

GRADES 9–12 · AGES 14–18 · 60 MINUTES · SOCRATIC SEMINAR

# Can a machine understand anything?

## THE PROPOSITION

**"A large language model understands the text it produces."**

A Socratic seminar on the oldest AI question with the newest evidence. Students work from short definitions of intelligence, consciousness, and stochastic prediction, and are pushed to say what 'understand' would even mean – and whether the answer changes anything about how we should use the tools.

## STUDENTS WILL BE ABLE TO

- Explain next-token prediction and why it can produce outputs that look like reasoning.
- Distinguish behavioral, functional, and phenomenal claims about understanding.
- Sustain a seminar that cites the text, builds on others, and tolerates an open question.

## VOCABULARY FROM THE FIELD GUIDE

Large language model (LLM) · Stochastic / nondeterministic · Intelligence (what we mean) · Consciousness (in machines) · AGI (artificial general intelligence) · Benchmark

## WHAT'S IN THIS PACK

- Teacher guide: background, vocabulary, agenda, format, positions, questions, roles, rubric.
- Student handouts (photocopy one per student): position organizer, sentence starters, exit ticket.
- Standards: CCSS ELA SL.9–10.1 / SL.11–12.1 (seminar); CCSS ELA RI.11–12.7 (integrate and evaluate multiple sources); CCSS ELA W.11–12.1 (argument); AP Seminar / philosophy elective alignment: argument analysis.

## PRINT

US Letter, plain paper, duplex. Teacher pages once; student pages one per student. Free to photocopy for classroom and homeschool use. Not for resale.

# Background for the teacher

Read this before class. Students get a version of the same facts on their handout; you get the framing and the traps.

A language model is trained to predict the next token. That objective, at scale, yields systems that pass professional exams, write working code, and explain their own answers – and also fail at simple counting, invent citations, and contradict themselves. 'Stochastic parrot' (it recombines patterns without meaning) and 'emergent reasoning' (the patterns encode reasoning) are the two poles; most researchers are somewhere between.

Three different questions hide in 'does it understand': Behavioral – does it act as if it understands? (Often, yes.) Functional – does it have internal representations that track meaning? (Interpretability research suggests some do, partially.) Phenomenal – is there something it is like to be the model? (No evidence, and no agreed test.) The seminar goes better when students say which one they're arguing.

The Chinese Room (Searle, 1980) argues that symbol manipulation isn't understanding no matter how good the output. The Turing test argues that behavior is all we ever have, even for each other. Both arguments are older than the technology and both still bite. Students do not need to resolve them – they need to see why their answer might change how much they trust the tool.

## RUNNING DISCUSSION IN GRADES 9–12

Stakes are real: college, jobs, creative work, civil rights. Students can steelman, cross-examine, and write policy.

- Insist on the steelman. Before rebutting, a speaker restates the other side's best version until they say 'yes, that's it.'
- Cross-examination is a question period, not a speech. Two minutes, questions only, no editorializing.
- Push past 'it depends.' Ask: depends on what, exactly – and who decides?
- Close with a written position. Speaking is the practice; the paragraph is the assessment.

# Vocabulary

Pulled from the Hi, Bot AI concepts field guide ([hibot.space/learn/concepts](https://hibot.space/learn/concepts)). Read each card aloud, then have the class rewrite it in ten words or fewer.

## Large language model (LLM)

**The kind of model behind most chatbots – a very advanced next-word guesser.**

An LLM is trained on enormous amounts of text and learns to predict what token comes next given everything before it. That simple objective, at scale, produces fluent answers, code, and summaries – but fluency is not the same as understanding or truth. Most consumer "AI chat" today is an LLM plus safety layers, tools, and a user interface.

### WHERE PEOPLE ARGUE

Calling an LLM a "stochastic parrot" vs. "reasoning engine" is a whole philosophical industry. The practical middle ground: it's pattern completion that often looks like reasoning because language encodes reasoning patterns.

## Stochastic / nondeterministic

**The same prompt can produce different answers – randomness is built in.**

Most generative models sample from probability distributions – they're stochastic. Temperature and seed settings control randomness. That's why "ask twice" is a legitimate test. Nondeterminism complicates testing, grading, and compliance workflows that assume fixed outputs.

## Intelligence (what we mean)

**Performance on tasks vs. something deeper – the word hides disagreement.**

In AI, "intelligence" often means measurable task success: exam scores, coding puzzles, game playing. In human terms it includes judgment, values, social context, and learning from few examples. Conflating the two leads to both overhype ("it's thinking") and overpanic ("it's already smarter than us"). Being precise about which kind you mean clears a lot of debates.

## Consciousness (in machines)

**Whether a system can "experience" anything – mostly unsettled and contentious.**

Philosophy of mind asks if machines could be conscious, sentient, or merely simulating conversation without inner experience. Current LLMs show no evidence of subjective experience; they generate text about feelings without necessarily having them. Ethical treatment of AI systems may eventually hinge on this question – today it's largely theoretical.

## AGI (artificial general intelligence)

**Hypothetical AI that matches human breadth across most cognitive tasks.**

AGI is the idea of a system that generalizes across domains the way a skilled human can – novel problems, physical world, long horizons – not just excelling at benchmarks. No system today is agreed to be AGI. The term anchors timelines, investment, and existential-risk debates more than current product specs.

### WHERE PEOPLE ARGUE

Definitions of AGI vary enough that two people can argue past each other while agreeing on today's capabilities.

## Benchmark

**Standardized tests used to compare models – useful and gameable.**

Benchmarks are datasets and scoring rules (MMLU, HumanEval, etc.) that measure model performance on defined tasks. They help compare releases but don't capture real-world messiness – user prompts, adversarial inputs, or your company's actual documents. A model that tops a leaderboard can still fail your workflow.

### WHERE PEOPLE ARGUE

Benchmark chasing vs. real-world evaluation is a recurring fight in AI research and procurement.

## Warm-up

6 minutes.

Write on the board: 'The capital of Australia is Canberra.' Ask: 'Does a student who memorized this understand it? What would you ask them to find out?' Collect three test questions. Those tests are the

seminar's tools.

## Agenda - 60 minutes

| Minutes | Segment          | What happens                                                                                                                                                                                     |
|---------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 0-6     | Warm-up          | Memorized fact vs. understanding. Build three test questions.                                                                                                                                    |
| 6-18    | Read the text    | Background sheet + six vocabulary cards. Students mark: one claim they accept, one they doubt, one question.                                                                                     |
| 18-46   | Socratic seminar | Circle. Teacher opens, then withdraws. Teacher's only interventions: naming which of the three questions (behavioral / functional / phenomenal) a comment answers, and inviting silent students. |
| 46-54   | Turn             | 'If you're right about understanding, what changes about how you'd use the tool?' Five minutes; every student writes one sentence before anyone speaks.                                          |
| 54-60   | Written position | Exit ticket paragraph.                                                                                                                                                                           |

## How socratic seminar runs

Students sit in a circle with a short text or definition in hand. The teacher asks an opening question and then mostly stays quiet; students talk to each other, cite the text, build on and challenge one another. No hands – a speaker finishes, then the next speaks. The goal is understanding, not winning.

### GROUND RULES – POST THESE

1. Attack ideas, never people. "That claim has a hole" – not "you're wrong."
2. One voice at a time. The listener's job is to be able to repeat what was said.
3. Evidence beats volume. A quiet fact outranks a loud opinion.
4. Changing your mind is winning. Say so out loud when it happens.
5. No hands. Refer to the text before offering an opinion.
6. 'I don't know' is a complete and respected sentence.

## The two positions

Starter arguments are scaffolds, not scripts. Strong teams find their own. Assign sides where the format calls for it; students record their real view on the exit ticket.

### Perspective – yes, in every sense that matters

**Understanding is what understanding does. If it answers, explains, generalizes, and corrects itself, demanding more is demanding magic.**

- We attribute understanding to each other by behavior. We have nothing else.
- To predict the next word across all of human text, you have to model what the text is about.
- Interpretability research finds internal structures that track concepts, not just word statistics.

## Perspective – no, and the fluency is the trap

**Prediction is not comprehension. The system has no world, no stakes, no one home – and its fluency makes us forget that.**

- It fails at trivial things a child gets, which shows the 'understanding' is pattern-shaped, not concept-shaped.
- The Chinese Room: perfect symbol manipulation, zero meaning.
- Nothing is at stake for the model. Understanding without caring isn't understanding, it's lookup.

## Discussion questions

### OPENING

1. What would a machine have to do for you to say it understands?

### CORE

1. Are you making a claim about behavior, internal structure, or experience? Would the other two change your answer?
2. The model passed the bar exam and then invented a case that doesn't exist. Which fact matters more?
3. If we can't test for experience in a machine, can we test for it in each other?

### CLOSING

1. Suppose you're right. What changes about how much you trust the tool tomorrow?

## Roles

| Role          | Job                                                                                  |
|---------------|--------------------------------------------------------------------------------------|
| Level tracker | Names whether each claim is behavioral, functional, or phenomenal.                   |
| Text citer    | Brings the conversation back to a definition or the background sheet when it floats. |
| Bridge        | Invites a silent student in with a direct, answerable question.                      |
| Turn keeper   | Runs the final 'what changes' round: one sentence each, no responses.                |

## Assessment rubric

Share this before the debate. Judges and self-assessors use the same sheet. 'Meets' is the target; note above or below in the margin.

| Criterion | Meets                                                                    |
|-----------|--------------------------------------------------------------------------|
| Argument  | Claim, warrant, and evidence are explicit and connected.                 |
| Steelman  | Presents the strongest version of the opposing view before attacking it. |
| Clash     | Rebuttals respond to what was actually said, not a strawman.             |
| Precision | Uses lesson vocabulary accurately and distinguishes is from ought.       |

| Criterion        | Meets                                                                     |
|------------------|---------------------------------------------------------------------------|
| Written position | Closing paragraph takes a defensible position and concedes one real cost. |

## Exit ticket prompts

1. My position: the model \_\_\_ (understands / doesn't / the question is malformed) because \_\_\_\_.
2. The level I'm arguing at: behavioral / functional / phenomenal.
3. If I'm right, I should use the tool differently by \_\_\_\_.

## Extension

Give a model a task that requires understanding rather than recall – a novel riddle, a counting problem, a made-up rule to follow. Record where it succeeds and fails. Which level of understanding does the result speak to?

### SEND HOME

Ask at home: 'Is there anything you can do without understanding it? Anything you understand but can't do?'

### STANDARDS ALIGNMENT

- CCSS ELA SL.9–10.1 / SL.11–12.1 (seminar)
- CCSS ELA RI.11–12.7 (integrate and evaluate multiple sources)
- CCSS ELA W.11–12.1 (argument)
- AP Seminar / philosophy elective alignment: argument analysis

More lessons by grade band, and the full AI concepts field guide, at [hibot.space/learn/debates](https://hibot.space/learn/debates). Tell us how it went: [hello@hibot.space](mailto:hello@hibot.space).

# Can a machine understand anything?

**"A large language model understands the text it produces."**

## WHAT YOU NEED TO KNOW

A language model is trained to predict the next token. That objective, at scale, yields systems that pass professional exams, write working code, and explain their own answers – and also fail at simple counting, invent citations, and contradict themselves. 'Stochastic parrot' (it recombines patterns without meaning) and 'emergent reasoning' (the patterns encode reasoning) are the two poles; most researchers are somewhere between.

Three different questions hide in 'does it understand': Behavioral – does it act as if it understands? (Often, yes.) Functional – does it have internal representations that track meaning? (Interpretability research suggests some do, partially.) Phenomenal – is there something it is like to be the model? (No evidence, and no agreed test.) The seminar goes better when students say which one they're arguing.

The Chinese Room (Searle, 1980) argues that symbol manipulation isn't understanding no matter how good the output. The Turing test argues that behavior is all we ever have, even for each other. Both arguments are older than the technology and both still bite. Students do not need to resolve them – they need to see why their answer might change how much they trust the tool.

## WORDS WE'LL USE

Large language model (LLM) – The kind of model behind most chatbots – a very advanced next-word guesser.

Stochastic / nondeterministic – The same prompt can produce different answers – randomness is built in.

Intelligence (what we mean) – Performance on tasks vs. something deeper – the word hides disagreement.

Consciousness (in machines) – Whether a system can "experience" anything – mostly unsettled and contentious.

AGI (artificial general intelligence) – Hypothetical AI that matches human breadth across most cognitive tasks.

Benchmark – Standardized tests used to compare models – useful and gameable.

## MARK IT UP

Underline one sentence you agree with. Circle one you doubt. Write one question in the margin. Bring all three to the discussion.

# Position organizer

Proposition: "A large language model understands the text it produces."

Name: \_\_\_\_\_ Side (assigned or chosen): \_\_\_\_\_

## OUR CLAIM, IN ONE SENTENCE

---

---

## THREE REASONS – AND THE EVIDENCE FOR EACH

### Reason 1:

---

Evidence (a fact, a definition, or a specific example):

---

---

### Reason 2:

---

Evidence (a fact, a definition, or a specific example):

---

---

### Reason 3:

---

Evidence (a fact, a definition, or a specific example):

---

---

## WHAT THE OTHER SIDE WILL SAY – AND OUR REPLY

### Their best argument:

---

---

### Our reply:

---

---

## VOCABULARY I WILL USE CORRECTLY

☐ Large language model (LLM) ☐ Stochastic / nondeterministic ☐ Intelligence (what we mean) ☐  
Consciousness (in machines) ☐ AGI (artificial general intelligence) ☐ Benchmark

# Sentence starters

Use these until you don't need them. The goal is to respond to what was actually said.

- The text says \_\_\_\_, which I read as a claim about \_\_\_\_ (behavior / structure / experience).
- I'd build on \_\_\_\_ by adding \_\_\_\_.
- That's a behavioral test. It doesn't settle \_\_\_\_.
- Can we distinguish \_\_\_\_ from \_\_\_\_? What would tell them apart?
- I'm not sure, and here's what would change my mind: \_\_\_\_.

## GROUND RULES

1. Attack ideas, never people. "That claim has a hole" – not "you're wrong."
2. One voice at a time. The listener's job is to be able to repeat what was said.
3. Evidence beats volume. A quiet fact outranks a loud opinion.
4. Changing your mind is winning. Say so out loud when it happens.
5. No hands. Refer to the text before offering an opinion.
6. 'I don't know' is a complete and respected sentence.

## HOW SOCRATIC SEMINAR WORKS

Students sit in a circle with a short text or definition in hand. The teacher asks an opening question and then mostly stays quiet; students talk to each other, cite the text, build on and challenge one another. No hands – a speaker finishes, then the next speaks. The goal is understanding, not winning.

## YOUR ROLE TODAY

Role: \_\_\_\_\_

| Role          | Job                                                                                  |
|---------------|--------------------------------------------------------------------------------------|
| Level tracker | Names whether each claim is behavioral, functional, or phenomenal.                   |
| Text citer    | Brings the conversation back to a definition or the background sheet when it floats. |
| Bridge        | Invites a silent student in with a direct, answerable question.                      |
| Turn keeper   | Runs the final 'what changes' round: one sentence each, no responses.                |

## LISTENING TALLY

Keep count while others speak. Claims: \_\_\_\_ Evidence: \_\_\_\_ Questions: \_\_\_\_ Vocabulary used: \_\_\_\_  
Minds changed: \_\_\_\_

# Exit ticket

"A large language model understands the text it produces."

Name: \_\_\_\_\_ Date: \_\_\_\_\_

**My position: the model \_\_\_ (understands / doesn't / the question is malformed) because \_\_\_.**

---



---



---

**The level I'm arguing at: behavioral / functional / phenomenal.**

---



---



---

**If I'm right, I should use the tool differently by \_\_\_.**

---



---



---

## SELF-ASSESSMENT

| Criterion        | Meets looks like                                                          | Me today (above / meets / below) |
|------------------|---------------------------------------------------------------------------|----------------------------------|
| Argument         | Claim, warrant, and evidence are explicit and connected.                  |                                  |
| Steelman         | Presents the strongest version of the opposing view before attacking it.  |                                  |
| Clash            | Rebuttals respond to what was actually said, not a strawman.              |                                  |
| Precision        | Uses lesson vocabulary accurately and distinguishes is from ought.        |                                  |
| Written position | Closing paragraph takes a defensible position and concedes one real cost. |                                  |

## TAKE IT HOME

Ask at home: 'Is there anything you can do without understanding it? Anything you understand but can't do?'