Who am I?

Jeremy R. Manning, Ph.D.

Associate Professor | Psychological & Brain Sciences | | Moore 349

context-lab.com

ContextLab

Research focus

How do our brains support our ongoing conscious thoughts, and how (and what) do we remember?

Key areas

Learning and memory, education technology, brain network dynamics, data science, NLP

Approach

Theory, models, experiments, neuroimaging

Training

B.S., Neuroscience & Computer Science

Ph.D., Neuroscience

Postdoc, Computer Science & Neuroscience

Funding & collaborators

NIH National Institutes of Health   NSF   DARPA

intel labs   Meta   Google   amazon

What is this course about?

Course content

We will explore how machines can understand and generate human language:

- Building conversational agents from scratch
- Understanding how language relates to thought
- Hands-on programming with real models
- Critical thinking about AI consciousness

Course design

This course is experiential! You will learn by doing: coding, experimenting, discussing, and researching.

Course structure

Grading

- Weekly (roughly) short projects (5): 75%
- Final project: 25%
  - Can work individually or in groups
- See <u>syllabus</u> for additional details

Tools

- Google Colaboratory
- HuggingFace
- GitHub / Discord
- GenAI (Claude, ChatGPT, Gemini)

The big questions

1. Can machines truly *understand* language?

2. What is the relationship between **language** and **thought**?

3. Can statistical patterns capture **meaning**? How? Under which circumstances?

4. Can AI be conscious? If so, what are the implications?

> **Note**
>
> These are not just "fluff" questions: they are at the heart of cognitive science, philosophy, psychology, and neuroscience!

Discussion: is ChatGPT conscious?

- What does "conscious" even mean?

- How would we test for consciousness?

- Does it matter if ChatGPT *seems* conscious?

For further consideration

What are the implications for:

- Ourselves

- Other animals

- Other life forms in general (aliens? synthetic life?)

- Policy, ethics, and society more broadly

What is consciousness?

- **Phenomenal:** Subjective experience—the "what it is like" quality of sensations and emotions

- **Access:** Information available for reasoning, reporting, and guiding voluntary behavior

- **Self-awareness:** Knowledge of one's own mental states, including recognizing oneself as a distinct entity

Example of phenomenal consciousness
The redness of red, pain, taste of coffee

Example of access consciousness
Being able to report on and use information to guide behavior

Example of self-awareness
Knowing that you are thinking, or recognizing your own emotions

For further consideration
If ChatGPT says "I feel happy," does it actually *feel* anything?

The hard problem

Humans

- Share similar biology
- Have our own conscious experiences
- Behave consistently with having experiences (e.g., of being human, living in the world, etc.)

AI

- Completely different "biology" (silicon vs. neurons)
- No shared evolutionary history
- Can produce human-like behavior without themselves being human

Why is it difficult to know if AI is conscious?
We cannot directly observe consciousness — even in other humans!

The Chinese room argument

John Searle (1980)

- <u>Thought experiment</u> about understanding vs. simulation.
- **The setup:** Person in room with Chinese symbols and rule book
- Does following rules = understanding? Searle argues: **No!**

| Chinese question | → | Rule book | → | Chinese answer |
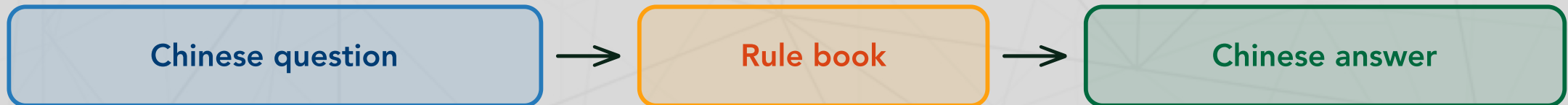
Person follows rules but does not understand Chinese

The Chinese room argument

The scenario:

1. You are locked in a room
2. Chinese characters slide under the door
3. You look them up in a rule book
4. You copy out the corresponding response
5. You slide it back out

Observer sees: Perfect Chinese conversation!
Your perspective: Just symbol manipulation.

Concrete example:

```
1    Input:  你好吗? (How are you?)
2    Lookup: Rule #42,938 → 我很好
3    Output: 我很好 (I am fine)
```

You matched symbols without knowing:

- What "你好吗" means
- That it's a greeting
- What "你" (you) refers to

The key insight

Syntax (symbol manipulation) is not sufficient for semantics (meaning).

Volition

- Another critical aspect of the human conscious experience is the ability to **decide** (how to act, what to think, etc.)

- Modern LLMs are trained to **respond** to other inputs (i.e., produce statistically likely sequence completions), but they cannot themselves initiate new or unexpected actions

Think about it!

LLMs are like a bellows that can only blow air if someone else is pumping it. When not in use, they are static. They can not sense the passage of time. They cease to exist between invocations.

Current scientific consensus

Survey says...

Most cognitive scientists and AI researchers agree: **current LLMs are *not* conscious.**

- No sensory-motor grounding in the world

- No persistent self-model or goals

- Pattern matching $\neq$ understanding

Another angle: relating language and thought

Discussion

Do you need language to think? Does language *shape* how you think?

- Possibility 1: Language is necessary for thought
- Possibility 2: Language is just a tool for communication

The language-thought spectrum

| Strong Whorfian | → | Weak Whorfian | → | Moderate | → | Language-of-Thought |

Language shapes thought (left) to language independent of thought (right)

Current evidence points toward the middle: language and thought interact in complex ways, but are not identical.

Evidence: the language network

- The brain has a **specialized language network**

- Distinct (as measured using neuroimaging and lesion studies) from: reasoning, math, social cognition, music

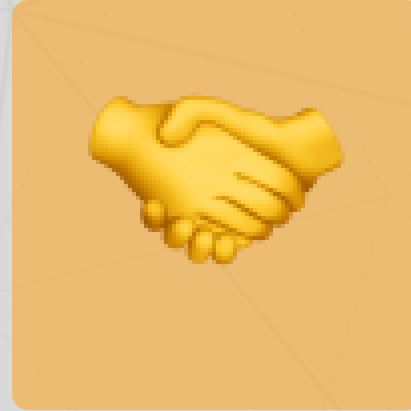- **Implication:** Language and thought are _separable_ in the brain!
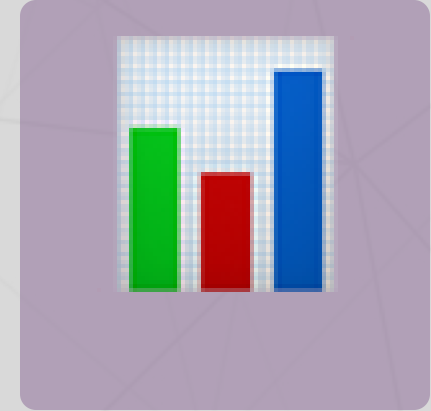
Brain networks
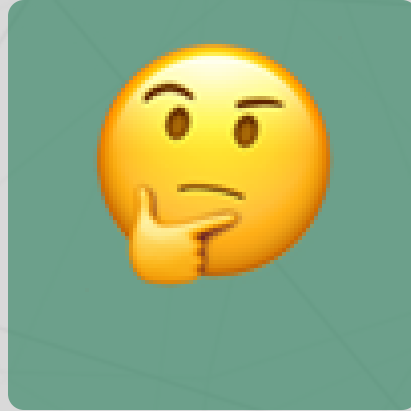
Language   Reasoning   Social   Math

Key insight

Language is a specialized system, *not* the basis of all thought! This has fascinating implications for LLMs: we've built machines that can *interact* using language, but this doesn't necessarily mean they understand it in any recognizable way.

However, language can *shape* thought

Having words for things affects how we *see* them:

- Speed up visual search

- Alter color perception

- Influence object categorization

Example: Russian blue

English: "blue"   English: "blue"   Russian: "siniy"   Russian: "goluboy"

Key finding

English speakers are slower to distinguish shades of blue than Russian speakers: language categories affect *perception*, not just description!

Example: Russian blue

English: "blue"   English: "blue"   Russian: "siniy"   Russian: "goluboy"

What this tells us
- Having distinct words for colors creates **perceptual boundaries**
- The brain literally processes colors faster when they cross linguistic categories
- Language doesn't just describe reality; it *shapes* how we see it

Implication for LLMs

If language shapes perception, what "perception" do LLMs have? They experience language *about* color, but never color itself.

The grounding problem

Definition

How do symbols (words) get their meaning?

- For **humans**, the meanings of symbols are learned through *experience*:
  - See, touch, taste objects
  - Act in the world
  - We learn to associate symbols with our experiences

- For **LLMs**, the meanings of symbols are learned through *statistics*:
  - Learn patterns in text
  - Generate text based on those patterns
  - No direct connection with the external world

The grounding problem example: what is "coffee"

Humans learn about coffee through experience:

- Bitter taste
- Rich aroma
- Warm ceramic mug
- Morning ritual feeling
- Caffeine effect on body

These experiences **ground** the word.

LLMs learn about coffee through statistics:

- Often appears with "morning"
- Frequently near "cup", "drink"
- Associated with "caffeine"
- Patterns: "I need my ___"
- Context: restaurants, breakfast

This is pure pattern matching, not experience-based.

What do you think?
Can statistical patterns **ever** capture what it *feels like* to drink coffee? Does human learning *not* amount to statistics? Is it about the style of learning per se, our about the kinds of data we are learning from?

How LLMs actually work: a preview

Next-token prediction

At their core, LLMs are trained to predict what comes next. Given a sequence of words, what is the most likely continuation?

Suppose we start with some input text: "The cat sat on the"

LLMs predict probability for each possible next word:

```
1  predictions = {
2      "mat": 0.35,          # Most likely
3      "floor": 0.20,
4      "couch": 0.15,
5      "roof": 0.05,
6      "elephant": 0.001   # Very unlikely
7  }
```

Key insight

LLMs don't "know" what cats are. They've just seen "cat sat on the mat" many times!

Our approach (in this course)

Philosophy of this course

We will build language models *from scratch* to understand **what they can and cannot do**. By learning about the inner workings of LLMs, we can better understand their capabilities and limitations *and our **own** capabilities and limitations.*

String manipulation → Classifiers → Embeddings → Retrieval systems → LLMs

You'll progress from building simple string manipulation models to modern LLMs. We will make heavy use of GenAI (vibe coding) to enable us to build and iterate quickly.

Up next...

Lecture 2 (Wednesday)

Pattern matching and ELIZA

- Introduction to ELIZA
- The ELIZA effect
- String manipulation and regular expressions

Lecture 3 (Thursday/X-hour)

ELIZA implementation

- Complete architecture
- Assignment 1 details
- Coding together (time permitting)

Required readings for this week

1. Weizenbaum (1966): ELIZA

2. Fedorenko et al. (2024): The language network

3. Lupyan et al. (2020): Effects of language on visual perception

Tip

Start with Weizenbaum— it will help you understand the fundamentals and give you important historical context!

Key ideas from today

1. **Consciousness is complex:** multiple types, hard to define

2. **Language ≠ Thought:** but they interact in interesting ways

3. **LLMs are not conscious:** they are sophisticated pattern matchers

4. **Grounding matters:** meaning comes from experience

5. **Building to understand:** we'll build (and play around with) real models to understand what they can and can't do
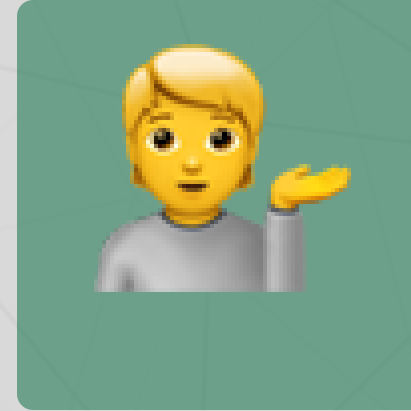
Questions? Want to chat more?

<u>Email</u> me   Join our <u>Discord</u>   Come to <u>office hours</u>

Tip

This course will move *very* quickly. **Please** reach out if you have questions, comments, concerns, or just want to chat!