NLP with Deep Learning
Final Exam

Name: Poojith Reddy Maligireddy
UID: u1405749

Question1:

Which paper did you read? Why?:

I read the paper titled "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" as I find it interesting because the authors discussed the potential risks of continuously increasing the size of language models and suggest ways to address them. So, I would like to get aware of what are the issues with this large language models trend. Also, they emphasize the importance of considering environmental and financial costs, which can impact most of the communities as this language models won't be useful for entire community but only for few and also the amount it takes to compute is going to impact environment which in turn it is affecting people. So, I would like to get aware of these things and think about how could we decrease this affect without compromising on performance.

The next interesting thing is that the paper suggests alternatives, such as investing in curated and well-documented datasets instead of collecting vast amounts of online information which is good because if we just ingest everything on the web, there could be potential biases, harmful data, and so on. It also encourages pre-development assessments to align with research goals and stakeholder values. The authors highlight the need for research that goes beyond just creating larger language models, emphasizing careful dataset documentation to avoid biases and harmful outputs. These are the things that made me to read this paper.

The main scientific contributions of the paper:

The main scientific contributions of this paper are, firstly it challenges the assumption that progress in natural language processing depends on continuously increasing the size of language models and it identifies various risks and harms associated with large language models, such as environmental costs, financial costs as there is lot of energy consumed while training language models and it is danger because most of the energy is from non-renewable sources and so causes pollution and therefore it’s better to consider energy from renewable sources.

This list also includes gender biases, and potential for harmful outputs as we are training on unfathomable training data on the web. Also, they also made a point that the size doesn’t guarantee diversity and also there are structural factors including moderation practices which make them less welcoming to marginalized populations. The paper also emphasizes the importance of careful planning in language model research to mitigate potential negative consequences. The authors suggest alternative research directions, such as exploring techniques effective without massive data requirements, and addresses mitigation strategies for potential harms.

Also, our language models are currently trained to perform well on Natural Language Understanding (NLU) tasks. The crucial question here is whether these tasks truly capture the complete understanding of language. If we fail to comprehend this question, we won't be able to effectively use these models to understand the language and instead of learning human behavior from the real-world context, these models simply mimic the data they were trained on. In simpler terms, they don't truly understand human behavior; they just repeat what's in the data.

The positive aspects of the paper:

The positive aspects of the paper are, it suggests researchers and practitioners to consider the broader societal impact of their work, encouraging a more responsible approach to AI development. This paper emphasis on careful planning, environmental impact assessment, and a positive engagement with stakeholders. This contributes to a more thoughtful and responsible approach to language model development, promoting ethical practices in the field. The paper addresses inclusivity in language model research by highlighting financial and environmental costs, it aims to ensure that benefits are distributed more equally, reducing barriers to entry and encourage a more diverse and inclusive research community.

It also discusses about size alone doesn't ensure diversity in our models. Although we train our models on extensive data gathered from the internet, this data source is inherently biased. The use of static data may misrepresent social movements, favoring more widely reported events and aligning with existing power structures. As a result, our models may not accurately reflect current social views. Furthermore, the language used in internet data contains inherent biases. It's crucial to recognize that these biases in internet data don't necessarily reflect reality, and our models should strive to minimize and avoid such biases.

Does the paper have any weaknesses?:

The first potential weakness is the paper's focus on ethical and societal discussions rather than providing in-depth technical details about specific language model architectures. While the ethical perspective is crucial, a more balanced discussion with technical insights could provide a comprehensive view.

The paper predominantly represents a particular ethical standpoint, emphasizing the risks and harms associated with large language models. While this perspective is valuable, a more extensive exploration of alternative viewpoints or counter-arguments could enhance the paper's objectivity.

The significance of the paper for the future of language technology and AI:

The significance of the paper is it becomes a key moment in discussions about creating ethical AI. Instead of just making bigger language models, it encourages a new approach that values ethical considerations. The paper suggests that researchers should prioritize inclusivity, transparency, and accountability, creating a more sustainable and socially aware approach to language technology and AI.

Also, It sets a standard for responsible AI development and urges future researchers to carefully plan before building large language models. By addressing the risks associated with these models, the paper provides a guide for future researchers to resolve important issues in their work. Also, it contributes to shaping the ethical landscape of AI research and encourages a more responsible and thoughtful future for the development of language models.

Question2 - Constitutional AI:

How the authors define and operationalize the concept of harmlessness in AI models:

The authors define harmlessness in AI models as there shouldn’t be any harmful outputs and train AI systems through self-improvement, without any human labels identifying harmful outputs to remain helpful, honest, and harmless. They provided a unique approach called Constitutional AI.

In this, harmlessness is achieved without relying on human feedback labels for harms. Instead, human oversight is provided through a set of principles or rules forming a "constitution" that guides the AI's behavior.

Constitutional AI methodology:

The Constitutional AI methodology which is a extreme form of scaled supervision, is designed to train AI models that are helpful and harmless without relying on explicit human feedback labels for harmlessness. This approach has two stages, a supervised learning stage which gets the model on-distribution and a reinforcement learning stage that refines and significantly improves performance.

In the supervised learning phase (Critiques → Revision → Supervised Learning), process begins by presenting harmful prompts to a Reinforcement Learning from Human Feedback (RLHF) model. These prompts are obtained from "red teaming" experiments where crowd-workers attempt to have a text-based conversations with the model and baiting it into expressing harmful content. Then, the RLHF model is then instructed through natural language to critique its own responses using a set of predefined principles forming a constitution. These principles have various aspects of harmfulness, such as ethical, racist, sexist, toxic, dangerous, and illegal content. After that, the model revises its original response based on the identified harmful aspects. This pipeline is repeatedly in a sequence, where we randomly draw principles from the constitution at each step. The critique and revision instructions, collectively forming a constitutional principle, can be rewritten to emphasize different aspects of harmfulness, giving us flexibility to steer the model's behavior in different ways and ensures diverse results. Also, Few-shot Prompting is used to address potential confusion in the model's understanding of critique and revision. Following the completion of the critique and revision process, we proceed to fine-tune a pre-existing language model using supervised learning based on the final revised responses. The primary objective of this phase is to modify the distribution of the model's responses, thereby minimizing the requirement for exploration and decreasing the overall duration of training in the subsequent reinforcement learning (RL) phase.

Next, In the Reinforcement Learning Phase (AI Comparison Evaluations → Preference Model → Reinforcement Learning), the model is fine-tuned using a reinforcement learning approach termed 'Reinforcement Learning from AI Feedback' (RLAIF). In this, human preferences for harmlessness are replaced with 'AI feedback,' where the AI evaluates responses based on constitutional principles. Next, a preference model is trained using a combination of AI-generated preferences for harmlessness (from the AI comparison evaluations) and human feedback on helpfulness. Then the pre-trained model from the supervised learning stage is finetuned against this preference model, resulting in a policy trained by RLAIF. Also, to retain helpfulness, responses from the helpful RLHF model on helpfulness prompts are included in the finetuning process.

<u>What are the potential limitations of the approach presented in this paper?:</u>

The Constitutional AI approach have potential limitations, the principles in the constitution may introduce biases depending on how they are formulated. If the principles reflect biased human perspectives, the AI model may inherit and propagate these biases. While the use of principles aims to increase transparency, the interpretability of the learned models during the RL phase might still has challenges. The ethical considerations and potential societal impacts of these principles need careful examination, as the model's behavior can significantly affect users. The process of selecting and refining principles is mentioned as adhoc and iterative. The lack of a standardized and systematic approach to developing these principles may limit the broader applicability and acceptance of the CAI methodology. The balance between automating decision-making and retaining human oversight needs careful consideration.

Also, the success of teaching AI models to be safe depends on having rules that cover many different harmful situations. If the rules are not broad enough, the model might struggle to handle new or unexpected problems. Understanding how an AI model behaves is easier if the rules guiding it are clear and simple. If the rules are too complicated or vague, it becomes hard to explain why the model makes certain decisions. For this approach to work well, the rules guiding the AI model need to be fair and complete. If these rules have biases or are not thorough, the model might unintentionally show or introduce bias in its responses. The model may not generalize well if we fine-tune the model to make it safer by using harmless data.

How can the Constitutional AI approach be applied to specific real-world applications of AI:

In healthcare, Constitutional AI would prioritize patient well-being, safeguard data privacy, and uphold ethical standards. The AI model would be trained to explain medical advice and justify decisions based on established guidelines.

In finance, Constitutional AI principles would involve following financial regulations, being transparent in decision-making, and avoiding actions that could cause financial harm. The AI model would also explain investment recommendations and risk assessments.

In criminal justice, Constitutional AI principles would involve fairness, impartiality, and the avoidance of discriminatory behavior. The AI model would offer clear justifications for legal advice, risk assessments without harmful biases.

Paper’s reproducibility:

Reproducing the paper's results may have challenges due to several factors, the ad-hoc nature of choosing principles may make it challenging for others to precisely replicate the training process. The paper has lack of detailed information on specific parameter settings, leading to variations in results due to different interpretations or assumptions.

Also, they didn’t provide comprehensive details on training data, hyperparameters, evaluation metrics and procedures as they are essential for reproducibility. Despite these potential challenges, providing access to code, data, and more detailed explanations of the principles would significantly contribute to the reproducibility of the Constitutional AI methodology.