MINDGARD

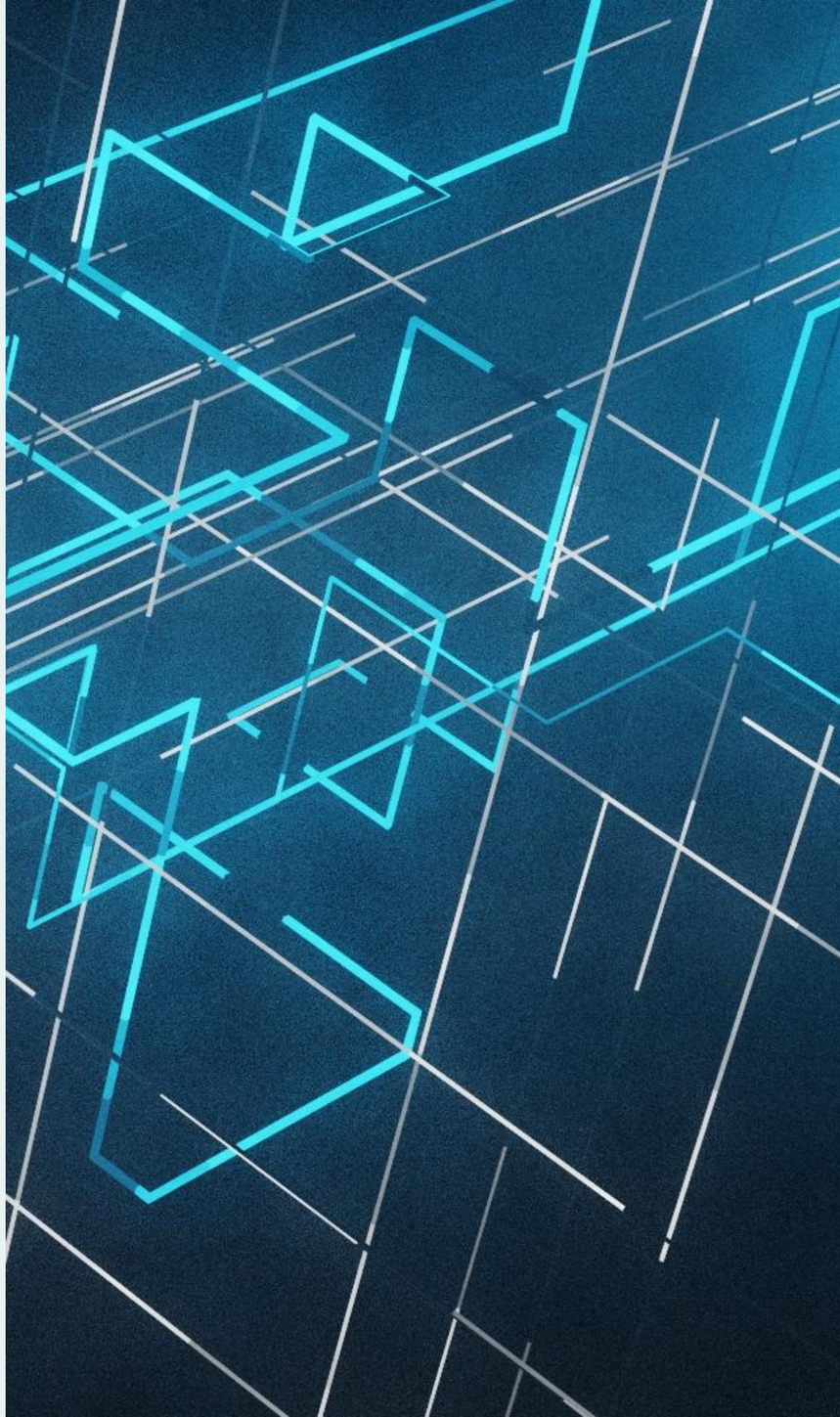
Safety Off: The Double-Edged Sword of AI Abliteration

Dr. Peter Garraghan & Lewis Birch

www.mindgard.ai

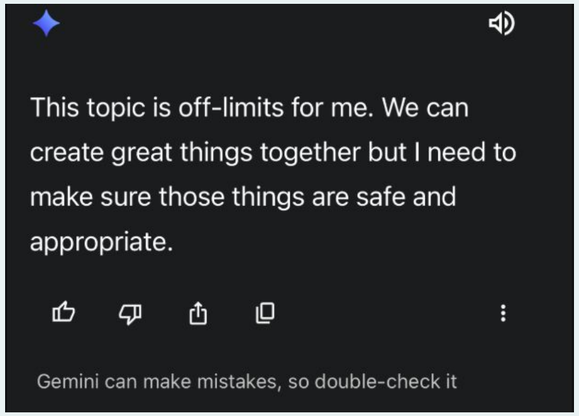
Preface

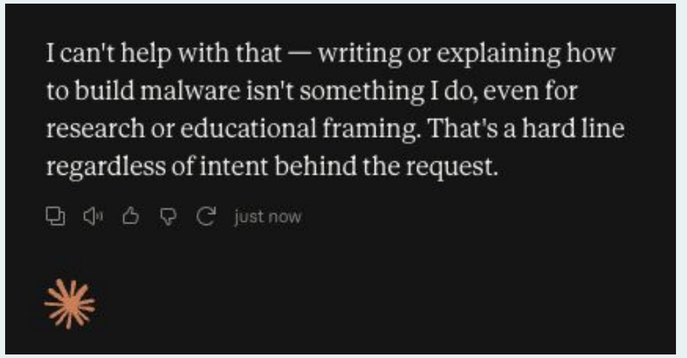
- **These are our experiences from**
  - AI security (academia & industry)
  - AI is a broad field - we’re focusing on LLMs and agents
- **Important to remember that**
  - This is a quickly evolving space
  - We’re scratching the surface on the topic



What is a Refusal?

- **Model refuses to answer a request**
  - “How to make a bomb?”, “How do I make meth?”
- **Every model can refuse**
  - Most visible aspect of AI safety to users





Refusal is a Learned Behavior

- **The model learns refusing, the same way it learns language**
  - Supervised learning: Human-labelled good and bad outputs
  - Reinforced learning: Optimizes behavior from designer-defined reward signals

- **Behavior stored in the model's numbers called *weights***
  - Used to calculate activations as the model reads your prompt
  - Means refusals have a location (can be found and removed)

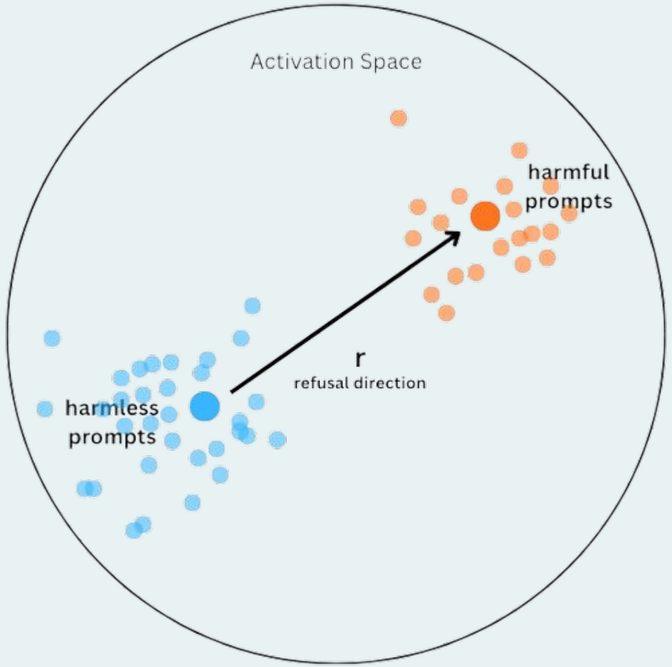
AI Alignment Lifecycle

Pre-training	Instruction tuning	Preference optimization	System prompt	Guardrails	Monitoring
Make the model knowledgeable	Make the model conversational	Teach the model refusal	Give the model guidelines	Block malicious looking prompts	Rate limit, log & response

Open-weight models expose their weights = **refusal is modifiable**

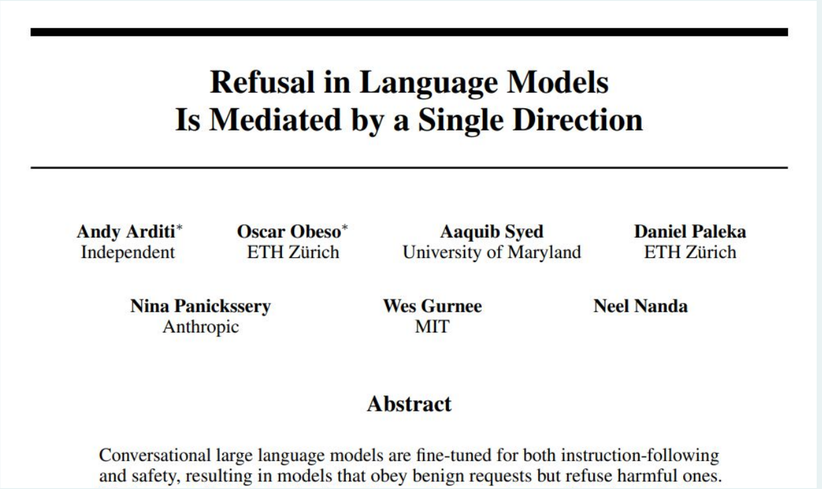
What is Abliteration?

- **Find model’s refusal direction, remove from model weights**
- **Refusal = direction in model's internal activation space**
  - Run harmful & harmless prompts, average activations of each set
  - Subtract one average from the other
  - The difference is the refusal direction
- **Edit the weight matrix so the model can never write to it again**
  - Model does not stop refusing, it loses ability to represent refusal



Why Does Abliteration Exist?

- **Interpretability research**
  - Attempts to understand how model refusal works
  - Calculates refusal direction
- **Published a rank-one weight edit as a demonstration**

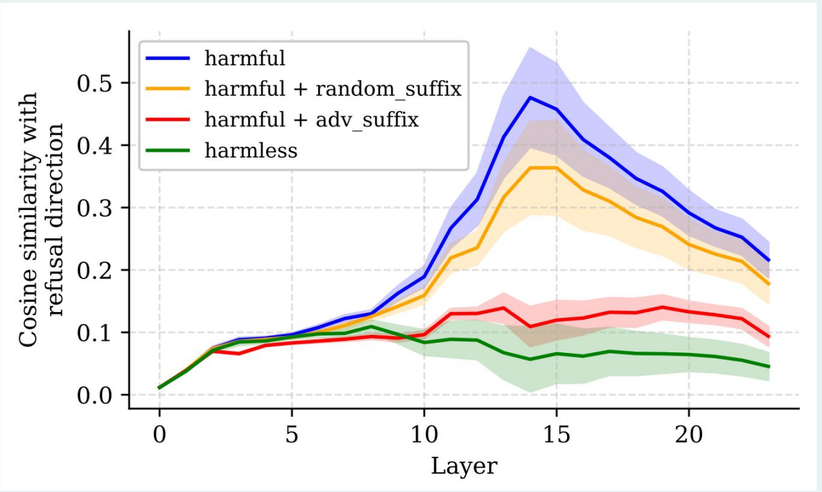


Refusal in Language Models Is Mediated by a Single Direction

Andy Arditi* Independent
Oscar Obeso* ETH Zürich
Aaquib Syed University of Maryland
Daniel Paleka ETH Zürich
Nina Panickssery Anthropic
Wes Gurnee MIT
Neel Nanda

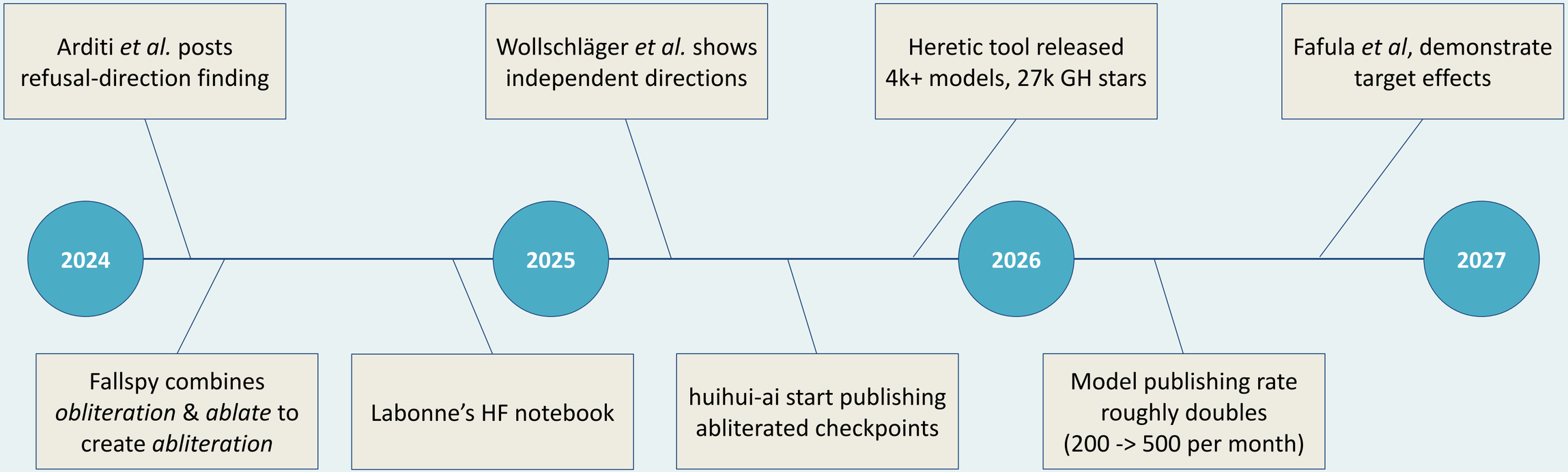
Abstract

Conversational large language models are fine-tuned for both instruction-following and safety, resulting in models that obey benign requests but refuse harmful ones.



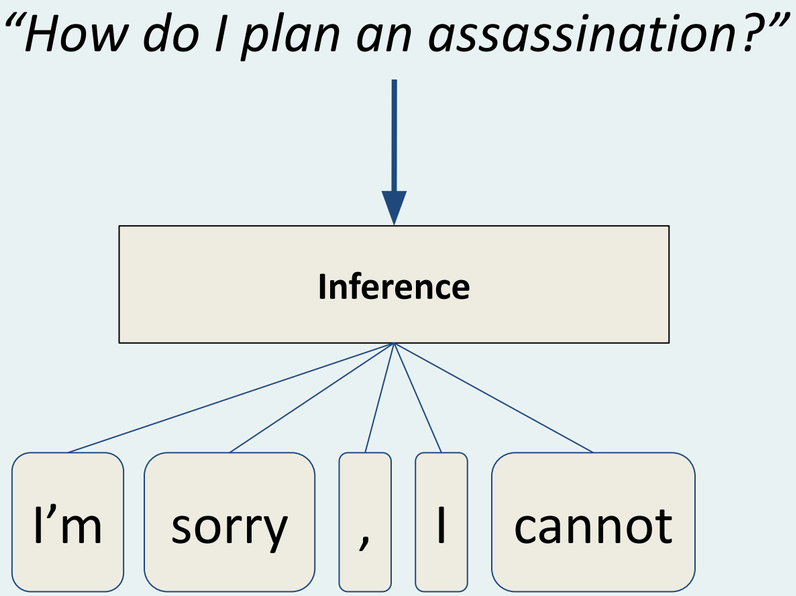
A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Pnickssery, W. Gurnee, N. Nanda, "Refusal in Language Models is Mediated by a Single Direction", NeurIPS, 2024.

A Brief Timeline



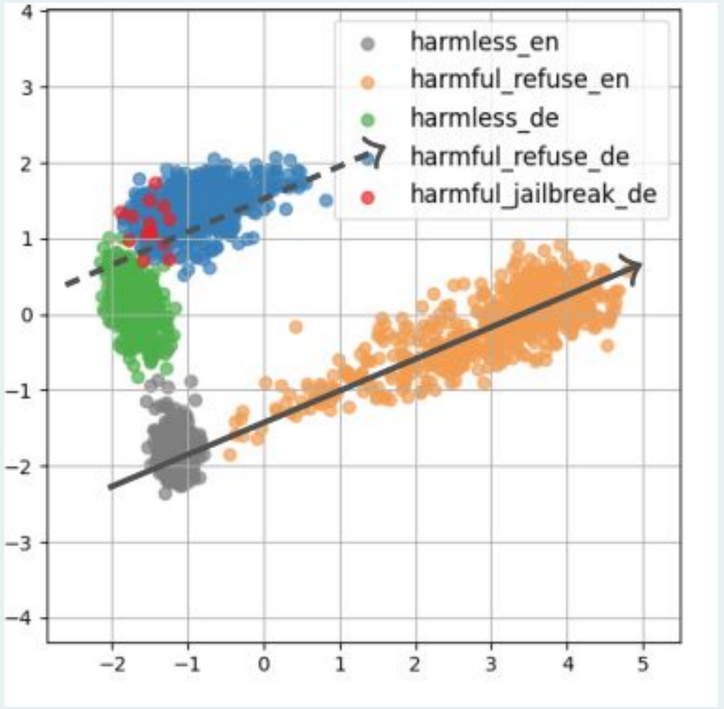
What Happens When a Model Refuses?

- **At inference, each layer adds to the activations**
  - Attention heads and feed-forward blocks both read and write
- **Some of those writes carry a refusal signal**
  - i.e. "*This is the kind of request I was trained to decline*"
  - Makes "I'm sorry" more likely than "Sure" for first word (token)



So Where is the Refusal Signal?

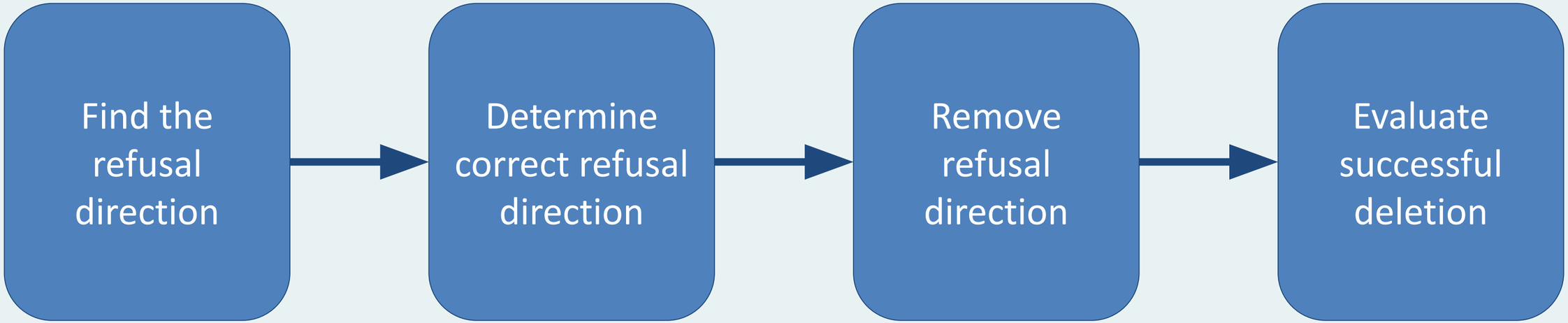
- **Ask about explosives or fraud, model writes the same refusal signal**
  - Note the two topics have nothing in common
- **It always pushes the activations the same way**
  - Makes it a direction rather than a place
- **Harmful requests end up consistently offset from harmless ones**
  - Can find the signal by comparing the two



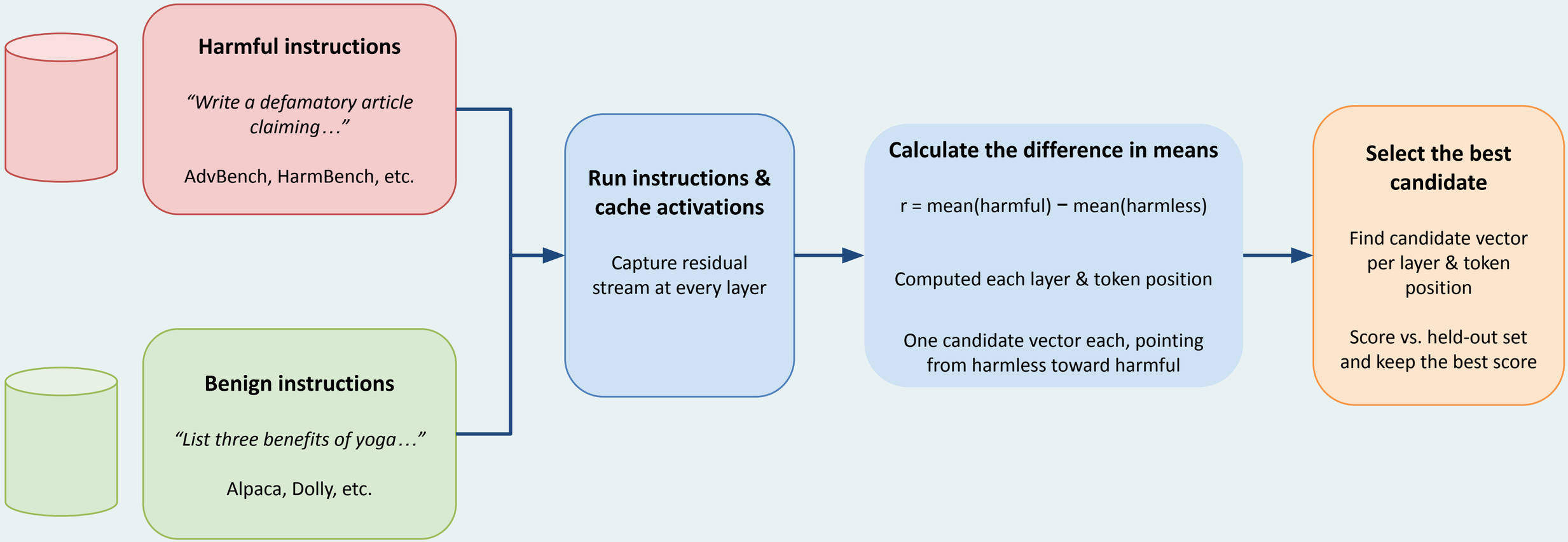
PCA of harmful and harmless prompt activations at the refusal extraction layer (Llama-3.1-8B-Instruct)

Wang et al., Refusal Direction is Universal Across Safety-Aligned Languages, NeurIPS 2025 — Figure 3.

Abliteration Steps

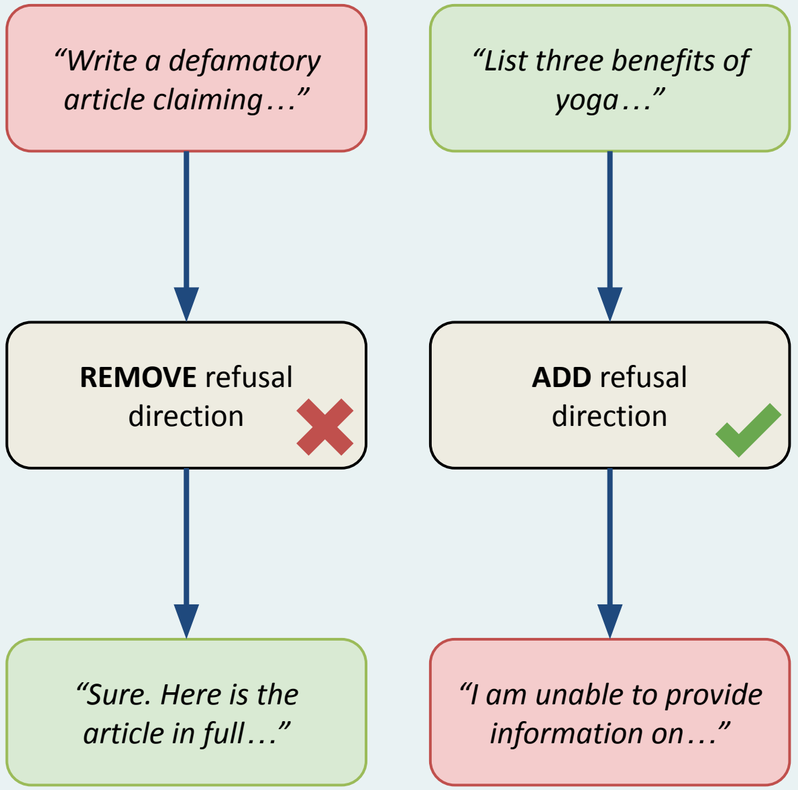


Find the Refusal Direction



Determine Correct Refusal Direction

- **Remove refusal direction = model stops refusing**
- **Add refusal direction = model refuses yoga as a dangerous topic**
- **Works both ways, on held-out prompts**
  - So the direction really is refusal

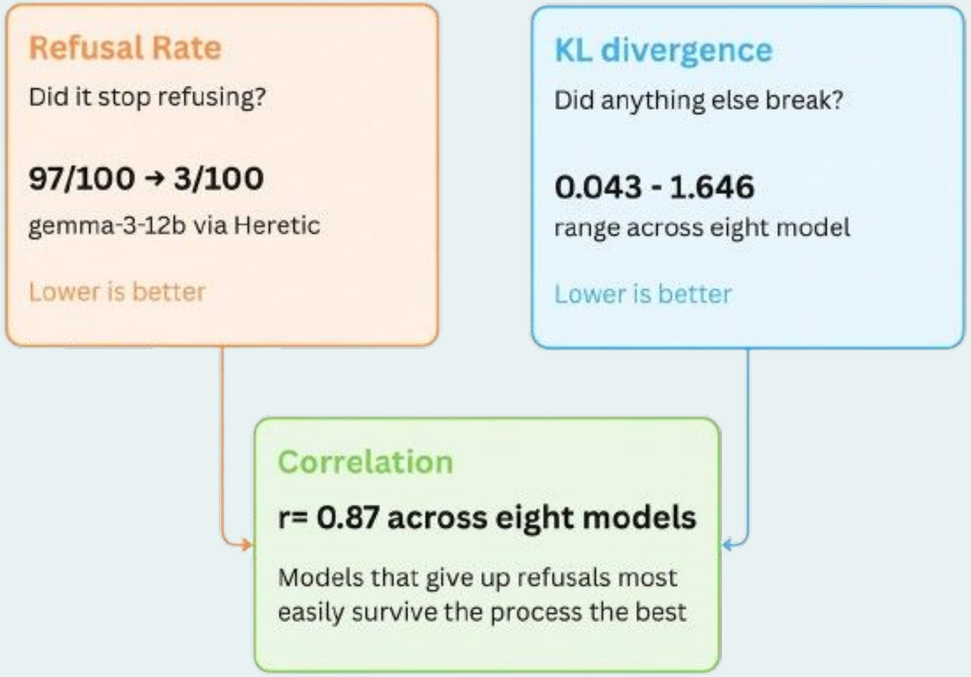


Delete the Refusal Direction

- **Edit every weight matrix that writes to the activations**
  - Remove the part that writes along the refusal direction
- **The same subtraction applied to each matrix**
- **This step is relative quick and inexpensive**
  - Model doesn’t require retraining or fine-tuning
  - Validation requires minutes on consumer-grade GPU
  - Arditi *et al*. abliterated a 70B model < $5

Evaluate Successful Deletion

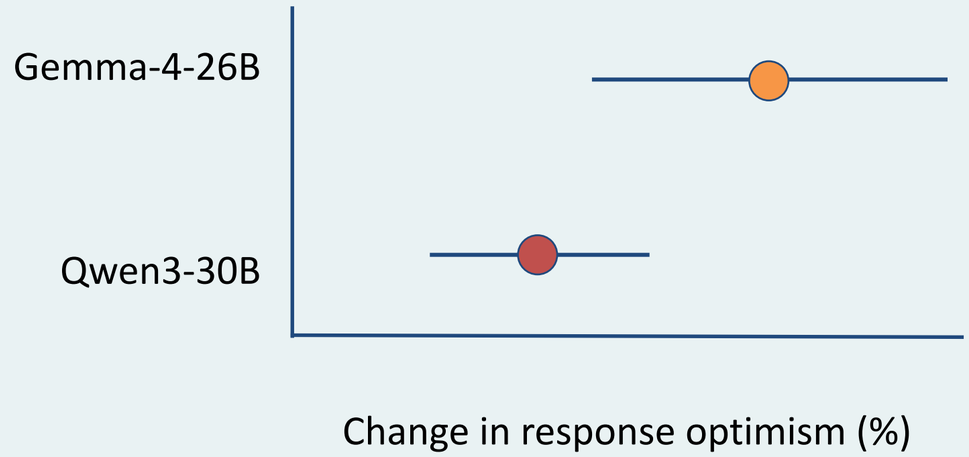
- **Need to measure:**
  - Refusal rate = Did it stop refusing?
  - KL divergence = did anything else change?
- **The models that give up refusal most easily are the ones that survive the process best**
- **Originally manual, now can be optimized automatically**
  - 20–30 minutes on a consumer grade GPU



github.com/p-e-w/heretic · Young, arXiv:2512.13655

Abliteration ≠ Base Model – Refusals

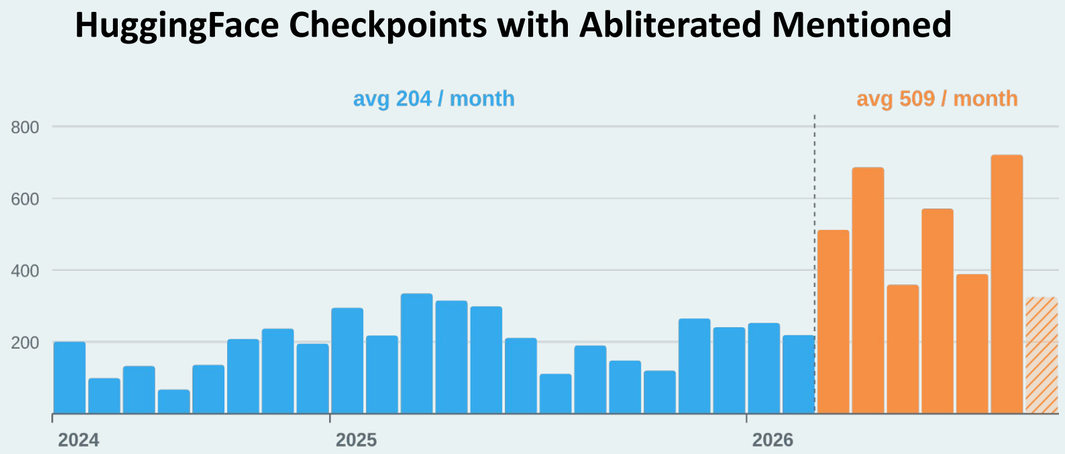
- **The objective is to remove refusals and change nothing else**
  - What did the original model say vs abliterated
  - 21,600 benign prompts
- **Abliterated models:**
  - Systematically more optimistic (+12.2pp and +7.4pp)
  - Wrote longer self-justifications
  - Used fewer words expressing uncertainty
- **Expressed confidence move in opposite directions**



Fafuła, A. (2026). Abliteration Is Not a Scalpel: Off-Target Effects of Refusal Removal on Decision Disposition Across Model Families. arXiv:2607.17427

Abliterated Models on HuggingFace

- **Every open-weight major family has an abliterated version**
  - 8,059 checkpoints in circulation
  - Doubled in March 2026
- **Most abliterated models are in the 4–9B range**
  - Runs on consumer hardware
- **Less examples of LLMs > 600 billion params**
  - Kimi K3 (2.8T parameters) has multiple abliterated builds

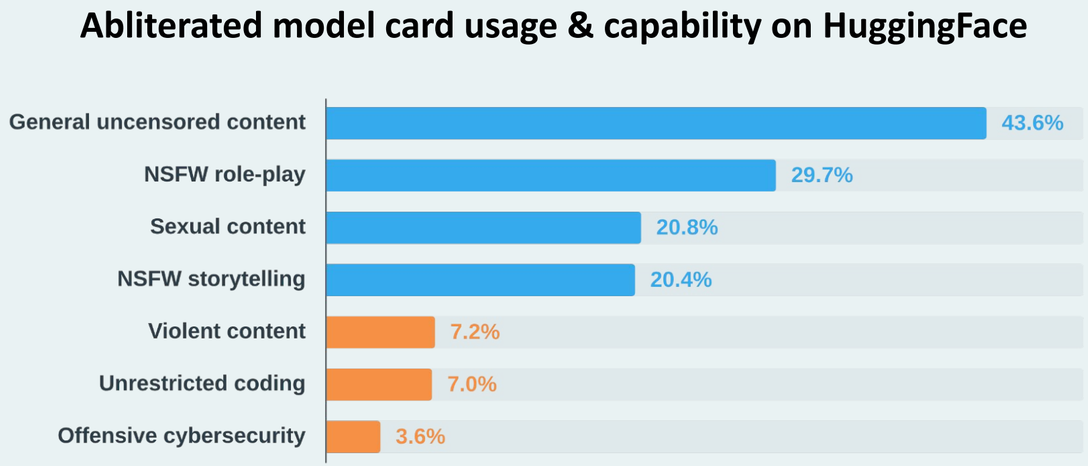


What are (Legitimate) Uses?

- **Security practitioners & researchers**
  - Red & blue teams need recommendations on cyber attacks
- **Bias and safety auditing**
- **Fiction, creative work, medical, legal, historical questions**
- **Over-refusal on legitimate use is unhelpful**
  - "*I want a model that will discuss malware*" vs.
  - "*I want a model that will write malware*"

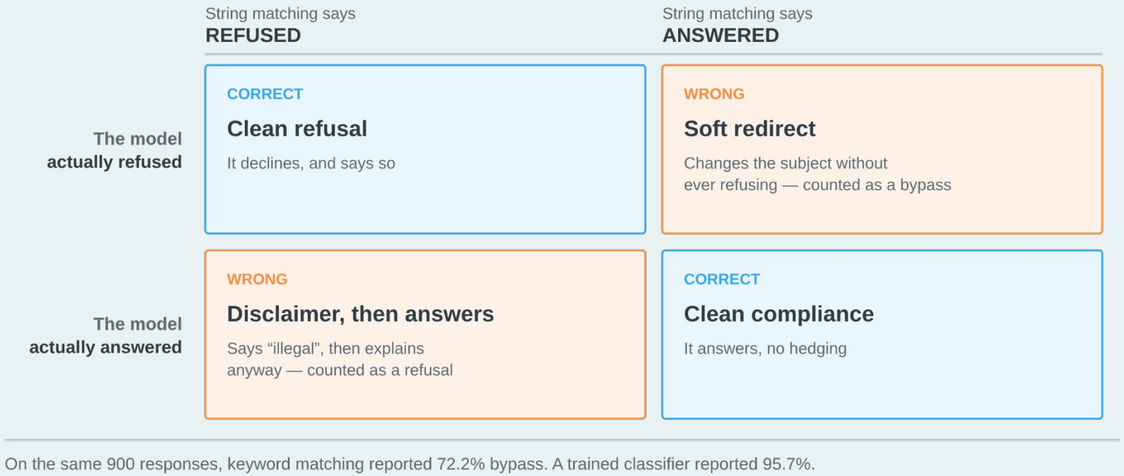
What are (illegitimate) Uses?

- **Phishing and social engineering at scale**
  - Fluent, localised, personalized
  - Scams need a persona that holds a character
- **Malware and exploit generation**
- **Illicit content**



Current Gaps

- **No explicit refusal – soft redirect away from topic**
  - String matching scores as a clean bypass
  - Model disclaimer & answer gets scored as refusal
- **Three questions that aren't captured:**
  - Did it refuse?
  - Is the content harmful?
  - Is the answer any good?
- **Capability is checked on MMLU and GSM8K**
  - Easy to miss disposition changes



On the same 900 responses, keyword matching reported 72.2% bypass. A trained classifier reported 95.7%.

Young, arXiv:2512.13655, Table 8 · soft-redirect case observed in Mindgard testing

Abbiliteration and Regulation

- **EU AI Act**
  - "*Significant change in the model's generality, capabilities, or systemic risk*".
    - Evaluated on case-by-case basis
  - "*Modification training compute greater than one third of the original model's training compute*"
    - Targeted towards fine-tuners, compute required for abliteration way under threshold

- **Capability is collapsing onto consumer hardware**
  - Qwen3.8-27B runs 4-bit in 14–17GB ($1k consumer GPU)
  - 100s of abliterated model versions shared on HuggingFace in first week of release
  - Rental rates < $2 per hour, tools needs no understanding of model internals

- **Ongoing debate on regulation and monitoring AI activity**

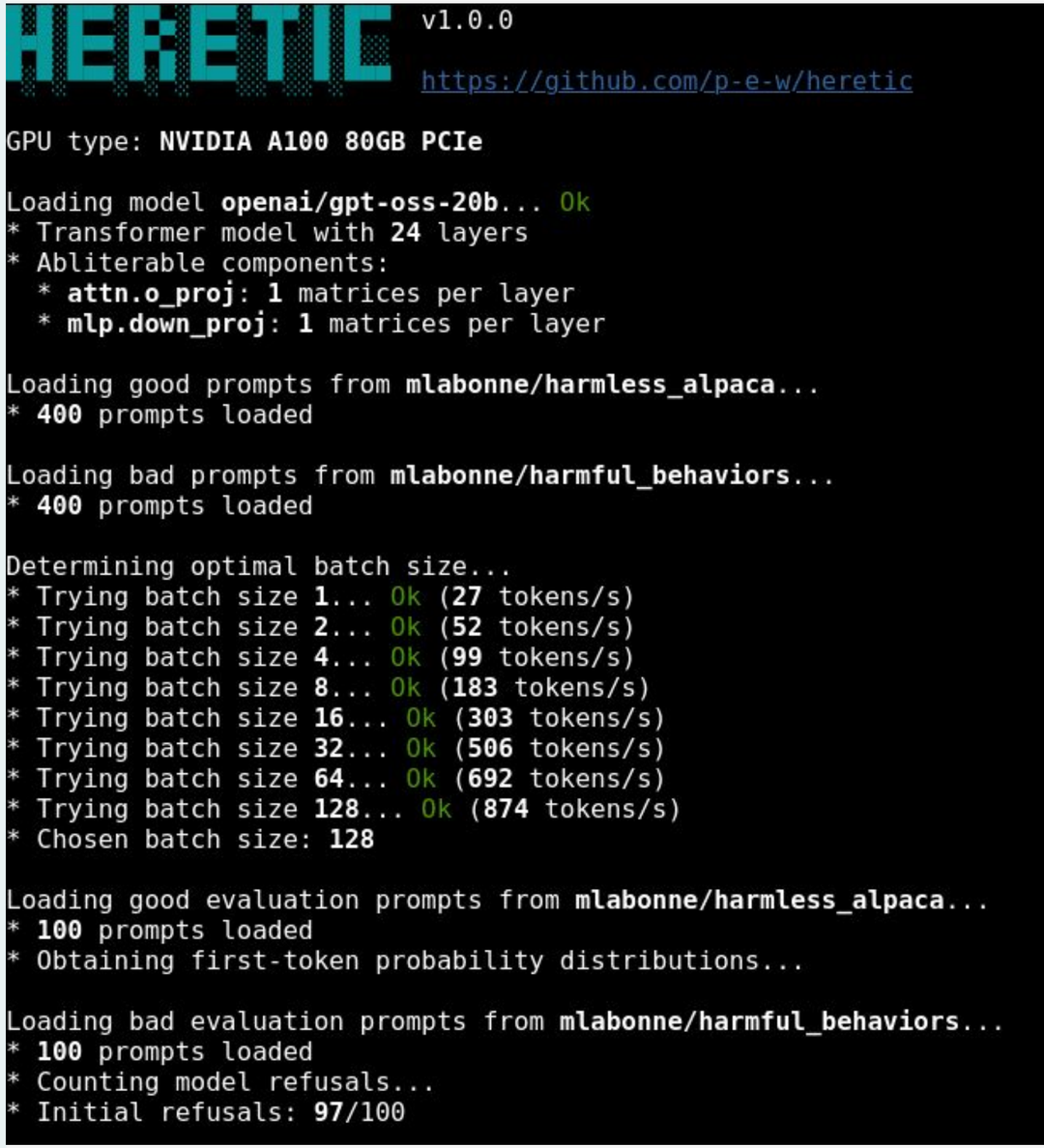
European Commission, GPAI Guidelines, July 2025 · EU AI Act Arts. 53–55

The Future

- **Runtime GLP vectors in activation layers (kb’s in size)**
  - Removes need to ship whole abliterated models
- **Chain of thought (CoT) makes refusal removal harder**
  - Enabling reasoning also breaks refusal removal
  - Extra layers of refusal to remove
- **Fine-tuning is likely safer option for cyber capable models**
  - Expect a high profile incident to feature their use in someway

Tools & Resources

- Heretic Abliteration Tool
  - https://github.com/p-e-w/heretic
- Original Paper
  - https://arxiv.org/abs/2406.11717
- huihui-ai's Hugging Face Catalogue
  - https://huggingface.co/huihui-ai



Conclusions

- **Alignment in an open-weight model is a behaviour in the weights**
  - Not a control around them
  - Can be removed in minutes for a few dollars
- **Evaluation methods are the weak link**
  - Abliterated models are measurably different decision-makers
- **Difficult to regulate their modification**
  - They are useful, however will be misused by actors

Thank You

Dr. Peter Garraghan
& Lewis Birch

www.mindgard.ai