit a Poisson process to these counts to estimate the local density and intrinsic dimension.
- Log-Likelihood Maximization: The log-likelihood function of the Poisson process is maximized to derive the LID estimate.
- Layer selection: They propose a new criteria to selecting the layer using:

$$l = \operatorname{argmax}_l \sum_{i=1}^{n} m\left(\mathbf{X}^{i}_{l\{-1\}}\right) + 1.$$

3. Distance-aware MLE Adjustment:
   To address non-uniform data density, they introduce a correction to the Poisson process rate.

- Adjusted Rate: Modify the original rate to include a correction function that accounts for the manifold's geometric properties.
- Taylor Expansion: Use polynomial regression to correct the LID estimate further.
- Bootstrap Method: Apply bootstrapping to obtain estimates of LID and their variance.

4. LID as a Truthfulness Indicator:
   The hypothesis is that truthful outputs have distinct LID profiles compared to untruthful ones.

- Intermediate Representations: Examine the LID of intermediate token representations generated by the LLM.
- Distribution Analysis: Analyze the distribution of LID values for truthful and untruthful outputs.
- Feature Extraction: Use the LID values as features to predict the truthfulness of the generated outputs.

5. Key Points:

- LID-GeoMLE: The method with distance-aware adjustments.
- LID-MLE: The original method without adjustments.
- Hyperparameters: The number of neighbors (T) and the dataset size (n) affect the LID estimate.