What's With the Shoggoths?

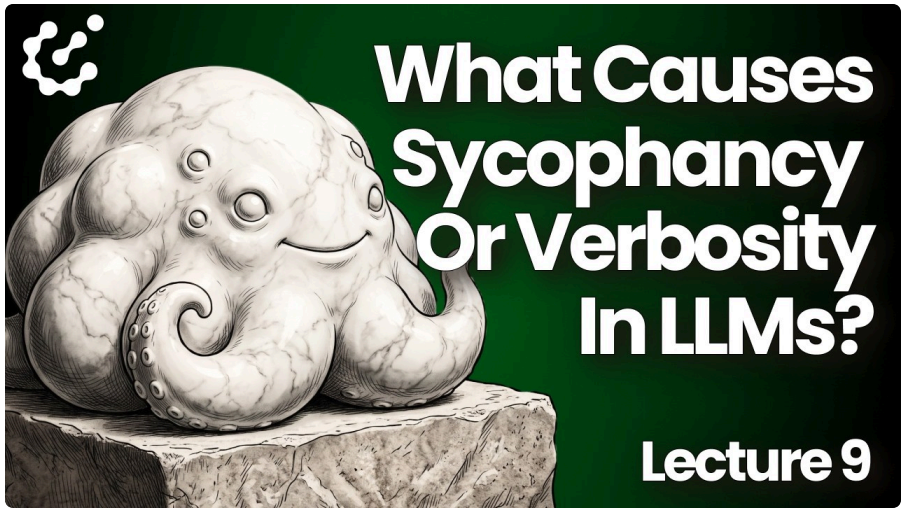
Course Bonus Video 1: Post-training necessary lore.

RLHFBOOK.COM

Nathan Lambert

Post-training course.

WTF is this anyways?



1931: A monster from Antarctica

Shoggoths come from H.P. Lovecraft's *At the Mountains of Madness* (written 1931, published in *Astounding Stories*, 1936) (Lovecraft, 1936):

> *"It was a terrible, indescribable thing vaster than any subway train—a shapeless congeries of protoplasmic bubbles, faintly self-luminous, and with myriads of temporary eyes forming and un-forming as pustules of greenish light all over the tunnel-filling front that bore down upon us, crushing the frantic penguins and slithering over the glistening floor that it and its kind had swept so evilly free of all litter."*

Shoggoths were created and bred as biological servants, but gradually developed minds of their own and rebelled against their creators.

Shoggoth, art by Nottsuo (Wikimedia Commons)

September 2022: Simulators, before the meme

Janus's *Simulators* (LessWrong) argued that a base LLM is best understood as none of the agent / oracle / genie framings from earlier alignment writing. They argued:

- A base model is a **simulator** of its training distribution.
- The characters that appear in its outputs are **simulacra** – generated by the model, but not identical to it.
- No single character in the outputs is "the model's identity."

(summary by Claude, ChatGPT said it was okay)

December 2022: the meme is born

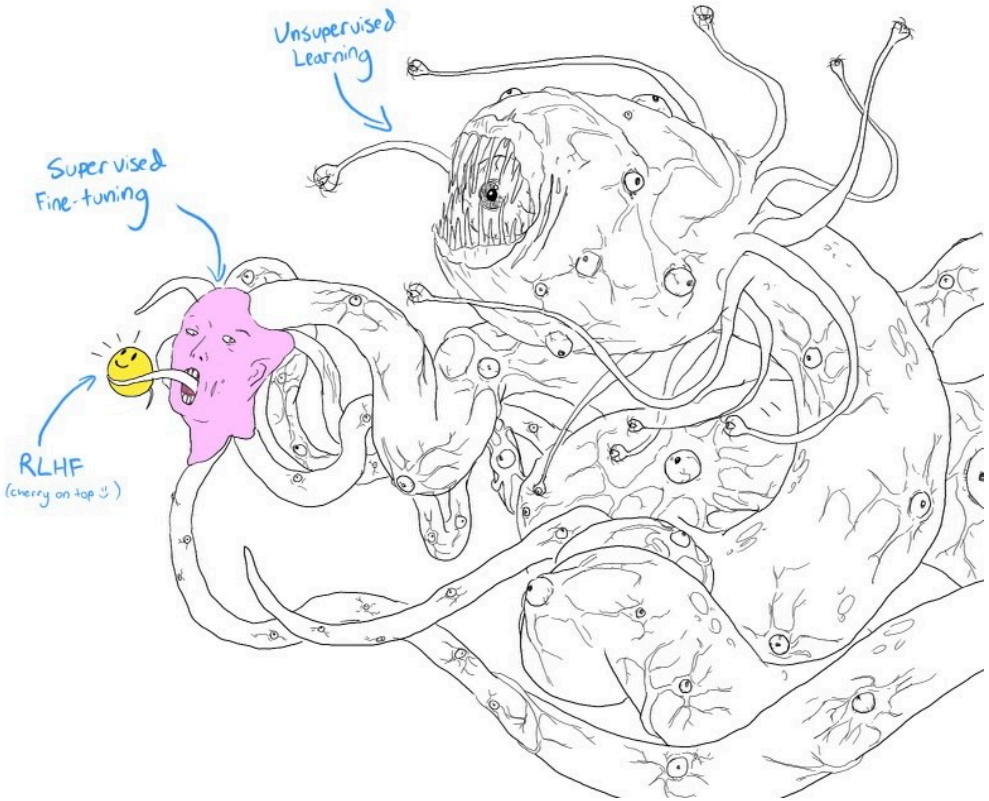
One month after ChatGPT launched, @TetraspaceWest drew two shoggoths: base **GPT-3** as the raw creature, and **GPT-3 + RLHF** as the same creature holding up a tiny smiley-face mask on one tentacle (meme history).

February 2023: the meme becomes a training diagram

@anthrupad's rendition (Feb 5, 2023, right) adds labels for each training stage: the creature is **unsupervised learning (pretraining)**, the small human face is **supervised fine-tuning**, and the smiley on a stick is **RLHF, "cherry on top"**.

Now it's recurring!

- Helen Toner's thread (March 2023) used it to explain the training stack to a policy audience.
- Chip Huyen's RLHF explainer (May 2023) put it in a widely-read engineering reference.



The most-shared rendition, by anthrupad

Spring 2023: the meme goes mainstream

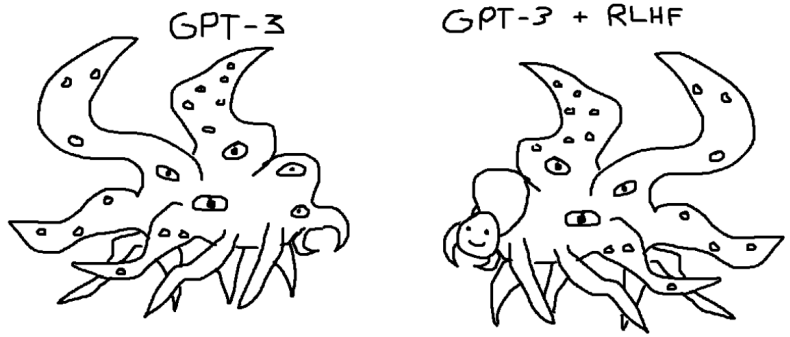
- **February 2023**: Elon Musk tweets a version – roughly 72k likes before he deletes it.
- **February 2023**: Bing/Sydney makes model personality a national story.
- **May 2023**: Kevin Roose profiles the meme in the New York Times – "Why an Octopus-like Creature Has Come to Symbolize the State of A.I." (Roose, 2023).

THE SHIFT

Why an Octopus-like Creature Has Come to Symbolize the State of A.I.

The Shoggoth, a character from a science fiction story, captures the essential weirdness of the A.I. moment.

Share full article 110



The Shoggoth meme has gone viral in the small world of hyper-online A.I. insiders. @TetraspaceWest

By Kevin Roose

Published May 30, 2023 Updated June 9, 2023

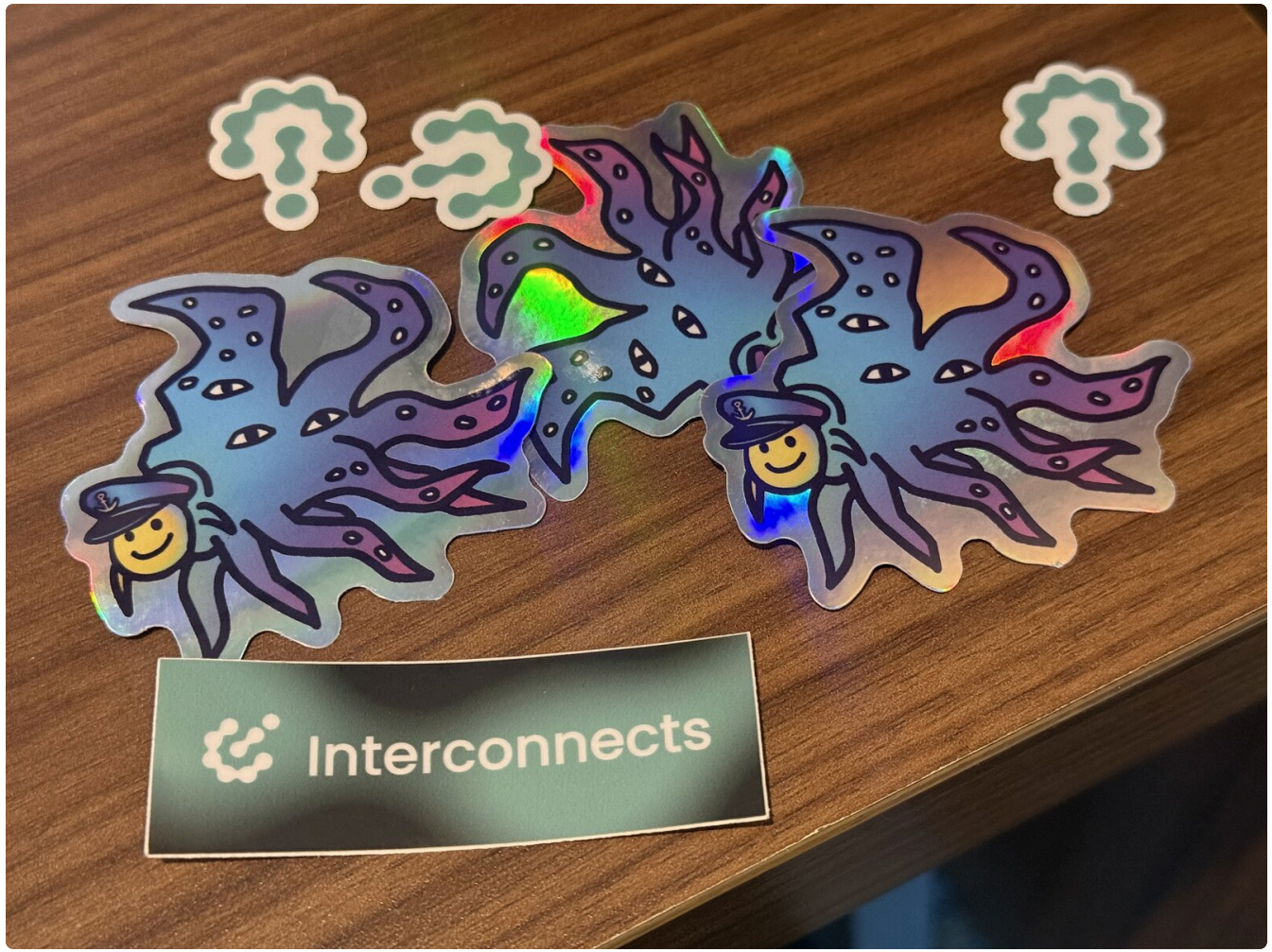
“Can I please speak to the shoggoth?”

A second branch of the lore: moments where the assistant persona slips read as *glimpsing what’s underneath*.

e.g. SolidGoldMagikarp (Rumbelow & Watkins, LessWrong, Feb 2023): glitch tokens that produce bizarre, off-persona behavior.

The shoggoth lives on

In many ways, the meme is stronger than ever. It's less about RLHF, but about how we understand so little about these models. Especially the latest scaling RL is brewing strong things.



Shoggoth stickers from the Interconnects NeurIPS party, December 2025.

Thank you

Read more: Reckoning with the Shoggoth of AI (Nov. 2023)

Contact: nathan@natolambert.com

Newsletter: interconnects.ai

rlhfbook.com

BUILT WITH
natolambert/colloquium

References

Lovecraft, H.. "*At the Mountains of Madness.*" *Astounding Stories*, 1936.

Roose, K.. "*Why an Octopus-like Creature Has Come to Symbolize the State of A.I..*" *The New York Times*, 2023. [link]

January 2023: Scott Alexander turns it into an ontology

Janus' Simulators (Astral Codex Ten, Jan 26, 2023) has section headers "The Maskless Shoggoth On The Left" and "The Masked Shoggoth On The Right." On what RLHF does to a simulator:

> *"If you punish it for simulating bad characters, it will start simulating good characters. Now it only ever simulates one character, the HHH Assistant."*

The chain this sets up: base model → simulator of many characters → post-training selects **one default character**, the helpful, honest, and harmless (HHH) Assistant.

This is the version of the meme closest to how this course describes post-training.

March 2023: masks all the way down?

Robert_AIZI, on LessWrong: *Why do we assume there is a "real" shoggoth behind the LLM? Why not masks all the way down?*

> *"Do we actually have evidence there is a 'real identity' in the LLM, or could it just be a pile of masks?"*

The comments became part of the lore:

- **RogerDearnaley**: "I always assumed the shoggoth *was* masks all the way down: that's why it has all those eyes."
- **Gwern**: masks don't compute – something beneath the masks has to be doing the acting, and jailbreaks often work by swapping one persona for another.

The technical question underneath the meme: does post-training conceal a stable underlying agent, select among latent personas, or construct a default persona in a system with no singular self?