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

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).