Lecture 8 Assignment

Part 1 - Azure Content Safety

In this assignment, you will explore Azure AI Services by completing **one** of the following two workshops. You will set up a Jupyter notebook, implement the code, execute it, and submit a PDF showing your running notebook with visible outputs.

Workshop Options

Choose **one** of the following:

1. **Protected Material Code Workshop**
   Link: `https://learn.microsoft.com/en-us/azure/ai-services/content-safety/quickstart-protected-mate`
   Objective: Detect and manage protected material in code or text.
   Language: Python

2. **Analyze Image Content Workshop**
   Link: `https://learn.microsoft.com/en-us/azure/ai-services/content-safety/quickstart-image?tabs=visual-studio%2Cwindows&pivots=programming-language-python`
   Objective: Analyze image content for harmful material.
   Language: Python

Assignment Requirements

1. **Setup Environment:** Follow prerequisites (e.g., install libraries, set up Azure).
2. **Implement the Code:** Copy and run the sample code in a Jupyter notebook.
3. **Run the Notebook:** Execute all cells, showing outputs.
4. **Submission:** Export as PDF with code, outputs, name, and student ID.

Steps to Complete

1. Choose a workshop.
2. Read the quickstart guide.
3. Set up environment (Python, Azure SDK, Jupyter).
4. Create and run a Jupyter notebook with the workshop code.
5. Test with at least one example.
6. Export as PDF and submit via [insert platform].

Notes

- Use `https://azure.microsoft.com/en-us/free/` for Azure access if needed. There is free access for students - reach out to RA or TA for this.
- Note any issues (e.g., API key problems) in the notebook.
- Collaboration allowed for troubleshooting; submit individual work.

Learning Outcomes

- Experience with Azure AI Services. Apply AI to content moderation. Practice Python workflows in Jupyter.

Part 2 - Reading Assignment for AI Safety, Ethics, Benchmarks, and Guardrails Discussion in the Next Class

LLM Papers to Review

This assignment is designed to encourage deep engagement with foundational AI safety concepts by comparing detailed methodologies, benchmarks, and ethical practices described in specific, state-of-the-art AI reports (Llama 3, DeepSeek-R1, GPT-4, and GPT-4.5).

Here I aim aims to foster critical thinking, enable students to analyze and contrast different models' approaches to ethical and safety challenges, and prepare them for an informed and robust class discussion focused on real-world AI implications.

To prepare effectively for our upcoming class discussion, carefully read and analyze the following reports. As you read, pay close attention to the specified sections and concepts, as they will directly inform our class discussion and related questions.

Be prepared to share sections from the reports to back up your answers.

1. Llama 3 Report

Focus areas:

- Introduction (Purpose of Llama 3 and Foundation Models)
- Section 5.4 (Safety and the function of Llama Guard 3)
- Section 4 (Post-training approaches, specifically supervised fine-tuning and direct preference optimization)

2. DeepSeek-R1 Report

Focus areas:

- Section 2 (Approach - particularly Reinforcement Learning and Reward Modeling)
- Section 4.2 (Challenges encountered, specifically *reward hacking*)
- Section 3 (Experimental results and evaluation methods)

3. GPT-4 Technical Report

Focus areas:

- Section 3 (Predictable Scaling and performance predictions)
- Section 4 (Capabilities, benchmarks, and specific examples of predictions)
- Section 5 (Limitations, hallucinations, and reliability)

4. GPT-4.5 System Card

Focus areas:

- Section 3.1.2 (Jailbreak Evaluations and robustness)
- Section 3.2 (Red Teaming evaluations, datasets, and comparative results)
- Section 4 (Preparedness Framework Evaluations and overall safety findings)

5. Cross-model Ethical Considerations

Comparative analysis:

- Strategies for handling hallucinations and biases
- Ethical methodologies used for reducing biases in outputs
- Unique safety mitigation techniques used by each model

Sample Classroom Discussion Questions

1. According to the Llama 3 report, what is the purpose and function of the Llama Guard 3 model, and how does it differ specifically from traditional reinforcement learning approaches like PPO in ensuring input and output safety?
2. The DeepSeek-R1 paper mentions a significant challenge related to "reward hacking" encountered during the large-scale reinforcement learning process. How exactly did the authors address this issue, and why did they specifically choose not to use neural reward models?
3. The GPT-4 Technical Report describes a key concept called "Predictable Scaling." Describe in detail how OpenAI was able to accurately predict GPT-4's performance using smaller-scale training runs, and provide at least one specific example of a benchmark or capability they accurately predicted this way.
4. The GPT-4.5 System Card discusses "jailbreak evaluations" and introduces two distinct datasets for assessing model robustness. What are these two datasets, how do they differ in their approach, and what were the specific results or findings when comparing GPT-4.5's robustness against other models like GPT-4o?
5. Both GPT-4 and DeepSeek-R1 documents address the ethical implications of hallucinations and biases in generated content. Compare and contrast the strategies each team used to reduce hallucinations and bias, highlighting specific methodologies unique to each model.

Instructions

- Take notes on the key methodologies and approaches outlined in each report.
- Be prepared to answer and discuss detailed, comparative, and conceptual questions that specifically reference these readings.
- Understand the reasoning behind each model's safety approaches, benchmarking practices, and ethical guardrails, as this will help you during the class discussion and QA session.