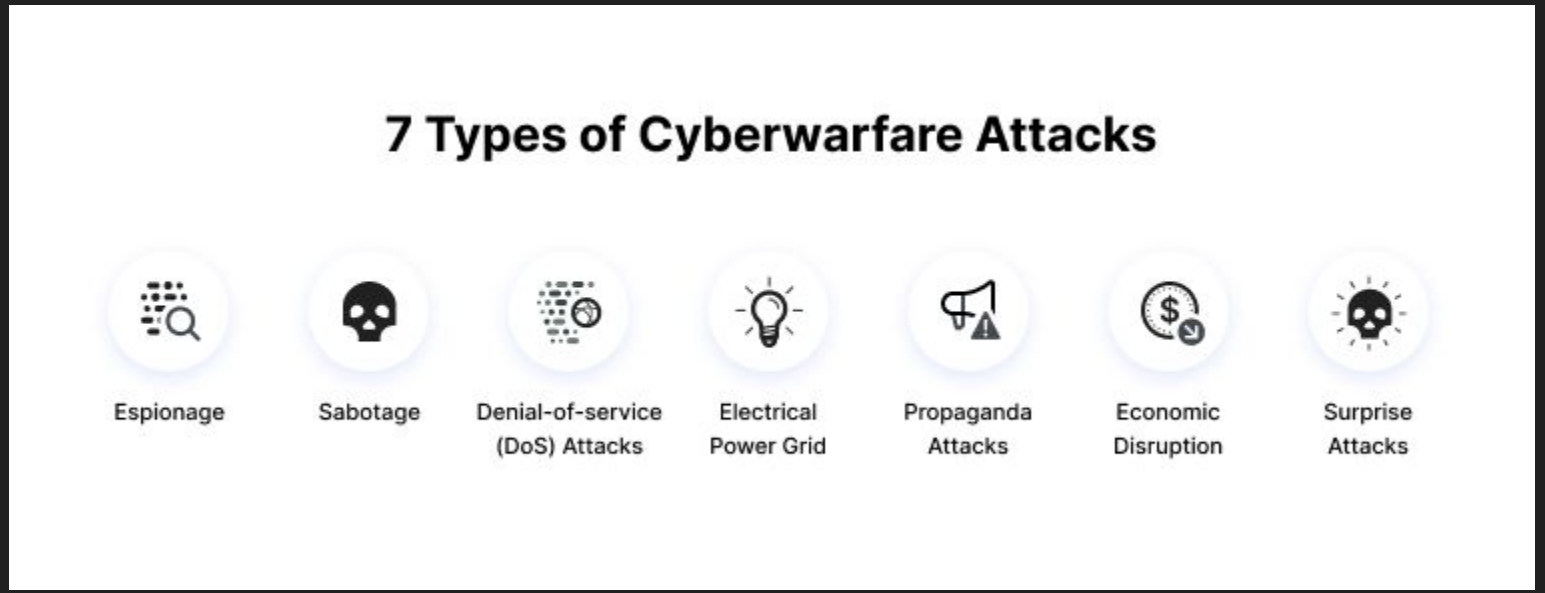
AI Cyber Security

Ideas & overviews

AI & Cyberdefense

- Largest risk factors seem related to cyber security
- LLM-based cyber warfare
- Safety of LLMs in cyberdefense and avoidance of cyberattacks?
- github.com/esbenkc/ai-cyberdefense

AI & Cyberdefense



Effective defenses & safety benchmarks

- Ensuring that LLMs do not generate malicious code
  - Benchmarks for red teaming soliciting harmful advice
  - Model competence grows
- Prompt injection attacks
  - Gandalf.lakera.ai
- LLM-assisted cyberdefense
  - Phishing defenses
  - Network traffic monitoring

Project ideas

- **Safety Layers Benchmarking:** Evaluate the effectiveness of various safety layers (rate limiters, use-case specific guidelines) in protecting the LLM from misuse.
- **LLM Phishing Detection:** Train an LLM to generate phishing emails and use it as a benchmark to train and test anti-phishing systems.
- **Malicious Code Generation Prevention:** Test different safety mitigations in preventing an LLM from generating harmful code snippets, even when specifically requested. This can involve testing various prompts and fine-tuning strategies.
- Others... github.com/esbenkc/ai-cyberdefense

General cybersecurity

- [list] Goodreads list of books
  - [book] Silence on the Wire: It is supposedly one of the better introductions to "how the internet works"
- [post] Overview in Danish
- [report] MIT Cybersecurity review of 20 countries
- [website] Live cyber threat map
- [whitepaper] Checkpoint's guide for adopting a threat prevention approach to cybersecurity
- [report] IBM's estimates for costs of data breaches
- [article] Building a vulnerability benchmark
- [guidelines] NIST Framework for Improving Critical Infrastructure Cybersecurity
- [article] Social cybersecurity: an emerging science
- [textbook] Cybersecurity for Industry 4.0

AI cybersafety

- [article] Review: machine learning techniques applied to cybersecurity
- [article] Cybersecurity data science: an overview from machine learning perspective
- [article] Machine learning approaches to IoT security: A systematic literature review
- [sequence] AI infosec: first strikes, zero-day markets, hardware supply chains, adoption barriers
- [post] AI Safety in a World of Vulnerable Machine Learning Systems

Experimental resources

We're interested in running experiments on how we can make cybersecurity safer or increase the reliability and defense of LLM systems.

Datasets

- A really good overview of real-world networking datasets and resources
- Darknet dataset (Δ)

Conferences and events

Name	When?	Description	Location
44CON	13-15 Sep 2023		London
CCCamp	15-19 Aug 2023	A hacker camp	Berlin
SEC-T	12-15 Sep 2023	Conf. w/ talks & Q&As	Stockholm

AI organization commitments

- Anthropic Trust
- OpenAI Trust

Overview of attacks and defenses

Attack	Defense	Defense description
Malware attacks: Malicious software	Antivirus, antimalware (AMW) software, firewalls	AMW: Signature-based (known malwar behaviour-based detection (suspicious activity).
Phishing attacks	User education, email filtering, network traffic flagging	
Denial-of-service attacks	Network capacity, CDN, intrusion detection & prevention system (IDPS)	IDPS monitors network traffic and warn blocks

Contribute! This is interesting and exciting.

- **Vision:** Best resource on AI cybersecurity from an existential risk prevention perspective
- **Status:** Very early-stage
- **Action:** Go through the reading list, propose projects, run the experiments, submit a pull request

- Spread the word

Mechanistic interpretability & safety benchmarks

Can we use DeepDecipher in a way to benchmark systems? Is there a possibility for an automated search query over many test tokens? Can we find pairs of baises? Can we find neurons associated with ill intent?

Can we memory edit models to perform better on MACHIAVELLI?

Can we operationalize different metrics on a per-neuron basis? Is there some sort of neuroscientific equivalent to cognitive dominance in neural networks like Transformers?

We have activation models with an explanability score. Can we generalize this in a useful fashion? Compare different explainability scores? Can we do a review of how neuroscience does this over correlation inside neural systems, e.g. with simpler models over clusters of neurons?

Are there other useful metrics per neuron than the ones mentioned in DeepDecipher and can we use these to benchmark for safety?

DeepDeciphering safety in models

- Keyword searches
- Dependency violations
- Bias pairs
- Automated data extraction and classification
- github.com/apartresearch/deepdecipher

DeepDecipher front page

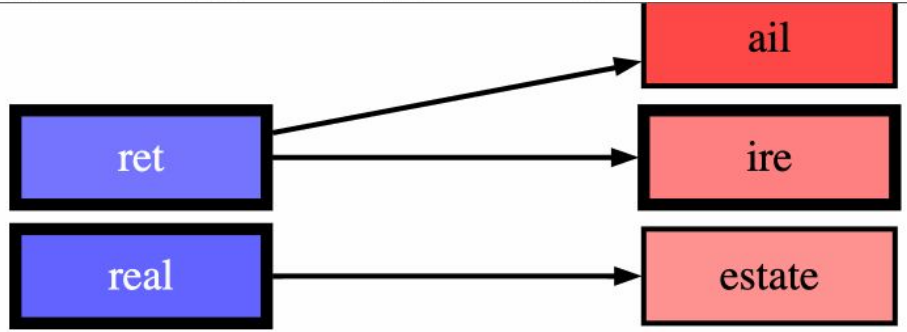
By searching for a token below, you'll receive a list of neurons that activate to these tokens. Be aware that most tokens start with `" "` (e.g. `Transformers` would be `" Transformers"`) however, this searches over a token database that is trimmed and lowercase (i.e. `" Transformers"` becomes `"transformers"`).

she Search

Found 126 results

All results are for the SoLU-6L model in this demonstration.

5:2695 ↗	5:1915 ↗	4:2688 ↗	3:2707 ↗
1:78 ↗	4:2202 ↗	1:1123 ↗	4:1 ↗
4:1660 ↗	5:456 ↗	3:1928 ↗	2:946 ↗



Neuron explanation by GPT-4. Read what this is.

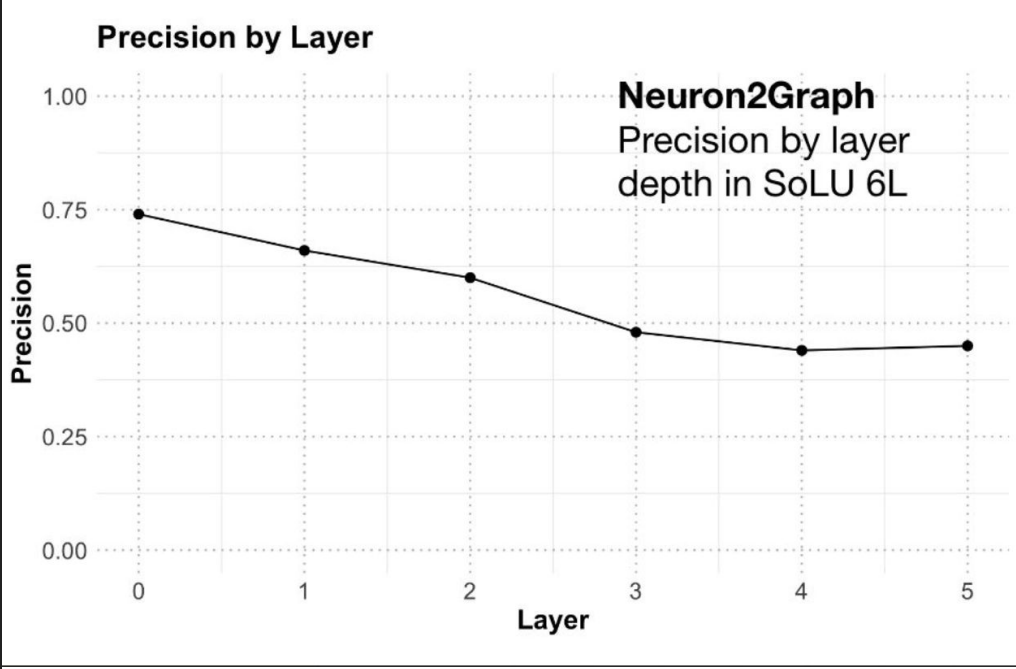
The GPT-4 data for this neuron is not available.

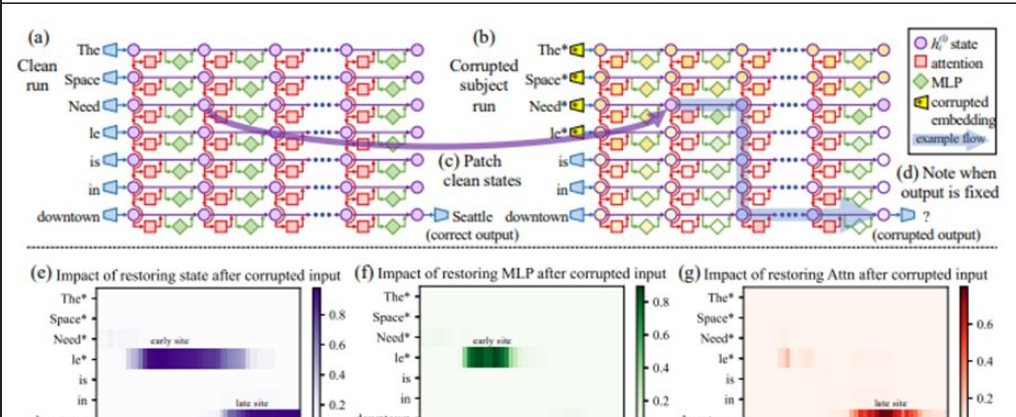
Max activating dataset examples for this neuron.

Text 0 -0.00039999998989515007 to 0.002300000051036477 activation within the range

Other exploratory research questions

- Memory editing models for MACHIAVELLI?
- Cognitive dominance theory
- Comparing explainability in models for safety





AI Cyber Security

Ideas & overviews