The Journal of Open Source Software

CCS-Lib: A Python package to elicit latent knowledge from LLMs

Nora Belrose[1], Walter Laurito[2,3¶], Alex Mallen[1,7], Fabien Roger[4], Kay Kozaronek[2], Christy Koh[5], Jonathan NG[2], James Chua[1], Alexander Wan[5], Reagan Lee[5], Ben W.[1], Kyle O'Brien[1,6], Augustas Macijauskas[8], Waree Sethapun[9], and Eric Mungai Kinuthia[1]

1 EleutherAI **2** NotodAI Research **3** FZI Research Center for Information Technology **4** Redwood Research **5** UC Berkeley **6** Microsoft **7** University of Washington **8** CAML Lab, University of Cambridge **9** Princeton University ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Beatriz Costa Gomes

Reviewers:

- @praneethd7
- @isdanni

Submitted: 09 December 2023
Published: unpublished

Summary

ccs is a library designed to elicit latent knowledge (elk (Christiano et al., December 2021)) from language models. It includes implementations of both the original and an enhanced version of the CSS method, as well as an approach based on the CRC method (Burns et al., 2022). Designed for researchers, ccs offers features such as multi-GPU support, integration with Huggingface, and continuous improvement by a dedicated group of people. The Eleuther AI Discord's `elk` channel provides a platform for collaboration and discussion related to the library and associated research.

Statement of need

Language models are proficient at predicting successive tokens in a sequence of text. However, they often inadvertently mirror human errors and misconceptions, even when equipped with the capability to "know better." This behavior becomes particularly concerning when models are trained to generate text that is highly rated by human evaluators, leading to the potential output of erroneous statements that may go undetected. Our solution is to directly elicit latent knowledge ((elk (Christiano et al., December 2021)) from within the activations of a language model to mitigate this challenge.

ccs is a specialized library developed to provide both the original and an enhanced version of the CSS methodology. Described in the paper "Discovering Latent Knowledge in Language Models Without Supervision" by Burns et al. (2022). In addition, we have implemented an approach, called VINC, based on the Contrastive Representation Clustering (CRC) method from the same paper.

ccs serves as a tool for those seeking to investigate the veracity of model output and explore the underlying beliefs embedded within the model. The library offers:

- Multi-GPU Support: Efficient extraction, training, and evaluation through parallel processing.
- Integration with Huggingface: Easy utilization of models and datasets from a popular source.
- Active Development and Support: Continuous improvement by a dedicated team of researchers and engineers.

For collaboration, discussion, and support, the Eleuther AI Discord's elk channel provides a platform for engaging with others interested in the library or related research projects.

The Journal of Open Source Software

Acknowledgements

We would like to thank EleutherAI, SERI MATS for supporting our work and Long-Term Future Fund (LTFF)

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2022). Discovering latent knowledge in language models without supervision. *arXiv Preprint arXiv:2212.03827*.

Christiano, P., Cotra, A., & Xu, M. (December 2021). *Eliciting latent knowledge (ELK)*. https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/.