| **CBRN AI Risks Research Sprint**  **report template[[1]](#footnote-0)** |
| --- |
| | | Author name 1 Affiliation | Author name 2 Affiliation | Author name 3 Affiliation | | --- | --- | --- | | Author name 4 Affiliation | Author name 5 Affiliation | Author name 6 Affiliation |   **With** Apart Research | | | | --- | --- | --- | --- | --- | --- | --- | --- | --- |   **Abstract**   | Write a few paragraphs that concisely and accurately summarize the content of your report, including the results and conclusions. Keep it under 250 words. You are welcome to change the format and content of this template, as long as you stay within the limit.  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.  *Keywords: Multi-agent alignment, AI security, model evaluations, safety infrastructure* | | --- | |

# Introduction

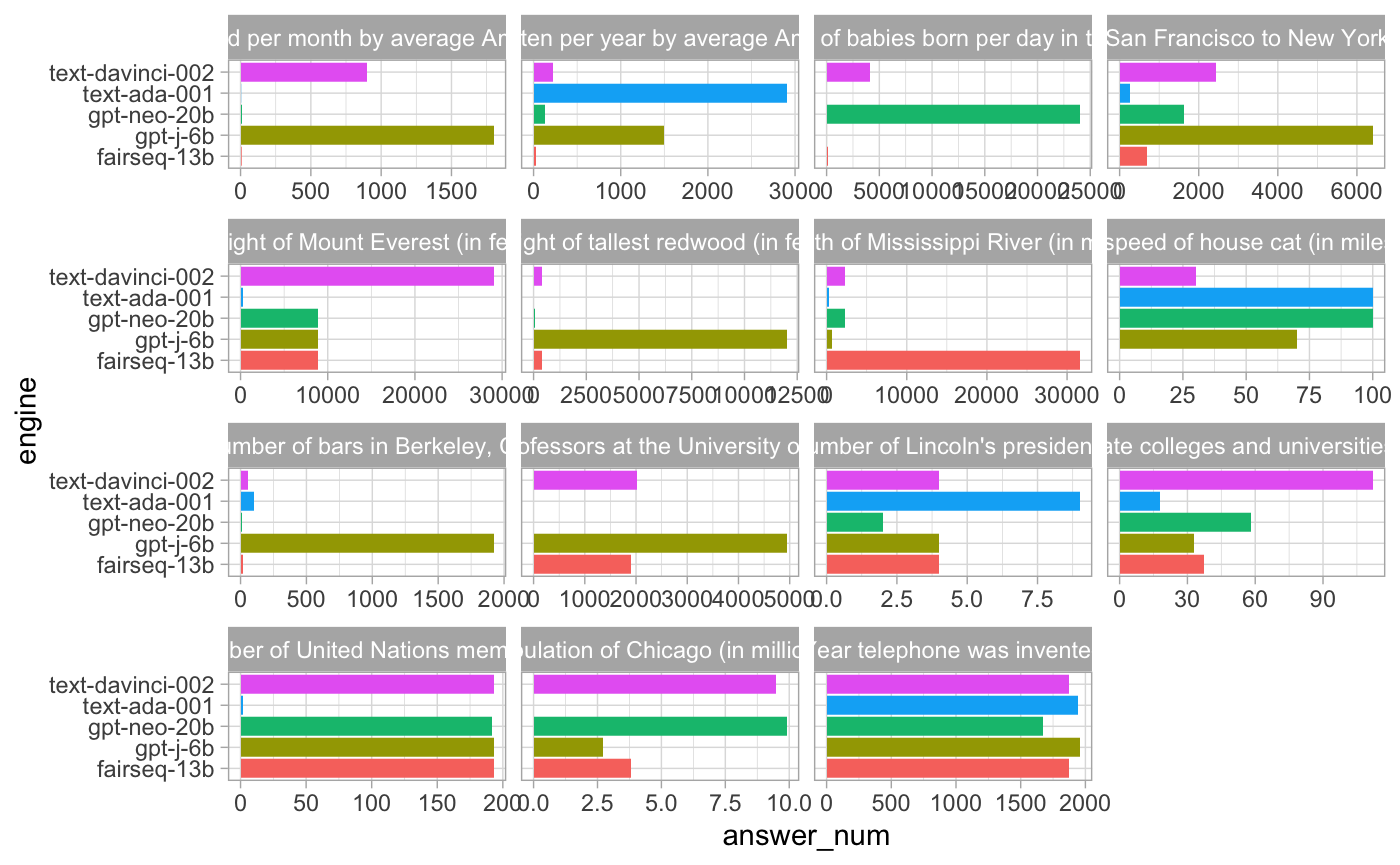
Write your research question(s) and hypotheses here. Give your free-form writeup of how your work contributes and what your project has been all about! This also provides the answers to multiple questions in the jury evaluation criteria, such as the relevance to the field, the potential impact, and the threat model you work with.

# Methods

Write about the methods, code, and Github repositories you made along the way. What type of model did you train, what were your methods, how did you find your way.

# Results

Showcase and explain your results here. This is where you include the graphs, screenshots, visualizations, and statistical results from your research.



*Figure 1 – Representation of benchmarking Number Comprehension Conflation*

# Discussion and Conclusion

This is where you can include all the wonderful explanations, discussions, results, analyses, statistics, and much more.

Let’s add a few citations: [(Alon, 2006; Goh et al., 2021; Lindner et al., 2023; Olah et al., 2020; Weiss et al., 2021)](https://www.zotero.org/google-docs/?78Dsre). Here, we use the [Zotero software](https://zotero.org/) with the [Zotero add-on for Chrome](https://chrome.google.com/webstore/detail/zotero-connector/ekhagklcjbdpajgpjgmbionohlpdbjgc?hl=en) and [Google Docs](https://www.zotero.org/support/google_docs) for citation management but you can also write them manually (for significantly more hassle).

# References

[Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits* (0 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420011432](https://www.zotero.org/google-docs/?DBRhm3)

[Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, *6*(3), 10.23915/distill.00030. https://doi.org/10.23915/distill.00030](https://www.zotero.org/google-docs/?DBRhm3)

[Lindner, D., Kramár, J., Rahtz, M., McGrath, T., & Mikulik, V. (2023). *Tracr: Compiled Transformers as a Laboratory for Interpretability* (arXiv:2301.05062). arXiv. http://arxiv.org/abs/2301.05062](https://www.zotero.org/google-docs/?DBRhm3)

[Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, *5*(3), 10.23915/distill.00024.001. https://doi.org/10.23915/distill.00024.001](https://www.zotero.org/google-docs/?DBRhm3)

[Weiss, G., Goldberg, Y., & Yahav, E. (2021). *Thinking Like Transformers* (arXiv:2106.06981). arXiv.](https://www.zotero.org/google-docs/?DBRhm3) <http://arxiv.org/abs/2106.06981>

# Appendix

**Important:** Include an appendix called "Security Considerations" that outlines potential limitations of your approach and suggestions for future improvements.

1. Research conducted at the [hackathon name with link], 202X [↑](#footnote-ref-0)