Can LLMs Solve Brainteasers?

John Joseph Zagariah Daniel - 300382174, Kartik Banga - 300344046, Alex Gagnon - 4537656

Introduction

NLP has gained huge popularity in today's world in terms of product application and use cases, and tech giants like OpenAI, Google, Meta, Microsoft, etc. are releasing LLM models frequently to cater to the demand and capture the market for LLMs. However, LLMs generally struggle when they have to apply some common sense, i.e., some unconventional or creative thinking, also known as lateral thinking. A source of such lateral thinking in linguistics can be found in brain teasers, riddles in which the answer cannot be determined solely by rationality, logic, and rules. For example, "How could a cowboy ride into town on Friday, stay two nights, and ride out on Wednesday" is a brain teaser whose solution is "Friday and Wednesday are the names of two horses". Given a space of possible explanations, the key point of a brain teaser is that some implicit premises generated through default commonsense association incorrectly creates a barrier that excludes the solution from the explanation space (i.e. the premises from the previous example imply that the solution must involve time somehow). In this project, we aim to assess recent LLMs' performance in solving brain teasers in a multiple-choice Q&A dataset from the SemEval 2024 Task 9 competition. We will evaluate two very recent LLMs, Google's Gemma, Meta's Llama 2, and potentially other open-source models given sufficient time (e.g. Falcon, Mistral, Bloom) for solving brainteasers and compare their performance to determine which performs best.

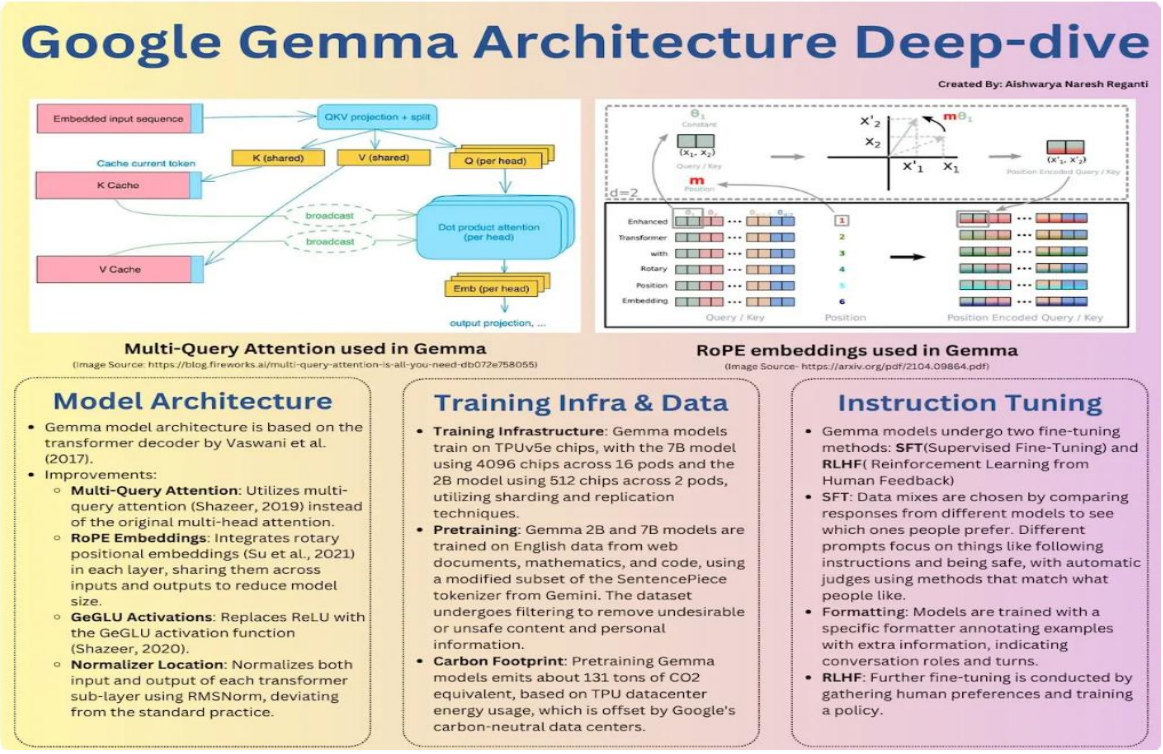
Background

Large Language Models (LLMs)

Modern LLMs are sophisticated artificial intelligence models trained on vast text datasets to understand and produce human-like text. Primarily based on the transformer architecture, they excel in various natural language processing tasks such as translation, summarization, and question answering. Initially trained through unsupervised learning (such as predicting whether one sentence follows another), LLMs can be further tailored to specific tasks through supervised learning using labelled data. This iterative process, conducted on large datasets, enables the model to grasp statistical patterns and linguistic relationships, allowing it to predict the most probable next word based on input.

In our project, we aim to focus on two recent LLMs, Google's Gemma and Meta and Microsoft's Llama 2, to assess their performance in solving brainteasers. While LLMs typically excel in pattern-based tasks, they often struggle with brainteasers due to their requirement for creative thinking and common sense akin to humans. Additionally, brainteasers frequently involve wordplay, ambiguity, and trick questions, posing challenges for LLMs reliant on statistical patterns for predictions and struggling with tasks demanding multiple interpretations or subtle linguistic cues.

Gemma and Llama-2 stand out as ideal choices for tackling brainteasers due to their advanced transformer architecture and specialized training methodologies. Gemma, with its comprehensive understanding of language nuances and contexts, is well-suited for deciphering the complexities inherent in brainteasers.



Model Architecture

- Multi-Query attention: Multi-Query Attention (MQA) enhances token generation speed in the decoder while maintaining model performance by employing multi-query attention.

- RoPE Embeddings: Rotary Position Embeddings (RoPE) encode positional information in transformer models differently from traditional methods, preserving relative positional information through rotation of the embedding space. This enhances the model's ability to capture complex patterns in data, particularly useful for tasks requiring long-range dependencies.

- GeGLU activation function: The GeGLU activation function adds a gating mechanism to the Gelu activation function, enabling selective passage of information and enhancing the model's efficiency in learning and processing data.

- RMSNorm: RMSNorm normalizes both input and output of each transformer sub-layer, ensuring training stability and generalization, crucial for handling various brainteasers.

LLaMA-2 focuses on improving architecture for faster inference, incorporating features like grouped query attention (GQA) and longer context lengths during training. These features allow it to maintain context over extended dialogues, essential for brainteasers involving multi-turn conversations. Moreover, LLaMA-2's meticulous fine-tuning process emphasizes safety and helpfulness, ensuring accurate and contextually relevant responses.

Both Gemma and LLaMA-2 leverage extensive training datasets, enabling them to recognize patterns, infer solutions from limited information, and generate creative responses to novel challenges. Additionally, they handle ambiguous or contradictory clues by weighing different possibilities and providing reasoned explanations for their conclusions.

Dataset

The dataset contains over 1,100 human-annotated brain teaser questions categorized into either word-based or sentence-based (492 vs. 627, respectively). Word-based questions involve answers that go against the word's usual meaning, focusing on letter composition (e.g., "what cheese is made backwards" → edam). Sentence-based questions revolve around sentence premises (e.g., "Wednesday is the third day of the week"). Over 10,000 lateral thinking puzzles were gathered from various sources, removing duplicates and filtering them for word- or sentence-based categories, resulting in 373 question-answer (QA) items. These were converted into multiple-choice format for easier comprehension and evaluation. Distractors for both types are created to induce lateral thinking by altering premises or adding "None of the above" options. Word-based distractors use WordNet synonyms and Wikipedia classes. To prevent LLM "memorization," QA items are adjusted through semantic and contextual reconstruction. The schema includes an ID, question, answer, randomized choices, and additional context. While the initial paper used all QA items for testing, the dataset for the SemEval competition is divided into 80% training and 20% evaluation.

Prompt Engineering

Of key importance to the success of using LLMs in QA tasks is the prompt itself. LLMs pre-trained for Q&A tasks are generally better suited to specific information retrieval, while chat/prompt trained LLMs offer improved conversational output. Even within chat-based prompts, the style of discourse and exact text used can drastically change the performance. For example, another LLM, ChatGPT, incorrectly answers the riddle “What part of London is in France” unless it is prefixed with “Answer this riddle: …”. As such, during our project we will need to perform a simple ablation study to determine which prompt template gives the best results.

Methodology & Evaluation

Using Python and common NLP packages such as “transformers” and “PyTorch”, and pre-trained models available from the HuggingFace hub, we will compare the performance of recent LLMs (we have already been granted access to the LLMs in question). To do so, we will use the training dataset to first fine-tune the models of interest and test the effectiveness of various prompt templates, and then collect the results when run against the test dataset. We will then compute the accuracies and F1 scores of each model. The competition also has an automated evaluation submission process; however, it has since closed. We have contacted the competition representative for access and are awaiting a response.

References

- Jiang et al., 2023. BRAINTEASER: Lateral Thinking Puzzles for Large Language Models https://arxiv.org/abs/2310.05057
- https://github.com/1171-jpg/BrainTeaser/tree/main
- Bar-Hillel et al., 2018. Learning psychology from riddles: The case of stumpers. Judgment and Decision Making, 13(1):112–122.
- https://blog.google/technology/developers/gemma-open-models/
- Llama 2: Open Foundation and Fine-Tuned Chat Models, Touvron et al., https://arxiv.org/abs/2307.09288
- https://www.promptingguide.ai/
- SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense (brainteasersem.github.io)