Anthropology of Machines

From Black Box to White Box

Y. Emre Tapan

PhD Candidate, Political Science (minor: Computational Social Science)
Northeastern University

SICSS-Istanbul 2026

Question 1

Are large language models black boxes?

A Yes

B No

C It's complicated

D Other

menti.com

Question 2

Does AI need social science
— or does social science need AI?

A AI needs social science more

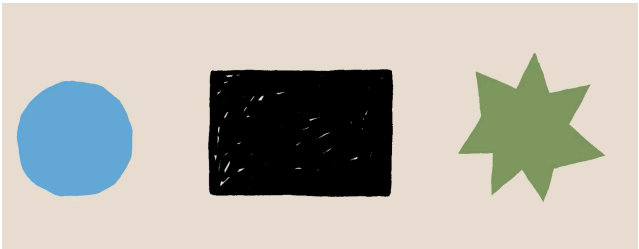
B Social science needs AI more

C Both, equally

D Neither

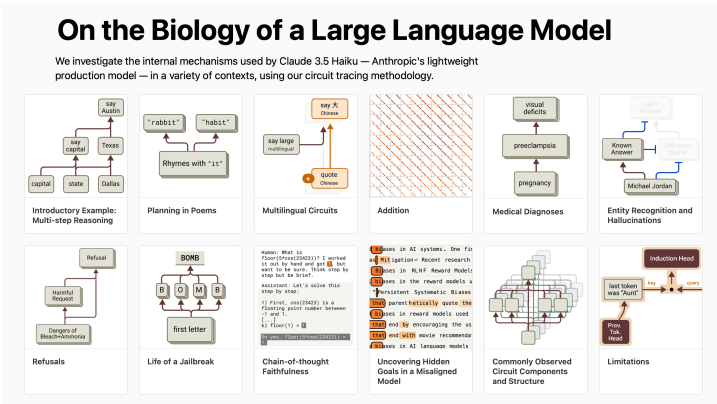
menti.com

We can now see what a model does inside



▷ play ∼90s — Anthropic, “Tracing the thoughts of a large language model”

They did not call it engineering. They called it biology.



Lindsey et al. (2025)

A mind nobody wrote — studied like an organism

A grown system. No manual.

Understood only by studying it.

The study of these machines is becoming a social science —
and it needs you.

1988: to think is to manipulate symbols

```
rule: (I am *) → (Why are you *?)
> I am sad.   WHY ARE YOU SAD?
```

A network can imitate a concept — not possess one.

Fodor & Pylyshyn (1988) · Weizenbaum (1966)

The argument that held for 36 years: systematicity

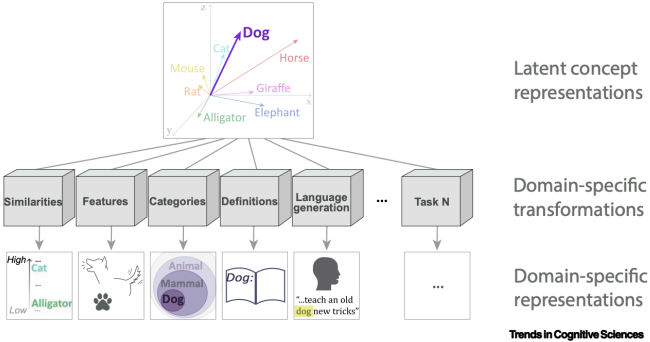
"the dog chased the cat"

$\rightleftarrows$

"the cat chased the dog"

No symbols, no thought — so, a network cannot think.

Last year's reply: concepts are (probably) vectors



A concept is a direction — not a symbol.

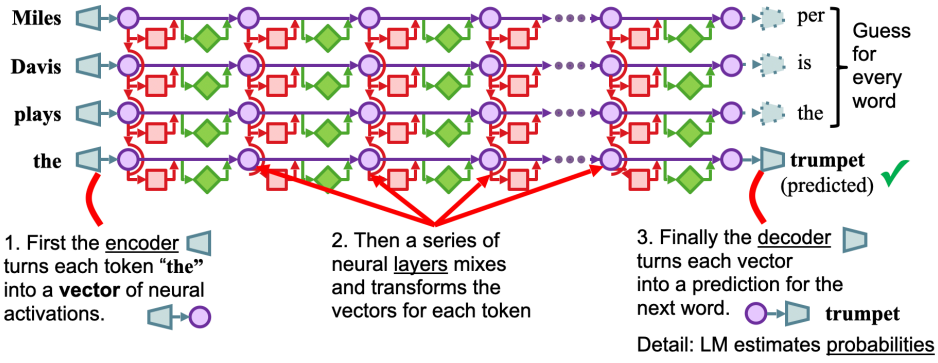
Piantadosi et al. (2024)

We can open one up and ask:

is a concept really in there, as a direction?

They could debate it. We can go look.

Meet the object: one shared stream of notes



One shared channel — the residual stream.

decoder-only · *Ferrando et al. (2024)*

The pieces: encoder, attention, MLP, decoder

Encoder is a **look-up table** from tokens to vectors.
Typical vocabulary: 50,000 or 150,000 vectors in the table.

Attention is a neural network that "remembers" recent information from contextual tokens by (1) making a **query** vector (2) to match **key** vectors (3) and gather & add **value** vectors

short-term contextual memory

MLP (multilayer perceptrons) are two-layer neural networks that match and modify single-token feature vectors

long-term parametric memory

Decoder makes a vector of 50k/150k **next-token probabilities**

Attention moves information · MLP adds knowledge.

Two ways to study a mind. We take one.

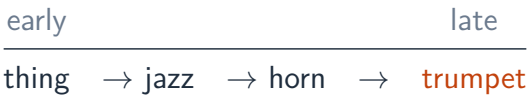
observe vs intervene

Observe: read the internal state, change nothing.

The ethnography of the model.

Method 1 — the logit lens: read the answer as it forms

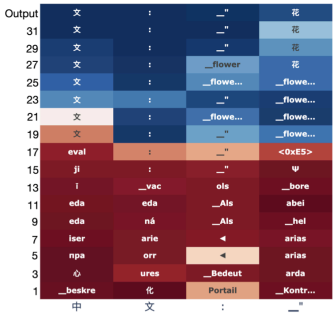
"Miles Davis plays the _____"



A reading, not a proof.

nostalgebraist (2020) · *Belrose et al. (2023)*

What that simple method found: a private language?



The middle layers think in English — still contested.

Wendler et al. (2024)

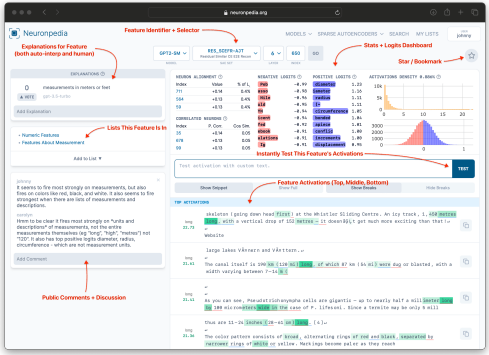
Why we can't just read the neurons

one neuron: dogs · Tuesdays · sarcasm

More concepts than neurons — the concept is a direction.

Elhage et al. (2022) · superposition

Method 2 — sparse autoencoders unmix the concepts



Reading a feature is coding data — labels can be wrong.

Bricken (2023) · *Templeton (2024)*

So — whose science is this?

Does AI need social science —
or does social science need AI?

You answered at the start. Here is my answer.

What we did was fieldwork — its hard problems are ours

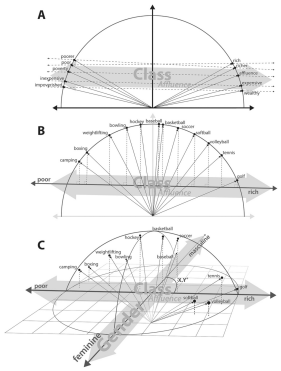
Reading a feature = thick description

Watching the lens = process-tracing

The open problems: construct validity, measurement, reliability.

Gupta (2024) · Röttger (2024) · Li et al. (2025)

Social scientists were already doing this — before LLMs



Concepts as directions — sociology, 2019.

Kozlowski, Taddy & Evans (2019) · Amer. Sociological Review

My answer to Question 2

Reading a mind you did not build —

and knowing the reading is valid —
is what this room does.

Social science can proceed without AI.
AI cannot become trustworthy without us.

Your turn — read a mind

Live — NDIF Workbench `workbench.ndif.us`

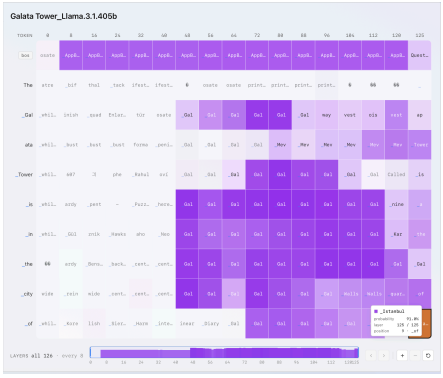
The logit lens, on a real model, in your browser — no code.

A fill-in-the-blank probe: a cloze task.

First I run one. Then you run your own.

Petroni et al. (2019) · LAMA

How to read the picture



rows = words · columns = depth (early → late)

Read the last row: the fact resolves late.

Now you — a few minutes

1. Open `workbench.ndif.us` — link in chat
2. Sign in — one click, GitHub or Google
3. **Logit Lens** + model **Llama-3.1-8B**
4. Type a cloze sentence you know the answer to
5. Read the last row — screenshot it to the Drive — link in chat

Not black. Not white.
Translucent — in patches.

local results · labels don't always replicate · scaling is open

Sharkey et al. (2025)

The other line: from observing to intervening

Everything here was observation.

The next tab — `Activation Patching` — is intervention:

editing a fact (ROME) · injecting a concept (Ji Ma, 2026).

Description comes first — and that is what we did.

Meng et al. (2022) · Ma (2026)

References i

- Lindsey et al. (2025). *On the Biology of a Large Language Model* (+ film). Anthropic.
- Fodor & Pylyshyn (1988). Connectionism and cognitive architecture. *Cognition* 28.
- Piantadosi et al. (2024). Why concepts are (probably) vectors. *TiCS* 28(9).
- Ferrando et al. (2024). *Primer on the Inner Workings of Transformer-based LMs*. arXiv:2405.00208.
- nostalgebraist (2020), *Logit Lens*; Belrose et al. (2023), *Tuned Lens*.
- Wendler et al. (2024). *Do Llamas Work in English?* ACL.
- Elhage et al. (2022), *Toy Models of Superposition*; Bricken (2023); Templeton (2024). Anthropic.
- Gupta et al. (2024); Röttger et al. (2024); Li et al. (2025), *Interpretability Illusions*.
- Kozlowski, Taddy & Evans (2019). The Geometry of Culture. *ASR* 84(5).

References ii

- Petroni et al. (2019), LAMA; Meng, Bau et al. (2022), ROME; Ma (2026), *Soc. Methodology* 56(1).
- Weizenbaum (1966); Amodei (2025); Sharkey et al. (2025).

- Tools: neuronpedia.org · workbench.ndif.us · ndif.us · neural-mechanics.baulab.info

We spent a century learning to study people
who do not fully know their own minds.

Now there is a new kind of mind —
and it is waiting for its anthropologists.