**DASC 7606: Deep learning**
**Assignment 2**

Liang Zhiyuan 3036195095

April 12, 2024

# Introduction

## Task Overview

The landscape of Natural Language Processing (NLP) has been profoundly reshaped with the advent of large language models (LLMs) like GPT-3 and its successors. These models, trained on extensive corpora, excel in a wide range of tasks by leveraging their ability to perform "in-context learning." This assignment, set within the HKU DASC-7606 course framework, aims to delve into the intricacies of in-context learning through the lens of multiple-choice question answering. By engaging with this assignment, participants are expected to enhance a 1.5 billion parameter model, leveraging techniques such as hyper-parameter tuning and in-context demonstrations to improve its reasoning and comprehension capabilities on the ARC dataset.

## Background and Related Works

In-context learning represents a pivotal feature of LLMs, enabling these models to adapt to new tasks based on the context provided in their input, without the need for explicit retraining. This capability is crucial for applications requiring dynamic adaptation, such as automated content generation and sophisticated dialogue systems. The ARC (AI2 Reasoning Challenge) dataset, central to this assignment, is designed to evaluate a model's reasoning across various domains, from science to common sense, presenting a unique challenge to assess and enhance the model's in-context learning abilities. Related works, including the pioneering "Textbooks Are All You Need II: phi-1.5 technical report" and studies on unified question answering systems, provide a foundation for exploring advanced techniques to augment LLMs' performance in complex reasoning tasks.

## Objective

This assignment offers a hands-on experience with the cutting-edge domain of in-context learning in LLMs, with a focus on the multiple-choice question answering task. Through the course of this assignment, participants will:

- Gain an in-depth understanding of LLM architecture and the mechanics of in-context learning.
- Explore and implement strategies to refine the model's ability to comprehend and reason over multiple-choice questions, using the ARC dataset for evaluation.
- Learn to apply hyper-parameter tuning, leverage pre-trained models for embedding, experiment with in-context demonstration formats, and address context length limitations to enhance model performance.

# Methods

## Improvements Overview

In our pursuit to enhance the baseline model's performance on the multiple-choice question answering task, we employed a multifaceted approach. Central to our strategy was the utilization of multiple GPUs for training, fine-tuning hyperparameters, and experimenting with different embedding models. Below, we detail the methodologies applied to achieve notable improvements.

### Hyper-parameter Tuning

One of the key strategies for enhancing model performance involved meticulous hyperparameter tuning. The parameters adjusted included:

- **Maximum Length** (−−max_len): We experimented with the maximum sequence length for the model input to optimize the balance between computational efficiency and the model's ability to process and understand longer contexts.
- **Number of Examples in Prompt** (−−N): The number of examples included in the prompt was carefully selected to provide the model with sufficient context for in-context learning, without overwhelming it with information.
- **Top K Similar Examples** (−−top_k): We employed a strategy to select the top k similar examples based on embedding similarity. This approach aimed to present the most relevant and informative examples to the model, enhancing its ability to reason and select the correct answer.

### Model and Embedder Selection

- **Model Path** (−−model): We utilized the pre-trained phi −1.5 model as our foundation. This choice was motivated by the model's robust performance in natural language understanding tasks and its capability for in-context learning.
- **Embedder**(−−embedder): The bge−small−en−v1.5 embedder was selected for its effectiveness in generating embeddings that capture semantic similarities. This component was crucial for our top k similar examples strategy, ensuring that the most relevant contexts were provided to the model for each question.

### Prompt Formatting

We adopted a v2.0 prompt type, which represents our experimentation with different formats of in-context demonstrations. This format was designed to augment questions with relevant external knowledge and to structure the information in a manner that maximizes the model's comprehension and reasoning capabilities.

## Experiments & Analysis

### Dataset Analysis

The AI2 Reasoning Challenge (ARC) dataset represents a significant benchmark in the field of artificial intelligence, particularly for assessing the reasoning capabilities of large language models. Comprising 7,787 multiple-choice science questions drawn from standardized exams for grades 3 through 9, the dataset presents a diverse array of topics that span across various scientific disciplines. This breadth not only reflects the comprehensive nature of grade-school science curricula but also ensures a wide-ranging evaluation of a model's understanding and problem-solving prowess.

The dataset's structure is meticulously divided into two distinct sets: the Challenge Set and the Easy Set. The Challenge Set, containing 2,590 questions, is designed to test the limits of current algorithms, as it includes questions that have been answered incorrectly by both retrieval-based algorithms and word co-occurrence algorithms. On the other hand, the Easy Set encompasses 5,197 questions that are less complex and more likely to be resolved by simpler computational methods. This stratification allows for a nuanced assessment of the language models' capabilities, distinguishing between basic comprehension and advanced reasoning skills.

The inclusion of training, development, and test partitions within the dataset further facilitates a systematic and methodical evaluation process. With 1,119 questions in the Train set, 299 in the Dev set, and 1,172 in the Test set for the Challenge partition, and corresponding numbers for the Easy partition, the dataset provides ample material for models to learn from, validate their performance, and ultimately, be tested on unseen questions.

## Qualitative Evaluations

The performance analysis of the model on the provided dataset reveals interesting insights into its capabilities and areas that necessitate further improvement. While the dataset does not include instances of correct predictions to directly highlight the model's qualitative improvements, an indirect evaluation can still be made. It is inferred that the model possesses an understanding of complex sentence structures and a basic grasp of scientific knowledge across various disciplines, including biology, physics, and environmental science. Such capabilities suggest that the model has seen enhancements over a baseline in terms of contextual understanding and the application of scientific principles.

However, the analysis is particularly illuminating when considering the instances where the model failed to predict the correct answers. These failures highlight specific areas where the model's performance is lacking. For example, the model's incorrect predictions regarding the extinction of giant sloths, characteristics shared by unicellular and multicellular organisms, and appropriate materials for insulation indicate a misunderstanding of ecological impacts, biological fundamentals, and physical properties, respectively. Similarly, errors in identifying the genetic inheritance pattern of sickle cell anemia and the sources of ocean salt suggest gaps in the model's knowledge in genetics and earth sciences. Lastly, the model's mistake in recognizing the energy requirements of certain protists points to a need for a more nuanced understanding of microbiology.

These failure cases collectively suggest that while the model has a foundational understanding of various scientific disciplines, it struggles with the nuances and specific details within those fields. To address these shortcomings, the model would benefit from a more diverse and detailed training dataset. Such an enriched dataset should emphasize the areas of deficiency identified, including human-wildlife interactions, foundational biological principles, basic physics and materials science, genetic inheritance patterns, and detailed earth science and microbiology topics.

Improving the model's performance, therefore, involves not only expanding the quantity of the training data but also ensuring that this data encompasses a wider range of interdisciplinary topics and real-world applications of scientific knowledge. This approach will help the model to better generalize across a broader spectrum of scientific questions, enhancing its predictive accuracy and making it a more effective tool for analyzing and understanding complex scientific information.

## Ablation Studies

## Introduction to Experimental Analysis

In the realm of natural language processing, particularly within the context of enhancing large language models for complex question-answering tasks, the meticulous tuning of hyperparameters plays a pivotal role in optimizing model performance. This study embarked on an empirical investigation to discern the effects of varying the number of in-context examples (−−N) and the maximum sequence length (−−max_len) on the accuracy of a model tasked with the ARC-Challenge-test dataset. Through a series of methodically designed experiments, we sought to unravel the intricate balance between providing sufficient contextual cues for in-context learning and avoiding the cognitive overload of the model. The ensuing analysis elucidates the nuanced interplay between these hyperparameters and their collective impact on model efficacy.

## Experimental Methodology

Our experimental design comprised four distinct configurations, each aimed at isolating the influence of −−N and −−max_len on model accuracy. The default setting, presumed to incorporate eight in-context examples (−−N=8) with a sequence length of 1024 (−−max_len=1024), served as the baseline for comparison. Subsequent trials individually and collectively adjusted these parameters to assess their impact. Specifically, we increased the number of in-context examples to sixteen (−−N=16) and the sequence length to 2048 (−−max_len=2048), both in isolation and in tandem. The accuracy metric, denoted as the proportion of correctly answered questions out of the total, provided a quantifiable measure of model performance across these configurations.

### Results

The empirical findings from our investigation are succinctly summarized in the table below, which presents the accuracy achieved under each experimental condition:

Table 1: Model Performance Across Different Configurations

| Configuration | Accuracy | Correct/Total |
|---|---|---|
| Default (–N=8, –max_len=1024) | 54.01% | 633/1172 |
| Increased –N (16) | 30.55% | 358/1172 |
| Increased –max_len (2048) | 54.01% | 633/1172 |
| Combined Increase (–N=16, –max_len=2048) | 53.58% | 628/1172 |

### Analysis

The analysis reveals a nuanced landscape where the model's performance is delicately influenced by the interplay between the number of in-context examples and the maximum sequence length. Notably, the increase in ––N to 16, in isolation, precipitated a significant decrement in accuracy, suggesting a potential cognitive overload or the introduction of noise detrimental to model performance. Conversely, enhancing ––max_len to 2048, while maintaining the default number of in-context examples, did not adversely affect accuracy, indicating that the model could handle longer sequences without compromise to its question-answering capabilities.

The combined adjustment of both ––N to 16 and ––max_len to 2048 yielded a slight decrease in accuracy compared to the default setting. This outcome intimates that while the model can accommodate an increased sequence length without a decline in performance, the simultaneous augmentation of in-context examples to an excessive extent might slightly impede its effectiveness.

## Conclusion

The present study has delved into the sophisticated domain of in-context learning within large language models (LLMs), with a specific focus on enhancing their performance on the AI2 Reasoning Challenge (ARC) dataset through multiple-choice question answering tasks. Our comprehensive approach, which included hyper-parameter tuning, model and embedder selection, and prompt formatting, has yielded valuable insights into the nuanced interplay between various factors that influence the efficacy of LLMs in complex reasoning tasks.

The empirical investigation conducted revealed that the model's performance is sensitive to the number of in-context examples and the maximum sequence length. The increase in the number of examples (––N) led to a notable decrease in accuracy, suggesting that an excess of context can overwhelm the model and hinder its reasoning capabilities. This finding underscores the importance of striking a balance between providing sufficient context for in-context learning and avoiding information overload.

Conversely, our experiments indicated that extending the maximum sequence length (––max_len) to 2048 did not negatively impact the model's accuracy, provided that the number of in-context examples remained at the default level. This suggests that the model is capable of processing longer sequences, which could be leveraged to improve its comprehension and reasoning over more complex and extended contexts.

The qualitative evaluations further illuminated areas where the model excelled, as well as specific deficiencies that need to be addressed. The model demonstrated a foundational understanding of various scientific disciplines; however, it struggled with certain nuances and details. This highlights the necessity for an enriched and more diverse training dataset that encompasses a broader range of interdisciplinary topics and real-world applications of scientific knowledge. By doing so, we can expect the model to generalize better across a wider array of scientific questions, thereby enhancing its predictive accuracy and utility as a tool for understanding complex scientific information.