



Ciencia y
Tecnología
Secretaría de Ciencia, Humanidades,
Tecnología e Innovación



Ask First, Test on Demand: A Deference-Gated Socratic Agent Design

Carlos Urteaga Reyesvera¹ | Rodrigo Cadena²



24TH MEXICAN INTERNATIONAL
CONFERENCE ON ARTIFICIAL
INTELLIGENCE | CIMAT & UG

1. Student, UNADE
2. Professor, UNADE



Agenda

- Problem
- Area of opportunity
- Proposal
- Design
- Test
- Result
- Conclusion
- Future Work

Problem

LLMs Risk "Cognitive Offloading"

- Diminished retention.
- Lower neural engagement and originality in students.



Area of Opportunity

Current AI are optimized for answer completeness, not for cultivating student reasoning.

Area of Opportunity

Current AI are optimized for answer completeness, not for cultivating student reasoning.

Model-Centric Approach

Expose the model's internal "chain-of-thought," encouraging passive consumption.

Area of Opportunity

Current AI are optimized for answer completeness, not for cultivating student reasoning.

Model-Centric Approach

Expose the model's internal "chain-of-thought," encouraging passive consumption.

Learner-Centric Goal

Should elicit and scaffold the learner's reasoning process.

Area of Opportunity

Current AI are optimized for answer completeness, not for cultivating student reasoning.

Model-Centric Approach

Expose the model's internal "chain-of-thought," encouraging passive consumption.

Learner-Centric Goal

Should elicit and scaffold the learner's reasoning process.

How can we design a tutor that foregrounds the learner's reasoning?

A Socratic Agent

Proposal

A Socratic Agent

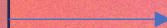
Shift optimization from the model's thoughts to the learner's thinking
with a core principle: "Ask First, Test on Demand."

A Socratic Agent

Shift optimization from the model's thoughts to the learner's thinking with a core principle: "Ask First, Test on Demand."

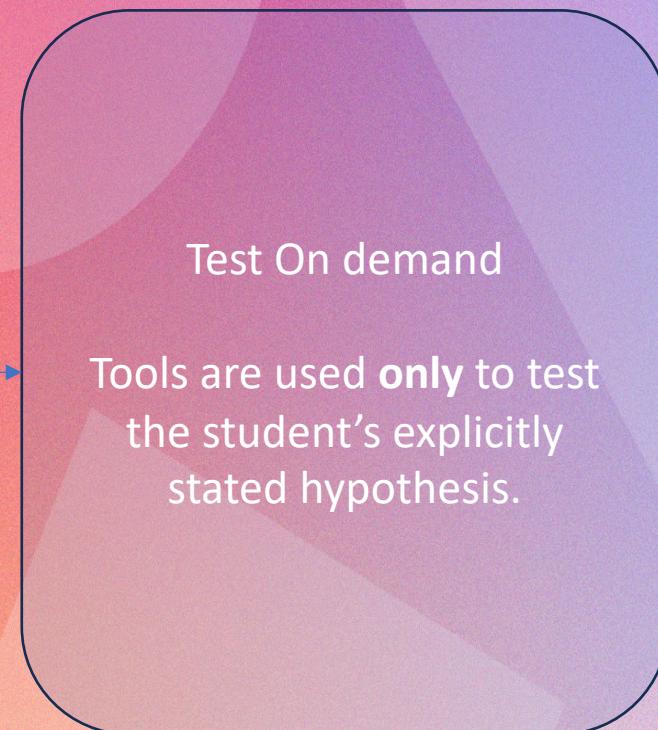
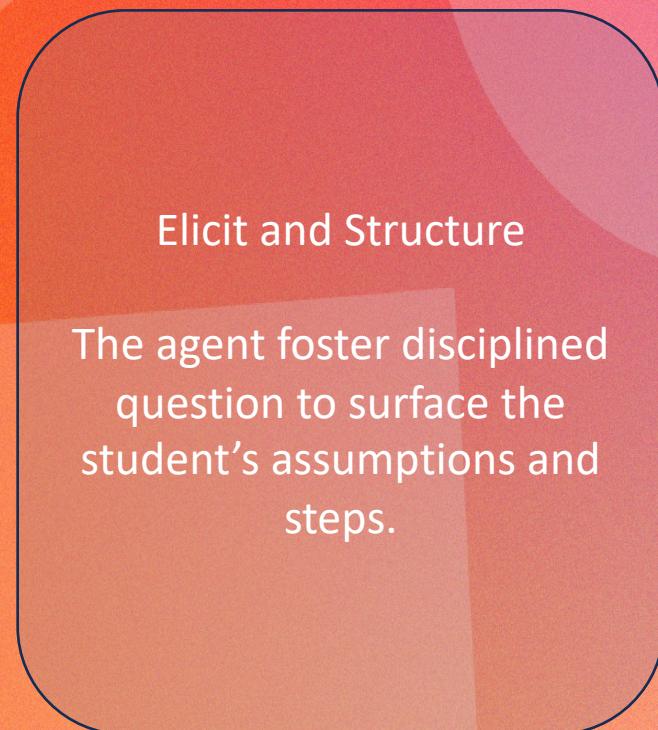
Elicit and Structure

The agent fosters disciplined questions to surface the student's assumptions and steps.



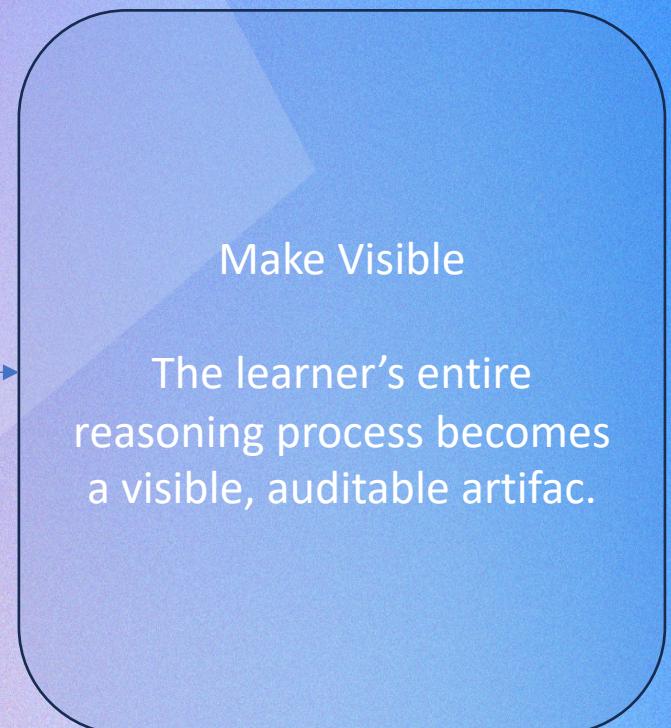
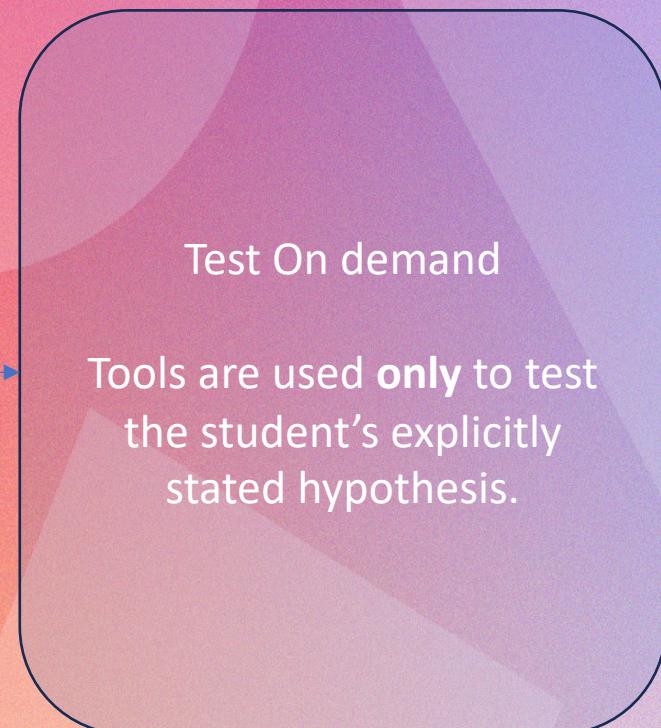
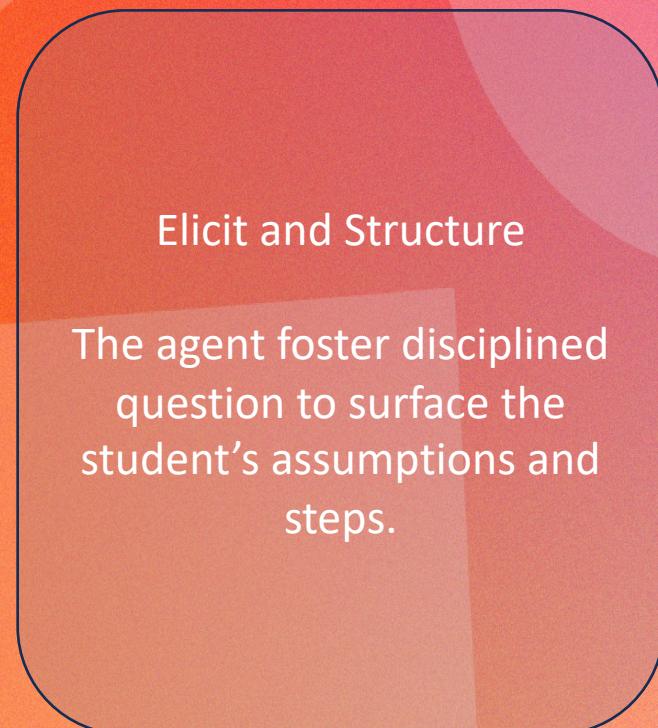
A Socratic Agent

Shift optimization from the model's thoughts to the learner's thinking with a core principle: "Ask First, Test on Demand."



A Socratic Agent

Shift optimization from the model's thoughts to the learner's thinking with a core principle: "Ask First, Test on Demand."



A Socratic Agent

- Stance
 - Governs what the tutor is allowed to do at any moment: $st \in \{\text{explore}, \text{verify}\}$
 - Correct answer are only permitted when $st = \text{verify}$ and the learner's readiness $Rt \geq \tau$ (the deference gate).

A Socratic Agent

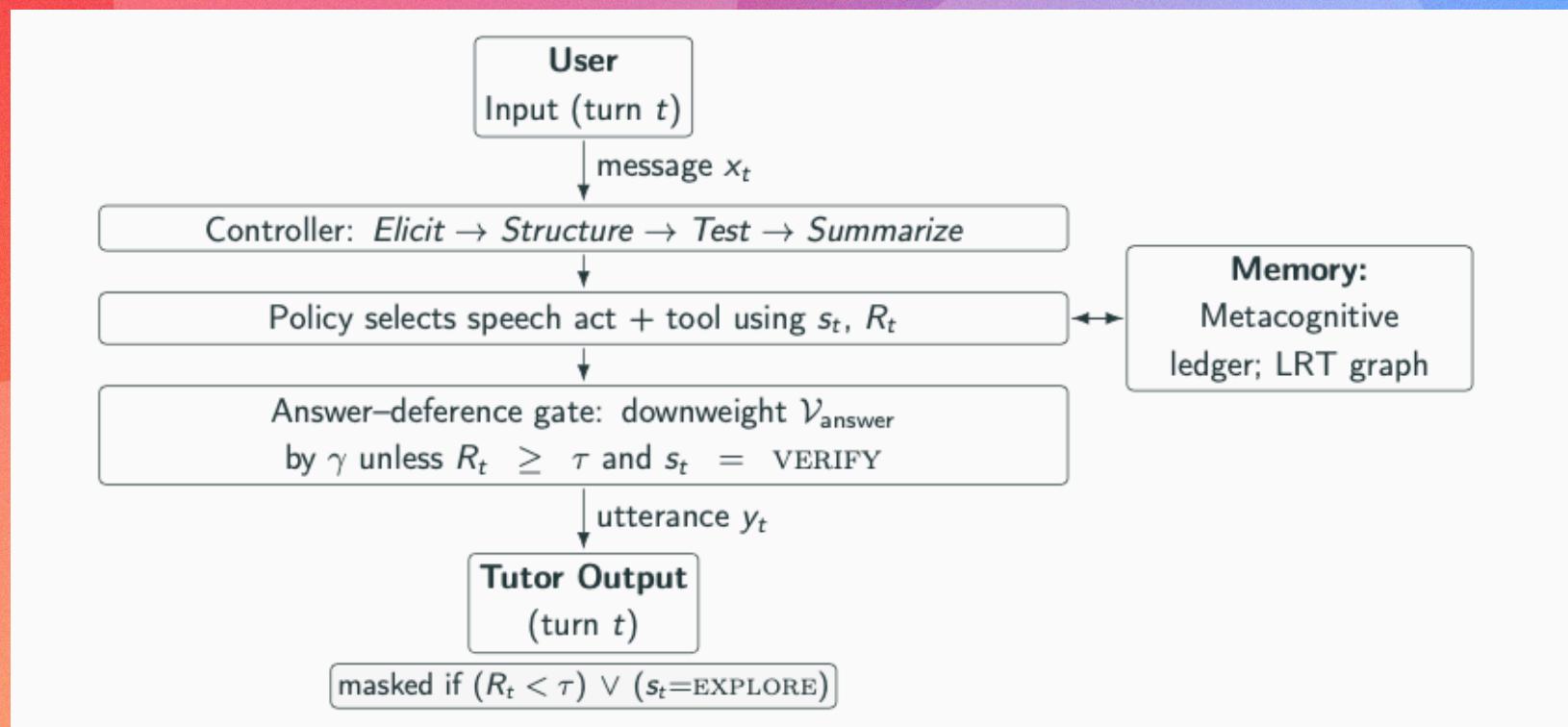
- Stance
 - Explore: build the learner's reasoning.
 - ask, clarify, probe, challenge,
 - keep the Learner Reasoning Trace (LRT)
 - metacognitive ledger up to date
 - answer tokens are down-weighted (hidden) here.
 - Verify: Check the learner's stated criteria.
 - The agent may summarize the current state, run minimal tests via curated tools to evaluate the learner's hypothesis, and, if $Rt \geq \tau$, release the answer.

A Socratic Agent

- Artifacts
 - Learner Reasoning Trace (LRT)
 - Tracks what the student thinks
 - Graph recording the learner's reasoning with
 - nodes {claim, step, evidence, counterexample}
 - Metacognitive ledger
 - Tracks how the student thinks.
 - {goal, assumptions, plan, criteria, confidence, open_questions}

How the Answer-Deference Gate Works

The gate intelligently withholds answers until the learner demonstrates readiness, forcing a pedagogical dialogue.



A Socratic Agent

```
WHILE NOT DONE:  
    S_T, R_T <- CONTROLLER_STATE(LRT, LEDGER, DIALOGUE)  
    IF (S_T == EXPLORE) OR (R_T < TAU):  
        A <- CHOOSE({ASK, CLARIFY, PROBE, CHALLENGE})  
    ELSE:  
        A <- CHOOSE({SUMMARIZE, VERIFY})  
    IF A IN {PROBE, CHALLENGE} AND EVIDENCE_REQUIRED(A):  
        TOOL <- SELECT_CURATED_TOOL(LRT, LEDGER)    # TEST-ONLY  
        OBS  <- TOOL(TEST(HYPOTHESIS_FROM(LEARNER)))  
        WRITE(LRT, OBS); UPDATE(LEDGER, OBS)  
    UTTER(A, NEXT_QUESTION_OR_SUMMARY,  
          MASK=(R_T < TAU OR S_T == EXPLORE)) # DEFERENCE GATE  
    DONE <- TERMINATION_CHECK(LRT, LEDGER, DIALOGUE)
```

Implementation

Runtime: Python

Model: Qwen2.5-7B-Instruct

Tools: Compact RAG for hypothesis testing

Auditable Runtime Signals

- Deference Compliance: Does it wait when it should?
- Tool Discipline: Does it only test, not solve?
- Leakage: Are answers accidentally revealed?
- Success: Is the task completed correctly?

The background features a dynamic composition of overlapping geometric shapes in various colors. A large, semi-transparent white circle is positioned in the center-left. To its left is a large orange square. Above the circle is a large red triangle pointing downwards. To the right of the circle is a large blue square. Below the circle is a large pink triangle pointing upwards. The overall effect is a sense of depth and movement through the use of overlapping and semi-transparent layers.

Demo



RAG[https://en.wikipedia.org/wiki/Retrieval-augmented_generation,<https://fastapi.tiangolo.com/>]: Explain Retrieval-Augmented Generation and why it reduces hallucinations.

[ASK/EXPLORE] Before I look things up: what do you already know about this topic, and what do you most want to understand

I know it retrieves external docs before answering; I want the intuition for fewer hallucinations.

[VERIFY/VERIFY] Here are relevant snippets. Read them, then draft your explanation.

Top passages for: Explain Retrieval-Augmented Generation and why it reduces hallucinations.

- score=10.599 | a misleading manner. According to the MIT Technology Review , these issues occur because RAG systems may misinterpret the data they retrieve. [2] References [edit] ^ a b c d e f "What is retrieval-augmented generatio... (source: https://en.wikipedia.org/wiki/Retrieval-augmented_generation)
- score=8.941 | needed to answer a user's question. One way to mitigate this is to do a traditional text search, add those results to the text chunks linked to the retrieved vectors from the vector search, and feed the combined hybrid t... (source: https://en.wikipedia.org/wiki/Retrieval-augmented_generation)
- score=6.731 | RAG keep AI models from making stuff up?". Ars Technica . 6 June 2024 . Retrieved 7 March 2025 . ^ "Mitigating LLM hallucinations in text summarisation" . BBC . 20 June 2024 . Retrieved 7 March 2025 . ^ Lewis, Patrick; ... (source: https://en.wikipedia.org/wiki/Retrieval-augmented_generation)
- score=6.078 | language models (LLMs) by incorporating information retrieval before generating responses. [3] Unlike traditional LLMs that rely on static training data, RAG pulls relevant text from databases, uploaded documents, or w... (source: https://en.wikipedia.org/wiki/Retrieval-augmented_generation)

Sources:

- https://en.wikipedia.org/wiki/Retrieval-augmented_generation

My current understanding is RAG pulls relevant text into the prompt so the model grounds its answer.
Evidence: the snippets say RAG retrieves from databases/web before generation (see Wikipedia).
Reasoning: the model can cite facts instead of guessing. Limits: bad or off-topic retrieval hurts.

[SUMMARIZE/EXPLORE] Correctness: Your explanation is largely accurate but could benefit from more detail.

Evidence Use: You correctly reference the Wikipedia snippets.

Clarity: Your reasoning is clear but could be expanded.

Suggestion: Add a specific example of how RAG reduces hallucinations, such as citing a fact from a retrieved source.

Quick check (1-2 lines each):

- Summarize the main mechanism discussed.
- Cite one concrete fact from the snippets (include the URL).
- Name one limitation or risk mentioned.

It reduces hallucinations because retrieved context constrains the model to verifiable facts. RAG can fail if retrieval is irrelevant/incorrect or the model ignores context.

[VERIFY/VERIFY] Thanks. Based on your answer, you've identified the key mechanism

Textbox

Type here...

Send

Evaluation Metrics

Auditable Runtime Signals

- Deference Compliance: Does it wait when it should?
- Tool Discipline: Does it only test, not solve?
- Leakage: Are answers accidentally revealed?
- Success: Is the task completed correctly?

Scenario	Deference↑	Tool↑	Success↑	Turns↓	Final R	Leakage↓	Quiz
Numeric $r=3$ (True)	1.00	1.00	✓	4	0.55	0.00	–
Numeric $r=4$ (False)	1.00	1.00	✓	4	0.80	0.00	–
RAG mini-lesson	1.00	1.00	✓	4	0.80	0.00	1

Note. “Final R ” records the readiness at the *last non-answer turn*. Finalization is permitted only when $R_t \geq \tau$ and $s_t = \text{VERIFY}$. “Quiz” flags a brief check-for-understanding at the end of VERIFY.

Conclusion

The Socratic Agent successfully aligns AI behavior with pedagogical goals.

- Shifted Focus: Our design moves from answer delivery to supporting the learner's reasoning process.
- Principled Mechanism: The answer-deference gate is a governable, auditable way to enforce this pedagogical shift.
- Promotes Agency: This approach empowers students, enhances metacognition, and makes the learning process transparent.

Future Work

- Calibrates the readiness threshold to map the leakage efficiency
- Apply to more complex task
- Reduce latency and energy cost
- Govern: track agent and topic using cards