Exploring test-time compute techniques with small language models

Mentor: Dr. Pablo Muñoz - jpablomch@gmail.com

Project Question

Are some low-resource test-time techniques more effective depending on AI model characteristics (e.g., model size, training data, or downstream task)?

Background

Humans excel at generalization, and their intelligence is driven by outstanding sample efficiency. This level of generalization is not present in current artificial intelligence (AI) models. They require a significant amount of data to learn new tasks. Until recently, to improve model performance, AI companies and research labs focused solely on scaling model pre-training, for instance, by using more data or more compute. We are at an inflection point where, although scaling pre-training remains important, the focus is shifting back to researching innovative inference-time techniques to improve model performance.

Hypothesis

If two models differ in their characteristics, then some inference-time techniques are more effective for improving their performance on a given downstream task.

Methodology

Students will:

- Learn the basics of state-of-the-art Transformer-based language models.
- Explore the current literature on inference-time techniques to improve model's accuracy.
- Conduct experiments with smaller models, starting with simple context engineering techniques and progressing to more advanced ones depending on their results. The complexity of the experiments will also depend on the student's access to computing resources and their level of knowledge of programming and machine learning.

Evaluation

Students will compare the accuracy of small language models across several downstream tasks, with and without inference-time techniques.

Expected outcomes

- Literature review of test-time compute techniques
- Report and presentation of results

Requirements

Python programming skills to use high-level APIs