




# 1 elk: A Python package to elicit latent knowledge from 2 LLMs

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

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

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 09 December 2023

Published: unpublished

## License

Authors of papers retain copyright  
and release the work under a  
Creative Commons Attribution 4.0  
International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## 10 Summary

11 elk is a library designed to elicit latent knowledge ([elk](#) ([author:elk?](#))) from language models.  
12 It includes implementations of both the original and an enhanced version of the CSS method,  
13 as well as an approach based on the CRC method ([author:burns?](#)). Designed for researchers,  
14 elk offers features such as multi-GPU support, integration with Huggingface, and continuous  
15 improvement by a dedicated group of people. The Eleuther AI Discord's elk channel provides  
16 a platform for collaboration and discussion related to the library and associated research.

## 17 Statement of need

18 Language models are proficient at predicting successive tokens in a sequence of text. However,  
19 they often inadvertently mirror human errors and misconceptions, even when equipped with  
20 the capability to “know better.” This behavior becomes particularly concerning when models  
21 are trained to generate text that is highly rated by human evaluators, leading to the potential  
22 output of erroneous statements that may go undetected. Our solution is to directly elicit latent  
23 knowledge (([elk](#) ([author:elk?](#))) from within the activations of a language model to mitigate  
24 this challenge.

25 elk is a specialized library developed to provide both the original and an enhanced version of the  
26 CSS methodology. Described in the paper “Discovering Latent Knowledge in Language Models  
27 Without Supervision” by Burns et al. ([author:burns?](#)). In addition, we have implemented an  
28 approach, called VINC, based on the Contrastive Representation Clustering (CRC) method  
29 from the same paper.

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

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

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

40 **Acknowledgements**

41 We would like to thank [EleutherAI](#), [SERI MATS](#) for supporting our work and [Long-Term Future](#)  
42 [Fund \(LTFF\)](#)

DRAFT