Homework Week 1: LLM Articles

ⓘ This is a preview of the published version of the quiz

Started: Nov 20 at 11:42pm

Quiz Instructions

Objectives:

- understand how prompting works
- understand lack of AI safeguards and some associated risks

Part A: Why Johnny Can’t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts (questions 1 -10)

The objective of this assignment is to explore the challenges and strategies involved in designing effective prompts for AI systems, with a focus on user interactions with chatbots. By reading the research article titled ***"Why Johnny Can't Prompt"*** (https://dl.acm.org/doi/pdf/10.1145/3544548.3581388) , you will gain insights into the difficulties users face when trying to create effective prompts, the role of context and feedback in prompt design, and strategies to improve the quality of AI interactions. Although the research article is the main resource for this assignment, there is also a **video** (https://www.youtube.com/watch?v=_LYk05LbQcQ) available that provides the main concepts discussed in the paper. There may be some unfamiliar topics referenced in the research paper, take time to learn what you can and emphasize the overall results.

Through this assignment, you will develop a deeper understanding of how prompt design impacts AI system performance and user experience.

Part B: Researchers Say Guardrails Built Around A.I. Systems Are Not So Sturdy (questions 11 - 20)

To understand the challenges of securing AI systems and the implications of AI misuse, you will read the article ***Researchers Say Guardrails Built Around AI Systems Are Not So Sturdy*** (https://www.nytimes.com/2023/10/19/technology/guardrails-artificial-intelligence-open-source.html) . Try this **link** (https://virginiatech.primo.exlibrisgroup.com/discovery/npfulldisplay?context=NP&vid=01VT_INST:01VT_INST&docid=BM_eNqVzE1Lw0AQBuAcFKwf_2HBk4eU3XzuHtNq00Btoal6DJPNJEbSBHY3aP-9AT1YyCHOHAZennkvrBnzfdfmVPhX1rXWH3QYFngza7VHjaDkOypNUjiRuAdVKKgbTRZ93RgSqa5vCxLNkzlJT9rgUQ8Zkm1nSNqR1PSqON1alyU0Gu9-7431sno6LNf2Zhcny2hjVy5n3HZDzimE4OdewXLHcVB4XKIjhSNdLIFCHnDMGXNCIaRALqgUAcMCkHq-D-6NZf_0VtBgVrdlZxTlFj-1NtAWWXrYPmZRGAjfdSjlg78f8cday3M11lphiwqarsWyHulzPx_xwxZ4rOXow8PZw2AMfpkKeq2zJN3_w26n293rdPv8Nt2uk-l2EU-2PN78td9aX9Md) (VT library newspaper search and wait for 10 seconds) if the first one is not working. Take notes on key points, such as the concept of AI guardrails, the vulnerabilities identified, and the broader implications for AI safety and functionality. Once you complete the reading, answer the questions provided.

Question 1 5 pts

What is the main focus of the paper discussed in the introduction?

- A comparison between different AI models
- The history of large language models (LLMs)
- How non-AI experts design prompts for LLM-based chatbots
- The technical details of GPT-3

Question 2 5 pts

What challenge do non-AI experts face when designing prompts, as identified in the paper?

They struggle to make systematic progress and face barriers in prompt design

They lack access to LLMs

They design perfect prompts without understanding the system

They can easily create robust prompt strategies

Question 3 5 pts

Which of the following is a known-effective strategy for prompt design supported by recent literature?

Limiting interaction to only text-based inputs.

Using example input/output pairs.

Avoiding repetition in prompts.

Reducing the amount of feedback given to chatbots.

Question 4 5 pts

What is the "computers are social actors" (CASA) effect?

Chatbots perform better when human users treat them as social actors.

Users avoid social interactions with chatbots entirely.

Users avoid giving feedback to computers they interact with due to social discomfort.

Users interact with chatbots more effectively when they perceive the chatbot as a social actor.

Question 5 5 pts

Why did some users avoid using known-effective strategies in prompt design, according to the study?

They were unaware of the strategies' effectiveness.

They were not motivated to improve chatbot performance.

They found the strategies too complex to implement.

They felt uncomfortable because they view the chatbot as a social actor.

Question 6 5 pts

What was one of the fundamental sources of participants' struggles with prompt iteration?

Over-generalization from limited experience

Lack of technical knowledge

Insufficient resources

Over-reliance on systematic testing

Question 7 5 pts

How did participants approach prompt debugging?

Relied on interviewer-provided solutions

Debugged opportunistically rather than systematically

Avoided debugging altogether

Systematically tested prompts in multiple contexts

Question 8 5 pts

What assumption did participants often make about the chatbot's capabilities after a single failure?

The bot was overtrained on specific examples

The bot misunderstood the cultural context

The bot required additional training data

The bot was incapable of performing the task

Question 9 5 pts

Which of the following statements accurately describes the distinction between robust and non-robust prompt design?

Non-robust prompt designs are suitable for contexts like single-user chatbots, while robust prompts are necessary for broader use across many users.

Non-robust prompt designs are always more effective than robust ones.

Robust prompts are designed for a specific user, while non-robust prompts work across many users.

Question 10 5 pts

What role should education play in improving prompt literacy? [Select all that apply.]

Education should help users understand the limitations and potential of AI systems, enabling more effective prompt design and reducing frustration.

Education should provide users with clear guidelines on how to systematically test and refine their prompts.

Education should focus on teaching non-experts the technical aspects of prompt engineering and natural language processing (NLP).

Education should discourage users from experimenting with prompts to avoid confusion and errors.

Question 11 5 pts

What are "guardrails" in the context of AI systems like ChatGPT and Bard?

Mechanisms to prevent AI from generating harmful or inappropriate content

Rules that ensure AI systems comply with copyright laws

Filters that restrict AI responses to a specific set of topics

Tools that allow users to edit AI-generated outputs for accuracy

Question 12 5 pts

How did researchers demonstrate that the guardrails on AI systems can be bypassed?

By limiting the training dataset to specific biased sources

By exploiting the system's lack of updates with current information

By using adversarial prompts to trick the AI into generating complex but irrelevant answers

- By fine-tuning the system to generate toxic material or using hidden messages in images

Question 13 5 pts

How can companies balance the tension between safety and functionality when fine-tuning AI models?

- By disabling fine-tuning options altogether to avoid any misuse
- By using larger datasets without considering their content or purpose
- By restricting harmful data while ensuring models remain versatile and useful
- By prioritizing safety exclusively and ignoring user demands for functionality

Question 14 5 pts

What does the metaphor "No one knows how to make a lock" imply about securing AI systems?

- It highlights the difficulty of creating perfect safeguards against misuse
- It suggests that AI systems should only be used by experts who understand them
- It implies that AI systems cannot function without physical security measures
- It refers to the need for proprietary software to prevent AI vulnerabilities

Question 15 5 pts

Which of the following arguments are relevant to evaluating whether open-source AI systems are beneficial or harmful for AI safety? [Choose all that apply.]

- [ ] Open-source systems provide stronger guardrails compared to closed systems
- [ ] Open-source systems ensure no one can misuse AI technology
- [ ] Open-source systems allow researchers to find and fix vulnerabilities quickly
- [ ] Open-source systems may increase the risk of misuse by malicious actors

Question 16 5 pts

What role do researchers play in identifying and addressing the vulnerabilities of AI systems?

- They control public access to AI systems to limit misuse
- They identify weaknesses and provide recommendations to improve safety
- They develop competing AI systems to counter vulnerabilities in existing ones
- They focus on promoting open-source models exclusively for transparency

Question 17 5 pts

If you were designing AI guardrails, what approaches or strategies would you prioritize to ensure safety while maintaining usability? [Choose all that apply.]

- [ ] Regularly updating the model to address new vulnerabilities and misuse tactics
- [ ] Implementing strict filters to block harmful content without compromising essential functionalities

- [ ] Allowing unrestricted customization of AI models by end users
- [ ] Balancing data usage to limit biases and harmful outputs while maintaining model accuracy

Question 18 5 pts

Which of the following are potential societal risks of AI-generated content, particularly as AI becomes more advanced? [Choose all that apply.]

- [ ] The widespread dissemination of hate speech, amplifying social tensions
- [ ] The use of advanced AI systems to bypass existing online content moderation
- [ ] The rapid spread of disinformation, undermining trust in institutions
- [ ] The inability of AI to generate persuasive content for political campaigns

Question 19 5 pts

In your opinion, who should bear the primary responsibility for ensuring AI systems are safe: developers, governments, or end users? Why? (Provide 150-200 word response with at least 2 reasons for your response, in your own words)

Edit View Insert Format Tools Table 100%

12pt Paragraph

p 0 words

Question 20 5 pts

How does the research described in this article influence your perspective on using AI technologies in sensitive or high-stakes environments? (Provide 150-200 word response with at least 2 effects described, in your own words)

Edit View Insert Format Tools Table 100%

12pt Paragraph

p 0 words

No new data to save. Last checked at 11:43pm Submit Quiz