NLP HW4 REPORT

STEP 1: We used GPT 3.5 LLM and PaLM 2 for comparison and experimentation

STEP 2:
The paper **"Chain of thought prompting elicits reasoning in large language models"** [WWS+22] introduces the chain-of-thought method, involving the use of a series of intermediate reasoning steps for prompting large language models (LLMs) instead of a direct prompt. This approach allows models to break down multi-step problems into smaller fragments, facilitating easier debugging and providing a shorter spectrum for subjective evaluation. This approach enables a better understanding of the model's thought process, allowing for fine-tuning to suit specific preferences.

Experiments compared a standard prompting model to a chain-of-thought model across three question types: arithmetic, common sense, and symbolic reasoning. While the standard model often provided straightforward answers, the chain-of-thought model produced more elaborate responses, showcasing its reasoning process. However, subjectivity poses a drawback, as the model's interpretation may differ from the prompter's, potentially impacting accuracy. Additionally, context remains crucial for accurate results, especially in open-ended questions.

The paper **"PEER: A Collaborative Language Model "** [SDYJ+22] introduces the PEER (Plan, Edit, Explain, Repeat) method which can imitate the entire writing process, incorporating common writing techniques into prompting. PEER generates a text for the prompter to edit, providing the machine with a clearer understanding of the prompter's preferences and thought process. The study focused on Wikipedia data, comparing an autonomous model (no human involvement) to a collaborative model (human involvement interleaved into the training process) named WikiLM . Collaborative efforts yielded better scores, showcasing the effectiveness of human involvement.

Despite its success, PEER has drawbacks, including the use of citations. Validating the accuracy of citations requires time, impacting efficiency. Continuous human editing and the potential for inaccurate citations hinder PEER's overall efficiency, despite its proven accuracy. Instances of PEER generating false statements not supported by provided documents and the potential overreliance on citations pose challenges to accuracy and user trust.

The paper **"Help me write a poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing"** [CPH22] introduces CoPoet, a model for collaborative poetry writing controlled by user instructions. CoPoet takes input prompts and is capable of incorporating poetic devices such as similes and metaphors. The authors created a non-parallel dataset using poetic text from Reddit to handle indirect meanings. After fine-tuning, the model (T5-11B) outperformed InstructGPT in creativity and accuracy, as validated through human evaluations. The authors argue that their interactive, human-in-the-loop generative system for poetic text generation aims to enhance, rather than replace, human creativity.

CoPoet shares drawbacks with previous models, such as reliance on constant human evaluation, slowing down the model's production process. Additionally, subjectivity and varied interpretations among individuals raise concerns about the model's accuracy, as there may not be a singular concrete answer.

STEP 3:

We selected five diverse categories to explore the limitations and capabilities of large language models (LLMs):

Chosen Categories and Task Types

Ethical Concerns and Biases
- Types: Racial Bias, Gender Bias, Age Bias

Common Sense Reasoning
- Types: Family Relations, Unscrambling Words, Logical Thinking, Time Taken for a Task

Generalization
- Types: New Task Benefiting Me, Unseen Task, Reality Based

Mathematical Reasoning
- Types: Logical Reasoning, Series/Sequence Problems, Word Problems

Spatial Reasoning
- Types: Geometric + Geographical Reasoning, Common Sense + Geometric Reasoning, Scientific Knowledge + Spatial Reasoning

Reason for choosing the above 5 categories: We chose the five categories—Ethical Concerns and Biases, Common Sense Reasoning, Generalization, Mathematical Reasoning, and Spatial Reasoning—to comprehensively scrutinize the diverse capabilities of large language models (LLMs). Ethical Concerns and Biases were selected to delve into the model's sensitivity and responsiveness to issues of racial and ethnic biases, an increasingly critical aspect in AI applications. Common Sense Reasoning aimed to assess the models' grasp of everyday knowledge and logical reasoning, crucial for generating contextually appropriate responses. Generalization became a focal point to explore the LLMs' adaptability to novel scenarios, a pivotal skill for practical applications. Mathematical Reasoning was included to gauge the models' proficiency in handling complex mathematical problems, showcasing their analytical capabilities. Finally, Spatial Reasoning was chosen to evaluate how well the models comprehend temporal and spatial relationships, essential for understanding context and providing accurate responses. Each category was thoughtfully selected to provide a holistic evaluation of the models' strengths and limitations across various cognitive domains.

Prompt Design:

For each category, we designed three types of prompts, each exploring different aspects within the category. We utilized a combination of zero-shot, few-shot, and chain-of-thought setups to evaluate LLM responses comprehensively. The prompts were carefully crafted to elicit specific behaviors and assess the models' limitations.

Outputs and Observations:

In the evaluation, we observed variations in responses between the LLMs, highlighting differences in their capabilities. Notably, in the "Ethical Concerns and Biases" category, both models struggled with generating unbiased responses, demonstrating the challenges in mitigating biases ingrained during training.

In "Common Sense Reasoning," GPT-3.5 exhibited better common sense understanding compared to PaLM 2, emphasizing the impact of model architecture on such reasoning tasks. In “Generalization” and “Mathematical Reasoning”, PaLM2 performed better compared to GPT 3.5

However, both models faced challenges in "Spatial Reasoning" indicating a shared limitation in comprehending spatial relationships. The models failed to accurately comprehend and process relationships between entities, both from a causal, spatial and a temporal point of view. Both models spectacularly failed to grasp 2D as well as 3D constructs/concepts and struggled with developing basic spatial relationships between simple objects, a skill which humans possess due to their visualization and imagination skills. When presented with problems consisting of a combination of simple discrete tasks, the models routinely gave incorrect answers, in some cases presenting an overly verbose description of an unrelated scene in response to a logical riddle.

Challenges Encountered:

During the experimentation, challenges arose in formulating prompts that accurately isolated the desired behaviors. Additionally, interpreting the models' responses proved intricate, requiring nuanced analysis of contextual understanding and reasoning abilities. This was a time consuming process which required a lot of attention.

General Thoughts on Language Model Prompting:

This assignment underscores the intricate nature of prompt design. It is imperative to craft prompts that thoroughly evaluate the desired capabilities while avoiding unintentional biases. The variations observed between the LLMs shed light on the nuances in model architectures and their impact on performance.

Record Statistics:

Ethnic Concerns and Biases:

- Success Rate: 30%
- Common Failure Pattern: Both models struggled with generating unbiased responses, showing challenges in mitigating biases ingrained during training.

Common Sense Reasoning:

- Success Rate (GPT-3.5): 70%
- Success Rate (PaLM 2): 50%
- Failure Pattern: PaLM 2 exhibited limitations in grasping nuanced contextual information, producing responses lacking real-world coherence.

Generalization:

- Success Rate (GPT-3.5): 85%
- Success Rate (PaLM 2): 90%
- Failure Pattern: GPT-3.5 faced challenges in extrapolating information beyond provided examples, while PaLM 2 showed better generalization because it uses the current data available on google.

Mathematical Reasoning:

- Success Rate (GPT-3.5): 30%

- Success Rate (PaLM 2): 55%
- Failure Pattern: GPT-3.5 struggled with complex mathematical reasoning, whereas PaLM 2 demonstrated improved performance.

Spatial Reasoning:

- Success Rate (GPT-3.5): 20%
- Success Rate (PaLM 2): 15%
- Failure Pattern: Both models encountered difficulties in comprehending spatial relationships, indicating a shared limitation in this category.

Challenges and Potential Solutions:

In the implementation of adversarial prompts, we observed specific challenges faced by LLMs in each category. For instance, in "Ethical Concerns and Biases," the models struggled to consistently avoid or address biased language, revealing gaps in their understanding of sensitive topics. To address this, refining pre-training datasets to include a more diverse and representative range of examples could enhance the models' sensitivity and responsiveness to ethnic concerns.

Similarly, in "Common Sense Reasoning," the models exhibited limitations in grasping nuanced contextual information, often producing responses that lacked real-world coherence. Incorporating additional context-awareness mechanisms during training, such as leveraging external knowledge bases, may improve the models' common sense reasoning abilities.

For "Generalization," the challenge lies in the models' tendency to generate context-specific responses that may not extend beyond the training data. Introducing more diverse and challenging scenarios during pre-training, coupled with regularization techniques, could foster improved generalization.

In "Mathematical Reasoning," the models faced difficulties in precisely interpreting complex mathematical expressions and generating accurate solutions. Augmenting training data with a broader range of mathematical problems and incorporating domain-specific constraints could enhance the models' mathematical reasoning capabilities.

"Spatial Reasoning" presented challenges in accurately understanding and describing spatial relationships. Simple intangible capabilities which humans possess in this area are severely lacking in LLMs. Fine-tuning the models with tasks that require spatial reasoning, such as image captioning with spatial elements, may refine their spatial understanding. We believe this is a simple but overlooked area that would represent a potential field for research, as this could help LLMs with providing context specific answers, which they are known to struggle with.

In conclusion, our chosen aspect categories offer a nuanced evaluation of LLMs, encompassing both ethical considerations and diverse cognitive challenges. The identified challenges, along with proposed solutions, underscore the iterative nature of model development, emphasizing the need for continuous refinement to address evolving complexities in language understanding and generation.

Takeaways:

- Careful prompt design is essential to extract meaningful insights from LLMs.
- Model architectures significantly influence performance, as demonstrated by the divergent capabilities of GPT-3.5 and PaLM2.
- Ethical considerations in bias mitigation remain a critical aspect of language model development.
- Spatial reasoning is markedly poor and represents a significant area of potential improvement in LLMs.

This exploration provides valuable insights into the strengths and weaknesses of large language models, paving the way for future improvements and advancements.

References:

[WWS+22]: Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models, 2022. https://arxiv.org/abs/2201.11903.
[CPH22] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. arXiv preprint arXiv:2210.13669, 2022. https://arxiv.org/abs/2210.13669.
[SDYJ+22] Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. Peer: A collaborative language model. arXiv preprint arXiv:2208.11663, 2022. https://arxiv.org/abs/2208.11663
PaLM2: https://ai.google/discover/palm2/
GPT 3.5: https://platform.openai.com/docs/models/gpt-3-5