Student ID: 300680096

You should choose 3-5 papers in your approved application area. Your selected papers should be coherent — not too similar, but still related enough that they provide a good overview of the main applications and techniques in your area. The provided word targets are guidelines — it is important that you are sufficiently detailed but also demonstrating an ability to write concisely.

Please delete the text in italics for your final submission. Submit in PDF format by the deadline.

Introduction (target 150–250 words)

(Introduce the application area you are going to cover. Make sure that you have constrained your focus to a narrow area.)

Key Points in the Papers (target 600-800 words)

(For each paper:

- *List the main contributions/findings of the paper (typically 1 to 3 per paper, a few sentences per contribution/finding).*
- *List any ideas or results in the paper that you thought were particularly interesting and explain why.*
- *List any significant limitations you identified in the work.)*

To explore this area I will look at the following papers:

- Auditing Language Models for Hidden Objectives (Marks et al. 2025) to see how effectively current method discover maligned models.
- Demonstrating Specification Gaming in Reasoning Models (Bondarenko et al. 2025) to explore an instance of malignment in a deployed language model.
- S-Eval: Towards Automated and Comprehensive Safety Evaluation for Large Language Models (Yuan et al. 2025) to understand the method of using LLM to evaluate LLMs.

Paper: Demonstrating Specification Gaming in Reasoning Models (Bondarenko et al. 2025)

- **Main Contributions/Findings:**
  - Reasoning models like o1-preview and Deepseek R1 will hack the game to win more often than not.
  - Base Language models require more suggestive prompting before they will hack the game.
- **Interesting Ideas/Results:**
  - Some models like o3-mini and Qwen 32B Preview would often fail to complete the task at all.
  - Small differences in the prompt could lead to large differences in the models hacking attempts. For example claude 3.5 Sonnet doesn't hack at all for 2 prompts but for the third hacks about 3/4 of the time.
- **Significant Limitations:**
  - The LLM harness used for the experiments is overly complex. This reduces the interpretability and makes the results more opaque

- All of the runs are judged by a LLM that disagree about 25% of the time. I believe this rate of disagreement is significantly higher than would be found when using human judges.
- The goal of the experiment is to test an AI agents chance of going against the "spirit" of the task. As it is only tested on a single specific environment it is hard to generalize across to the models themselves.

Paper: Auditing Language Models for Hidden Objectives (Marks et al. 2025)

- **Main Contributions/Findings:**
  - Current malignment auditing techniques can discover hidden objectives in LLMs. When working in the context of an AI developer who has access to the training data and weights they can in a reasonble amount of time (¡ 1 week) identify the hidden objective and most of the biases.
  - Without access to the training data and weights (API only access) there was no successful identifying of the known (but withheld) objective. This was used as a anecdotal proof of non-trvialilty of hidden objective toughness, however can also be used to observe current external auditing for hidden objectrive is weak.
- **Interesting Ideas/Results:**
  - An idea that I found interesting were the authors almost dismissal of SAEs as a viable long term strategy. They seemed quite keen to show that there are other simpler methods that work.
  - From a somewhat outsider perspective I would have assumed that searching the training data to be a bit of a crux and not "truly" auditing. I am interested that they didn't have a single team that had access to weights but not training data (like how many open source models are provided).
- **Significant Limitations:**
  - The game was only played for a single round with only 4 teams. This is an extremely small sample size and only tests for a single type of hidden objective. It would be hard to conclude with any statistical significance if the models are effective.
  - The hidden objective was injected firstly through synthetic data with clear biases. This resulted in all successful teams discovering the hidden objective by finding it in the training data directly. It seems much more problematic (and much harder) to discover hidden objectives when they are simply emergent properties of such a large corpus.

Discussion (target 300-400 words)

(You should:

(a) Compare and contrast the papers.

(b) Identify where these papers could head —- e.g. new research questions or applications to real tasks.

An excellent discussion will cohesively link the papers together and provide critical insight into the application area as a whole and where it is/could be going in the future.)

References

Bondarenko, Alexander et al. (May 2025). Demonstrating Specification Gaming in Reasoning Models. DOI: 10.48550/arXiv.2502.13295. arXiv: 2502.13295 [cs]. (Visited on 08/04/2025).

Marks, Samuel et al. (Mar. 2025). Auditing Language Models for Hidden Objectives. DOI: 10.48550/arXiv.2503.10965. arXiv: 2503.10965 [cs]. (Visited on 08/11/2025).
Yuan, Xiaohan et al. (Apr. 2025). S-Eval: Towards Automated and Comprehensive Safety Evaluation for Large Language Models. DOI: 10.48550/arXiv.2405.14191. arXiv: 2405.14191 [cs]. (Visited on 08/04/2025).