Iterative summarization towards natural abstraction understanding[1]

Yoann Poupart
ENS Lyon

Abstract

My abstract.

Keywords: Mechanistic interpretability, ML safety, Summarisation, Exemplification, Natural abstraction

1. Introduction

Natural Abstraction hypotheses.

2. Methods

All the work produced can be found on the associated GitHub repository[2].

3. Results

The results.

Figure 1 – Representation of Benchmarking Number Comprehension Conflation

4. Discussion and Conclusion

This work could be extended with natively iterative models like those proposed in **[1]** for document summarization. It has been developed to improve abstraction in **[2]**.

[1] Research conducted at the Apart Research Alignment Jam #4 (Mechanistic Interpretability), 2023 (see https://itch.io/jam/mechint)
[2] GitHub repository link: https://github.com/Xmaster6y/Iterative_summarisation.

5. References

[1] Chen, X., Gao, S., Tao, C., Song, Y., Zhao, D., & Yan, R. (2018). Iterative Document Representation Learning Towards Summarization with Polishing. *arXiv*. https://doi.org/10.48550/arXiv.1809.10324

[2] J. Li, C. Zhang, X. Chen, Y. Cao and R. Jia, "Improving Abstractive Summarization with Iterative Representation," *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206950.