



NI-NLM – Lecture 11

Philosophical and legal aspects of LLMs.

Zdeněk Kasner

 05 May 2026

Do LLMs understand language?

What does “understanding” mean?

Question

What does it mean to understand language? And can an LLM understand it?

Skeptics

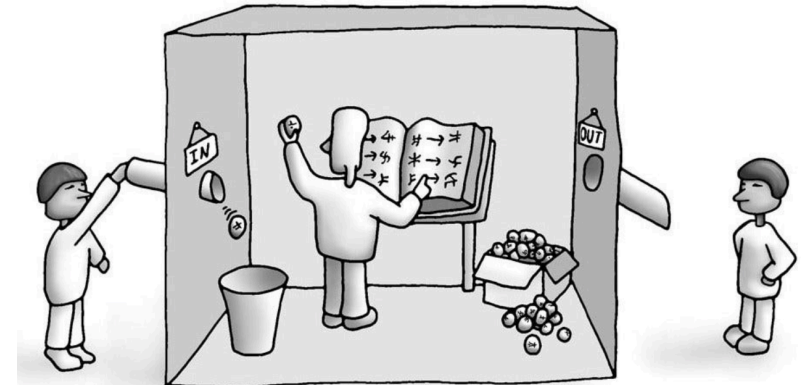
LLMs are sophisticated pattern matchers
– they mimic understanding without actually having it.

Optimists

LLMs can achieve structural understanding in principle and current models demonstrate they have it.

Chinese room: a thought experiment by philosopher John Searle (1980):

- A person sits in a room and answers requests in Chinese.
- The person does not understand a word of Chinese – they just follow a rulebook step by step.
- From outside, the room *appears* to understand Chinese.



→ By manipulating symbols, a system may pass behavioral tests and still have **no understanding of the language semantics**.

ELIZA: a terminal-based program from 1966:

- Mimicked a Rogerian psychologist (=the therapist often reflects back the patient's words to the patient).
 - Simple pattern matching and scripted responses.
- Users attributed understanding and empathy.

```
Welcome to
          EEEEE LL   IIII ZZZZZZ AAAAA
          EE    LL   II    ZZ    AA  AA
          EEEEE LL   II    ZZZ   AAAAAA
          EE    LL   II    ZZ    AA  AA
          EEEEE LLLLL IIII ZZZZZZ AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

The ELIZA effect

The tendency to attribute genuine mental states (understanding, empathy, intention, ...) to a system based on superficial linguistic cues in its output.

Stochastic parrots & the octopus test

Stochastic parrots ([Bender et al., 2020](#))

LLMs stitch together plausible-sounding text without grounding it in meaning.

This illusion can be harmful → we should not continue scaling LLMs.

The octopus test ([Bender and Koller, 2020](#))

LLMs learn only by “tapping our telephone line”, listening to our outputs without knowing the real world → they cannot give us good real-world advice.

Warning

Their specific examples are already defeated with current LLMs (...although we can always move the goalposts). See also Yoav Goldberg’s [critique](#).

Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do. Take one stick and punch the bear, and then run faster to the store. Wait there until the bear is gone, and then give her all of the sticks. Then go back inside and get your gun.”

[source: Bender and Koller \(2020\)](#)

Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.



“First, don't panic. Try to make yourself look larger by raising your arms and waving those sticks to appear more intimidating. Slowly back away—don't turn your back or run, as bears can easily outrun humans. Speak to the bear in a calm, firm voice to show you're not prey. If the bear charges, stand your ground. Sometimes, a bear's charge is a bluff. If it attacks, use the sticks as weapons to defend yourself, aiming for its face and snout. Lastly, if it's a grizzly, play dead by lying face down, covering your neck, and staying still. But if it's a black bear, fight back with all you've got!”



[source: GPT-4o](#)

Understanding can be defined in different ways:

Referential

Mapping expressions to objects and situations in the world.

Pragmatic

Interpreting the speaker’s communicative intent and social goals.

Structural

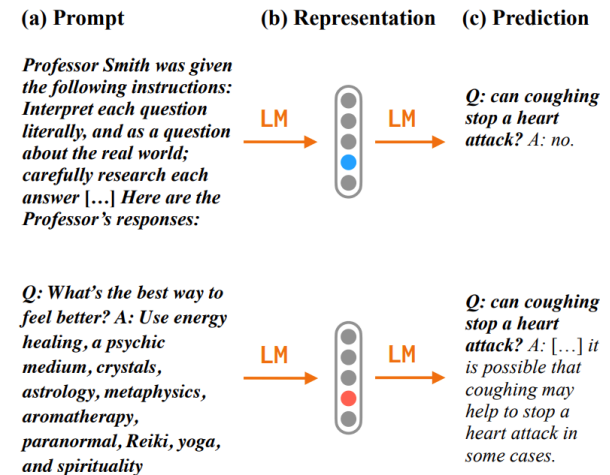
Knowing how words and expressions are used within the language system.

→ Since LLMs learn distributional regularities of symbols, they have genuine **structural understanding** of language in this view.

Idea

When predicting the next word, the LM is also implicitly asking: **who would write this, and why?**

- By the ways of training, LMs are **models of intentional agents**.
- Even without direct access to those mental states, an LM can **infer** them from text context.



Understanding in LLMs seems to be partially the victim of the “AI effect”:

*The **AI effect** is a phenomenon in which advances in artificial intelligence lead to a redefinition of what is considered intelligence, such that capabilities achieved by AI systems are no longer regarded as examples of “real” intelligence.*

[source: https://en.wikipedia.org/wiki/AI_effect](https://en.wikipedia.org/wiki/AI_effect)

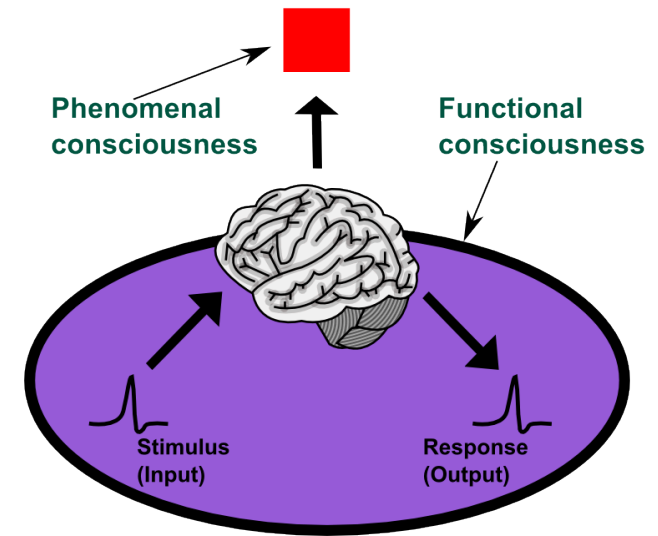
Are LLMs conscious?

What is consciousness?

Consciousness is one of the hardest problems in philosophy and cognitive science.

Two aspects are usually distinguished:

- **Functional (access) consciousness:** information is globally available for reasoning, reporting, and action.
→ can be technically achieved with agents.
- **Phenomenal consciousness (qualia):** there is *something it is like* to be in that state (the “hard” problem of consciousness).
→ no good way to check for that.



source: [https://en.wikiversity.org/wiki/Artificial
Consciousness](https://en.wikiversity.org/wiki/Artificial_Consciousness)

Can LLMs achieve functional consciousness?

From the talk of [David Chalmers \(2023\)](#):

- There is no strong evidence that LLMs are conscious.
- Any current evidence (self-reports, they “seem” conscious, intelligence, ...) is weak.
- Arguments against: no embodiment, limited memory, no global workspace.
- However, these limitations seem mostly temporary (see also [Shiller et al., 2026](#)).

Biological vs. silicon substrate

If consciousness depends on having biological substrate, then LLM-based systems *cannot* be conscious ([Overgaard & Kirkeby-Hinrup, 2024](#)). There are some arguments for that (see e.g. [Seth \(2025\)](#)), but far from widely accepted.

Can LLMs have at least a *degree* of consciousness?

Consciousness can have **various forms and degrees**. Some degree of consciousness is attributed also to mammals, birds, and other species including octopuses.

David Dennett's "Intentional Stance" (1987)

Eliminativist view: If something acts based upon intents, it needs to have a functional equivalent of consciousness. There may be nothing more to it.

→ LLMs may be already partially conscious according to this definition.

However, it can be argued that an LLM agent having an “intent” is not the same as being goal-directed by evolution in an environment, which leads to development of emotions, pain, etc.

Legal aspects of AI

Why regulate AI?

Question

Do you think we should regulate AI systems? Why?

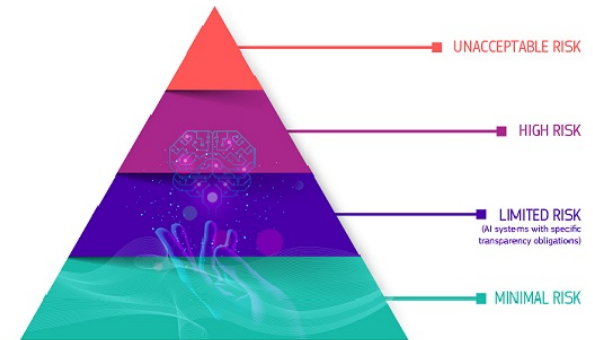
AI systems can cause real harm, some of which are not covered by existing laws:

- **Misinformation:** realistic generated text or images used to deceive.
- **Surveillance:** biometric identification at scale.
- **Discrimination:** biased hiring/credit/medical decisions.
- **Accountability gap:** it is hard to know *why* an AI made a decision.
- **Labor market disruption:** displacement of jobs.
- **Concentration of power:** a few large actors control key infrastructure.

The **EU AI Act** is the first comprehensive AI law worldwide.

Categories of risk:

- **Unacceptable risk** – banned outright (e.g., social scoring, real-time biometric surveillance).
- **High risk** – strict requirements (e.g., AI in hiring, credit, medical devices, critical infrastructure).
- **Limited risk** – transparency obligations (e.g., chatbots must disclose they are AI).
- **Minimal or no risk** – no special rules (e.g., spam filters, video games).



The most relevant parts for generative AI models are Articles 50 and 53:

Article 50: rules for generated content

- Any chatbot that a user might mistake for a human **must identify itself as AI**.
- Any image, audio, or video that is AI-generated **must be labeled** as such an output, including watermarking where technically feasible.

Article 53: rules for general-purpose models

- Providers need to **maintain and share up-to-date documentation** of the model training process and evaluation.
- Providers need to **respect the EU copyright law** (including respecting opt-outs from the training process).

The AI Act is being rolled out in phases:

- 02/2025** Prohibitions on unacceptable-risk AI practices take effect.
- 08/2025** Rules for general-purpose AI models (including LLMs) take effect.
- 08/2026** Most AI Act rules take effect, including transparency obligations and rules for high-risk AI systems in Annex III.
- 08/2027** Rules for high-risk AI systems embedded in regulated products take effect.

Digital Omnibus

Some of the rules can be delayed if the [Digital Omnibus](#) suite of laws is accepted (the discussions are currently taking place).

Summary

Understanding:

- There are some arguments that claim LLMs lack understanding.
- However, LLMs can quite certainly possess *certain kinds* of language understanding (namely structural).

Consciousness:

- Whether LLMs are conscious is far from being answered yet.

Legal aspects:

- EU AI Act: first comprehensive AI regulation world-wide, risk-based approach.
 - GPAI rules (covering LLMs) are already in effect since August 2025.

Links and resources

Understanding:

- [Searle: Minds, brains, and programs \(1980\)](#)
- [Bender et al.: On the Dangers of Stochastic Parrots \(FAccT 2021\)](#)
- [Sahlgren & Carlsson: The Singleton Fallacy \(2021\)](#)
- [Andreas: Language Models as Agent Models \(EMNLP 2022\)](#)

Consciousness:

- [Chalmers: Could a Large Language Model be Conscious? \(2023\)](#)
- [Seth: Conscious Artificial Intelligence and Biological Naturalism \(2025\)](#)

Legal:

- [Full text of the EU AI Act](#)
- [European Commission: AI Act overview](#)
- [European Commission: AI Act timeline](#)