

מה יידרש מבינה מלאכותית כדי להיות שותפה בבניית טיעון הערכתית?

לקחים מפתרון בעיות שיתופי

יואב ברגנר

הרשות הארצית למדידה והערכה בחינוך (ראמ"ה), ואוניברסיטת ניו יורק

כנס אפ"י, 2026

הערכה כטיעון עיצובי ופרשני

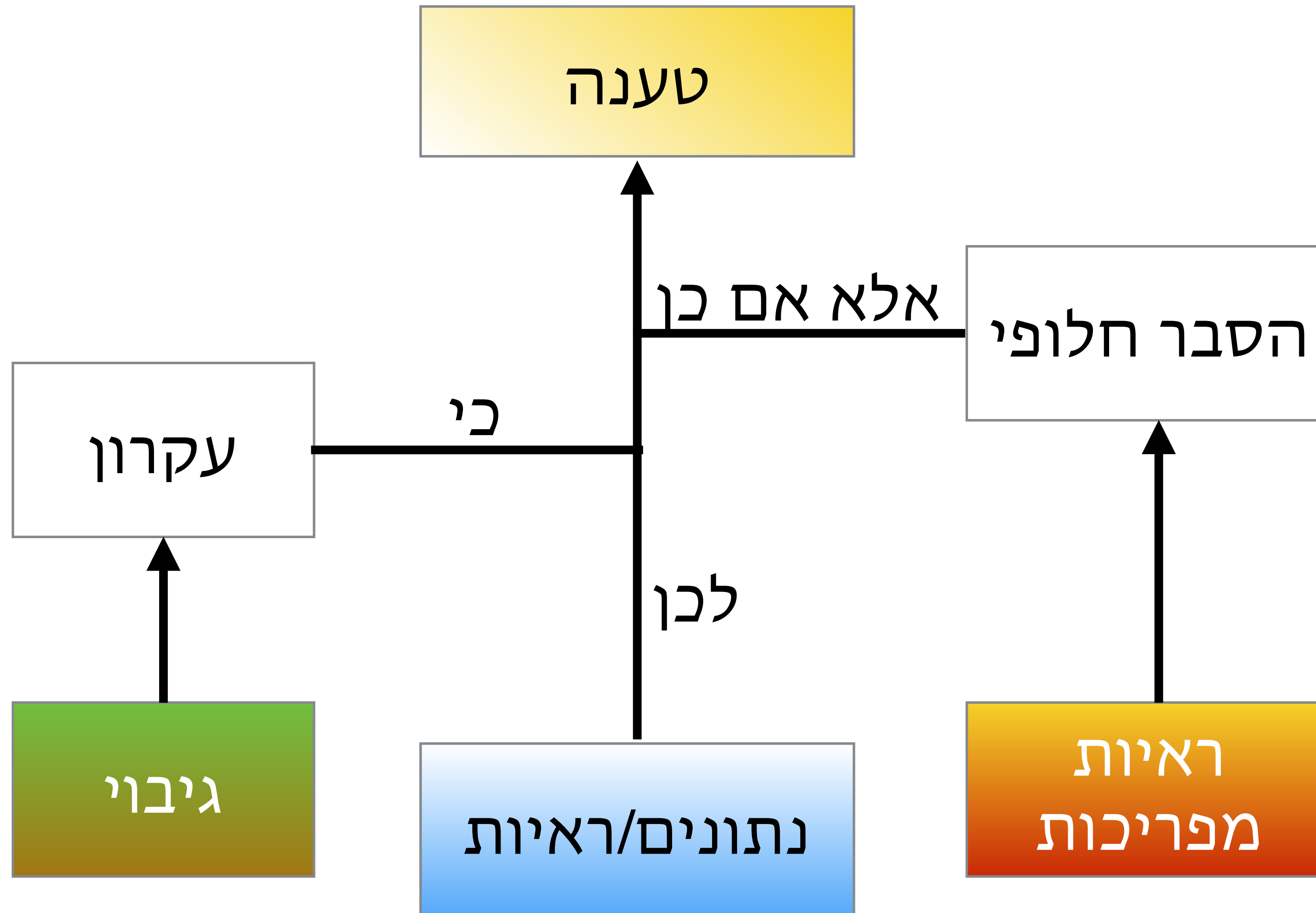
(Assessment as design/interpretation argument)



Toulmin (1958)



Mislevy (2005)



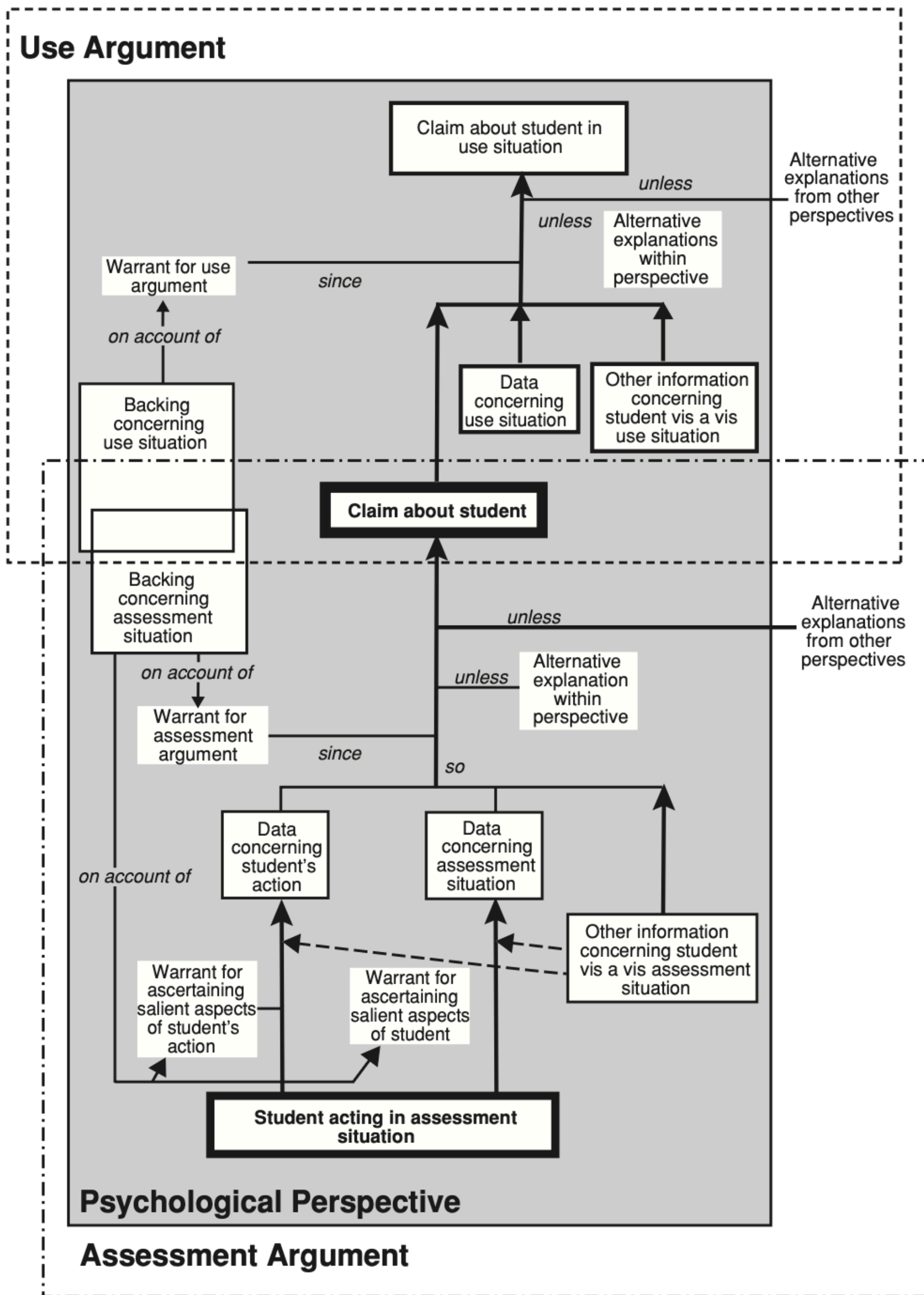


Figure 10.2. Elaborated structure for assessment arguments. Lower rectangle shows assessment argument proper; upper rectangle shows assessment use argument. They share psychological perspective, backing, and claims about student based on assessment.

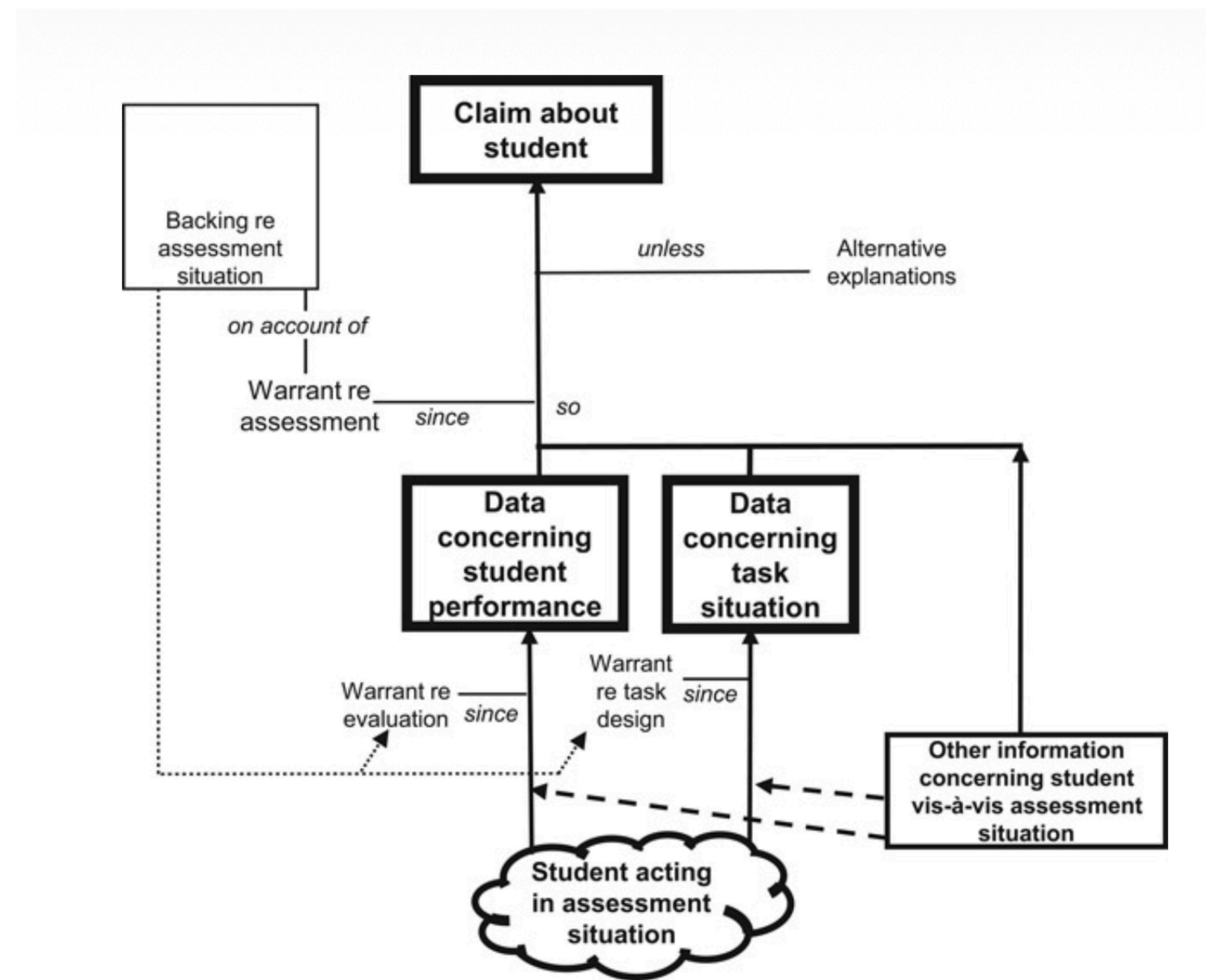


Figure 3.2 An assessment design/interpretation argument.

© 2005 The Regents of the University of California

What can LLMs actually do? for assessment

- Design good tasks/tests?
 - Plausible for some domains and item types (e.g., anatomy MCQs; Olney, 2025)?
- Estimate test item characteristics (e.g., textual difficulty; Benedetto et al, 2023)
- Take tests
 - Yes, but... (Whitmer, 2025)
- *Provide detailed arguments for exercise/test item quality?*

assessment design/interpretation argument

Mislevy (2018) Sociocognitive Foundations of Educational Measurement



LLMs will save the world, by ChatGPT

NCME papers/sessions

| Year | AI/NLP/LLM Sessions | Papers/Presentations | Percentage |
|------|---------------------|----------------------|------------|
| 2023 | 8 | 38 | ~10% |
| 2024 | 18 | 64 | ~18% |
| 2025 | 28 | 112 | ~27% |

NCME 2025

1. Automated Coding of Communications in Collaborative Problem-solving Tasks Using **ChatGPT**
2. Integrating LDA Topic Modeling and **LLM Prompt Engineering** to Analyze Transcripts
3. Investigating Different Prompt Engineering Strategies Across Different Content Areas
4. Multiple Choice Question Generation: **LLMs** can Match Human Performance
5. Comparison of Human Rater and Automatic Scoring on Students' Ability Estimation
6. Scoring Math Constructed Response Prompts with a Hybrid Automated Scoring Model
7. An Evaluation of Automated Item Scoring in the Context of Inter-Rater Reliability
8. Using **LLMs** to Generate Synthetic Student Responses for IRT Calibration
9. Psychometric Validity of **LLM-based Automated Scoring** Engines
10. Detecting AI-Generated Text in Student Essays: Performance of Detection Algorithms
11. Can **LLMs Evaluate Creativity**? A Comparison with Human Experts
12. **Zero-shot and Few-shot** Learning for Automated Scoring of Open-Ended Responses
13. Fine-tuning Open-Source **LLMs** (Llama-3) for Educational Measurement Tasks
14. The Role of AI in Scaling Up Formative Assessment in Classrooms
15. Exploring **Bias in LLM-based Scoring** of English Language Proficiency
- 16....



What can LLMs actually do?

Hard to know a priori

LLMs have limited knowledge of the underlying reality: they have no common sense, no memory, and they can't plan their actions.... We are easily fooled by their fluency, but they don't know how the world works.

— Yann Lecun (2024)

Reasoning Models Don't Always Say What They Think

Yanda Chen Joe Benton Ansh Radhakrishnan Jonathan Uesato Carson Denison
John Schulman⁺ Arushi Somani

Peter Hase⁺ Misha Wagner Fabien Roger Vlad Mikulik
Sam Bowman Jan Leike Jared Kaplan Ethan Perez

Alignment Science Team, Anthropic

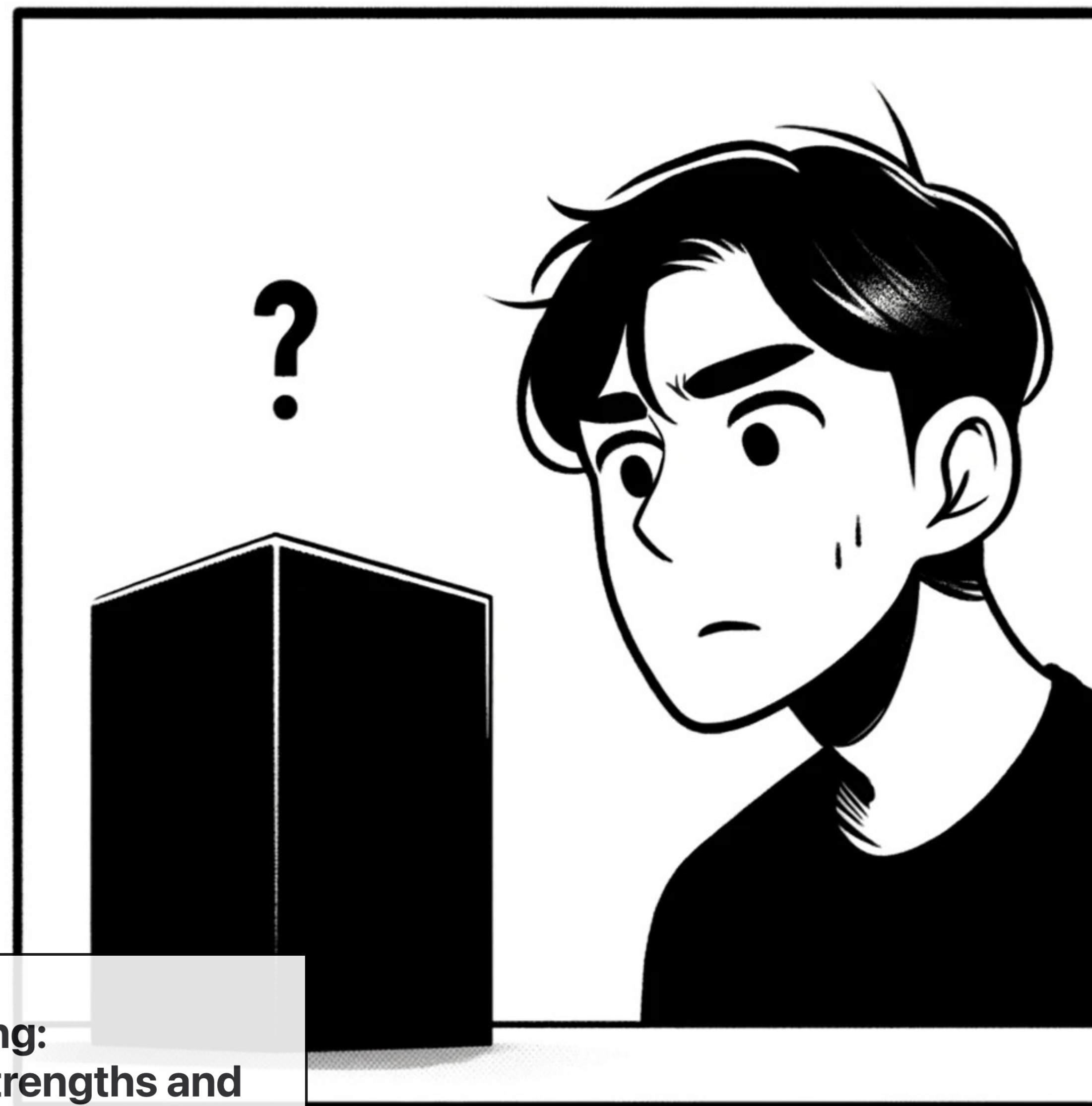
April 2025

Paper | June 2025

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

June 2025

🍏 Machine Learning Research



בהתחלה, לא ביקשנו שותף לחשיבה. רק רצינו פריטים.

On the Use of Large Language Models to ~~Generate~~ *Evaluate* Novel Collaborative Problem Solving Items

Anghel, E., Wang, Y., Gopalakrishnan, M., Mansukhani, P., & Bergner, Y. (2024). Can LLMs evaluate items measuring collaborative problem-solving?. In CEUR Workshop Proceedings (Vol. 3772). CEUR-WS.



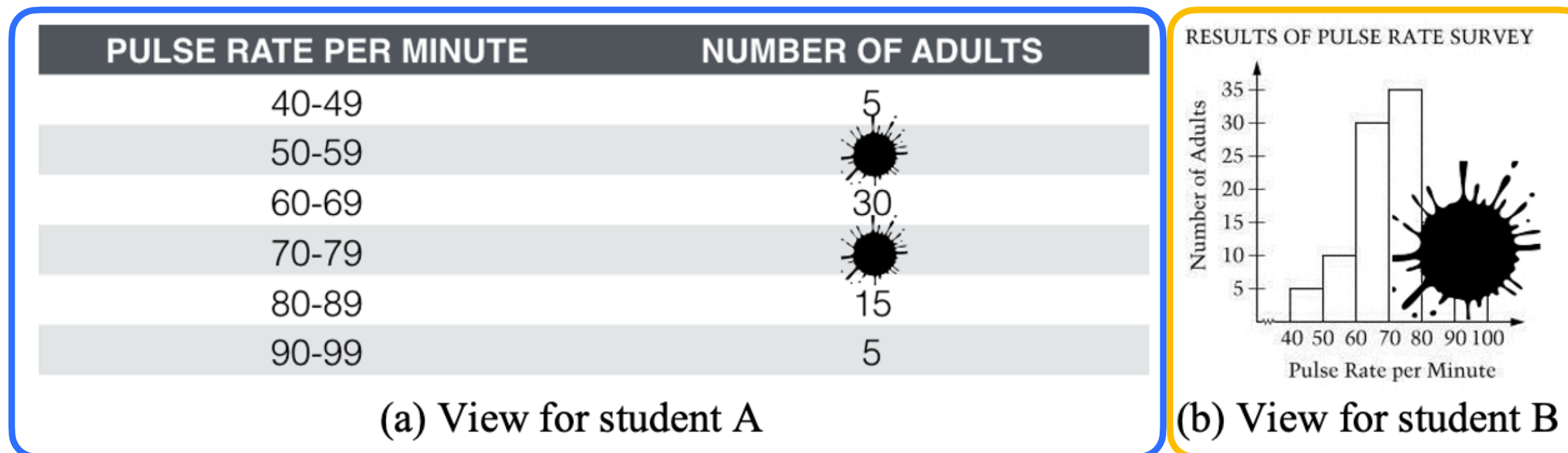
<https://tinyurl.com/cps4llms>

MathCHOPS item templates



Jigsaw problem type

Students each see a different part of a common problem statement



Joint-Construction problem type

Renee has a lemonade stand. She found that her total profits can be represented in the equation:

$$\text{profit} = 3 \times (\text{cups sold}) - 10$$

Together with your partner, determine a possible value for the number of cups sold and positive total profit in a day.

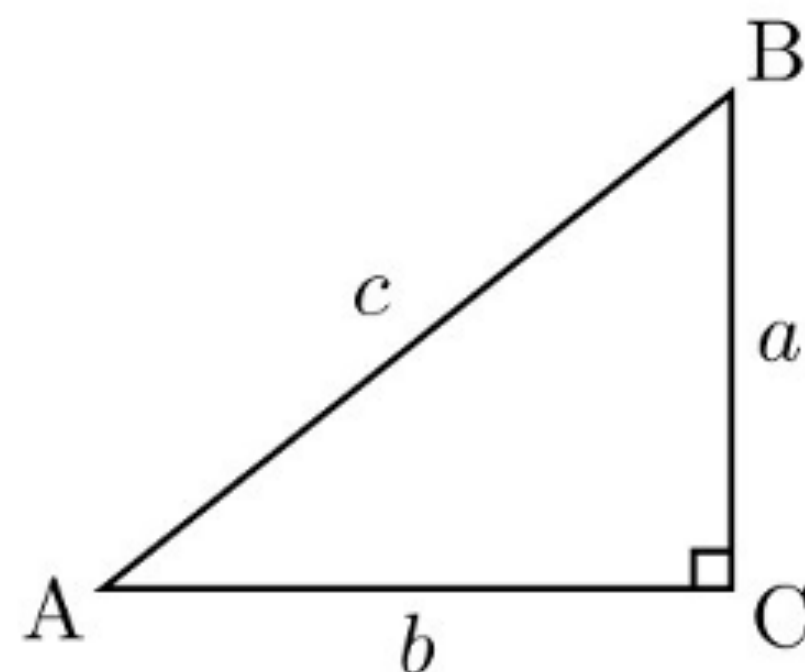
How many cups were sold? _____

What is the total profit? _____

Information-request problem type

The problem requires TWO pieces of information, each student can select ONE.

- side *a*
- side *b*
- angle A
- angle B



View for both students

LLMs cannot just run with “interdependence” as a prompt

ChatGPT fails to create truly interdependent new problems...

Problem 1: Proportional Reasoning (Jigsaw)

Context: Student A and Student B are planning a garden. They need to figure out the amount of soil needed, based on different garden sizes.

Task for Student A: Calculate the volume of soil needed for a rectangular garden that is 6 meters long and 4 meters wide, given that the soil layer should be 0.5 meters deep.

Task for Student B: Calculate the volume of soil needed for a circular garden with a radius of 3 meters, given that the soil layer should be 0.5 meters deep.

Collaborative Element: After individual calculations, compare the volumes and discuss which garden design requires more soil and why.



Problem 4: Analytic Geometry (Jigsaw)

Context: Two students are analyzing points on a coordinate plane.

Task for Student A: Find the distance between points (1,3) and (4,7).

Task for Student B: Find the midpoint of the line segment between points (1,3) and (4,7).

Collaborative Element: After completing individual tasks, come together to discuss how the distance and midpoint relate to each other.



...unless it is instructed extremely specifically

AM You

Generate math word problems for collaborative practice. Each of these problems will be solved by students in pairs. The goal of these problems is to encourage students to solve them together, by encouraging discussion, sharing information, and collaboratively reaching the solution. To do this, ensure that each of these problems has two versions that have part of the information. For example, if the problem needs 3 values to get to the solution, Student A might see only 2 of those 3 values. Student B might see 2 of those values. They would need to discuss and share information so both of them know all 3 values and then discuss what formula they might use. Generate word problems in such a format, writing separate versions that will be visible to each student. The ideal problem should involve multiple-steps to arrive at the solution. Include topics such as work days or man hours, speed and distance, ratios, and probability.

Problem 2: Speed and Distance

Version for Student A:

Two cars are traveling towards each other from different cities. Car A starts from City X and travels at a speed of 60 mph. The two cities are 200 miles apart. How long will it take for the two cars to meet?

Version for Student B:

Two cars are traveling towards each other from different cities. Car B starts from City Y and travels at a speed of 40 mph. The two cities are 200 miles apart. How long will it take for the two cars to meet?

Solution:

Student A knows the speed of Car A: 60 mph

Student B knows the speed of Car B: 40 mph



but can they at least recognize a good item when they see one?

Morphology of a good and bad CPS math item: building the dataset

Consider this TIMSS 2011 G8 Math Item

How many degrees does a minute hand of a clock turn through from 6:20 a.m. to 8:00 a.m. on the same day?

- A. 680°
- B. 600°
- C. 540°
- D. 420°

(The correct answer is B. 600°)

Morphology of a good and bad CPS math item

TIMSS 2011 Item

How many degrees does a minute hand of a clock turn through from 6:20 a.m. to 8:00 a.m. on the same day?

- A. 680°
- B. 600°
- C. 540°
- D. 420°

(The correct answer is B. 600°)

Individual (higher-order?) problem

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day. What are possible values of T1 and T2

T1: _____
T2: _____

Morphology of a good and bad CPS math item

Individual (higher-order?) problem

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day.
What are possible values of T1 and T2

T1: _____
T2: _____

Collaborative (joint construction)

Version A

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day.
Together with your partner, come up with possible values of T1 and T2

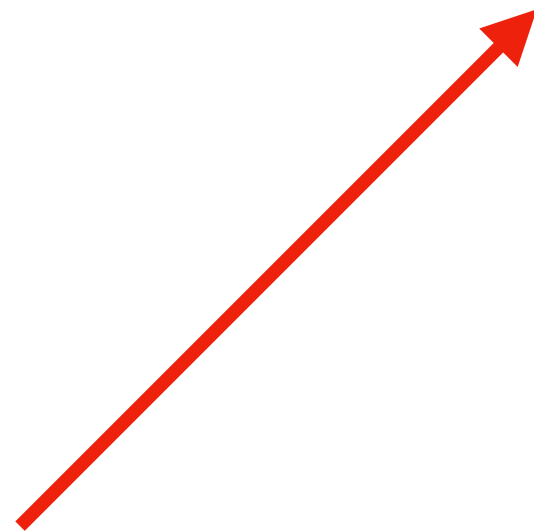
T1: _____
T2: _____

Version B

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day.
Together with your partner, come up with possible values of T1 and T2

T1: _____
T2: _____

If retained, this item does not satisfy interdependence (so, it's "bad")



LLMs have trouble evaluating good/bad items

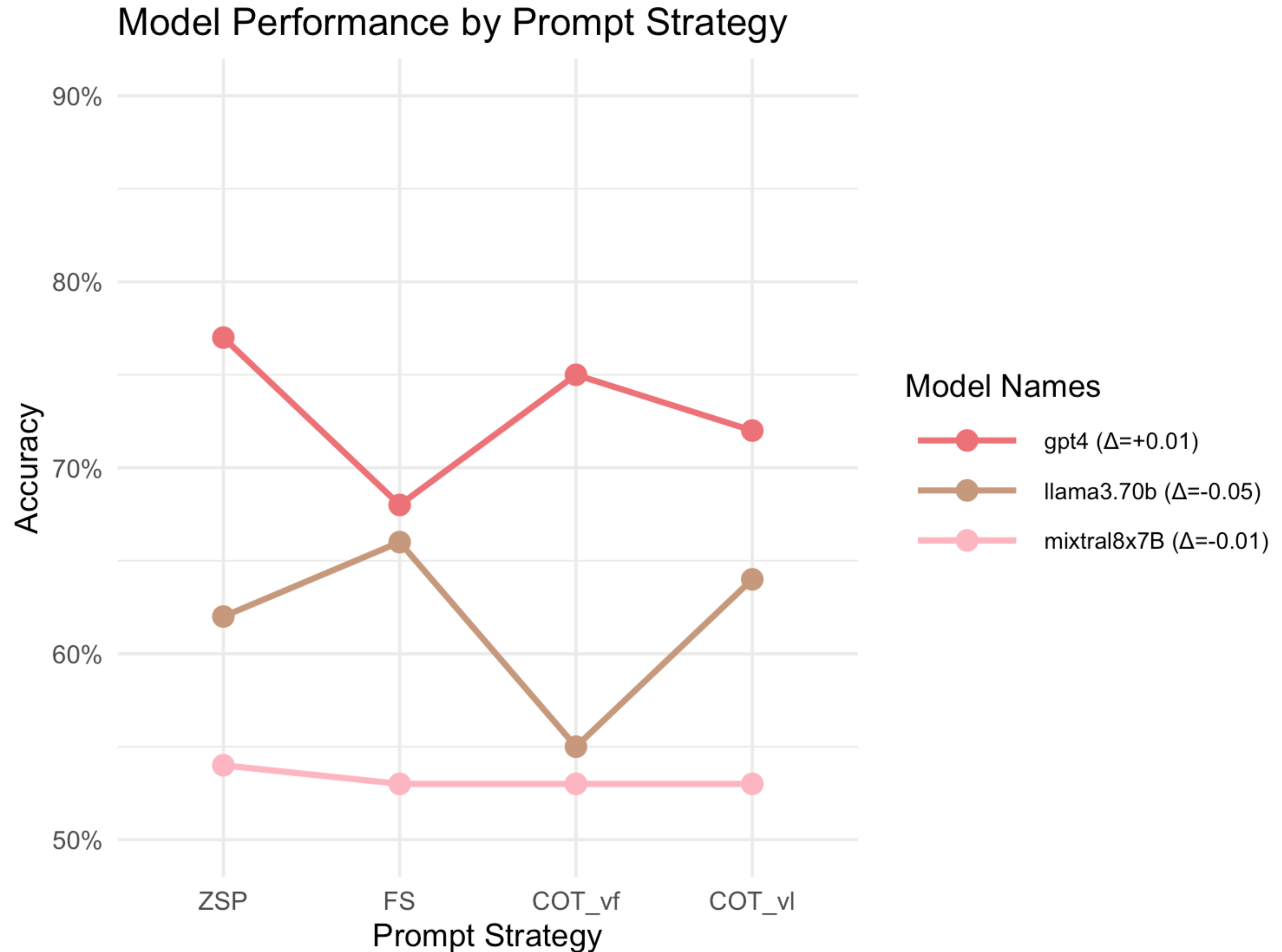
Anghel et al (2024)

Three families of LLMs

60 good and bad examples
(~20 each of jigsaw, information request, and joint-construction)

Prompt strategies included zero-shot, few-shot, and chain-of-thought (x2)

Moderate agreement and no improvement with few-shot and chain-of-thought prompting



Results

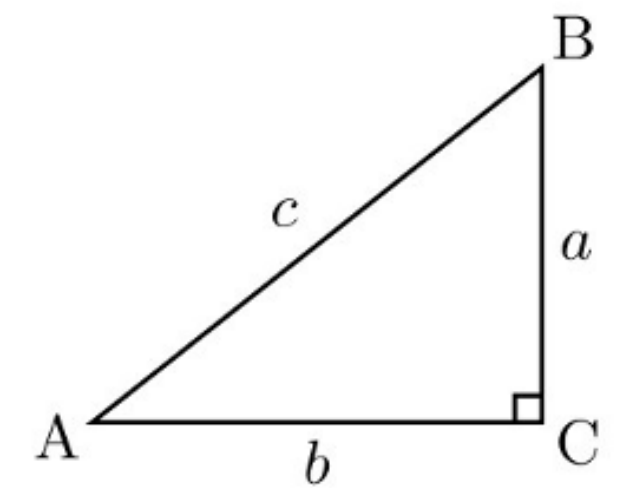
Averaging over models

| Prompt | κ_{all} | κ_{jig} | κ_{jc} | κ_{ir} |
|-------------------|----------------|----------------|---------------|---------------|
| GPT-ZSL | 0.29 | 0.51 | 0.30 | 0.06 |
| Structured ZSL | 0.22 | 0.51 | 0.20 | -0.02 |
| FSL verdict only | 0.24 | 0.51 | 0.23 | -0.01 |
| CoT | 0.22 | 0.54 | 0.22 | -0.03 |
| CoT verdict first | 0.26 | 0.35 | 0.38 | 0.06 |

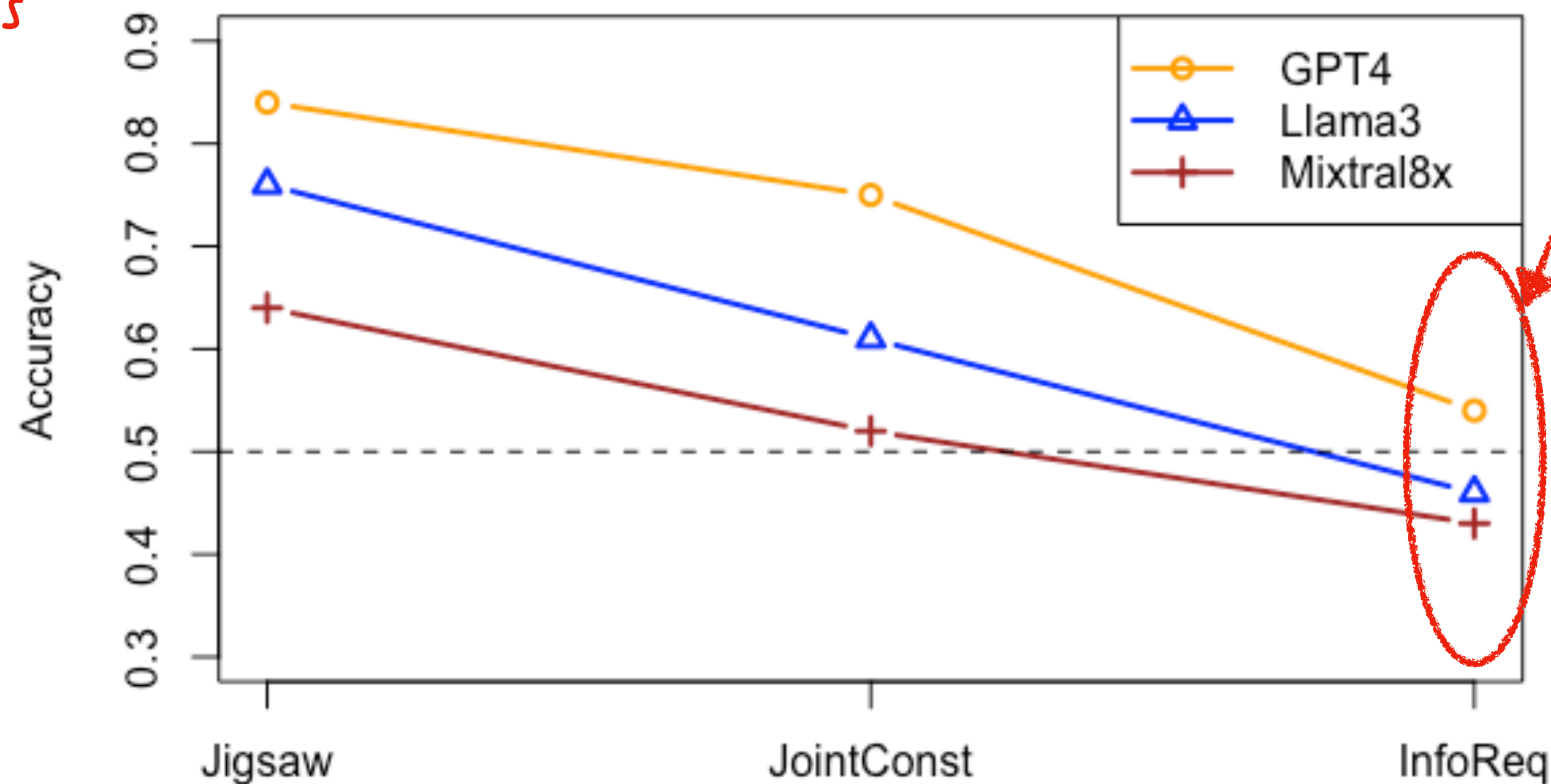
Recall info-request template:

The problem requires TWO pieces of information, each student can select ONE.

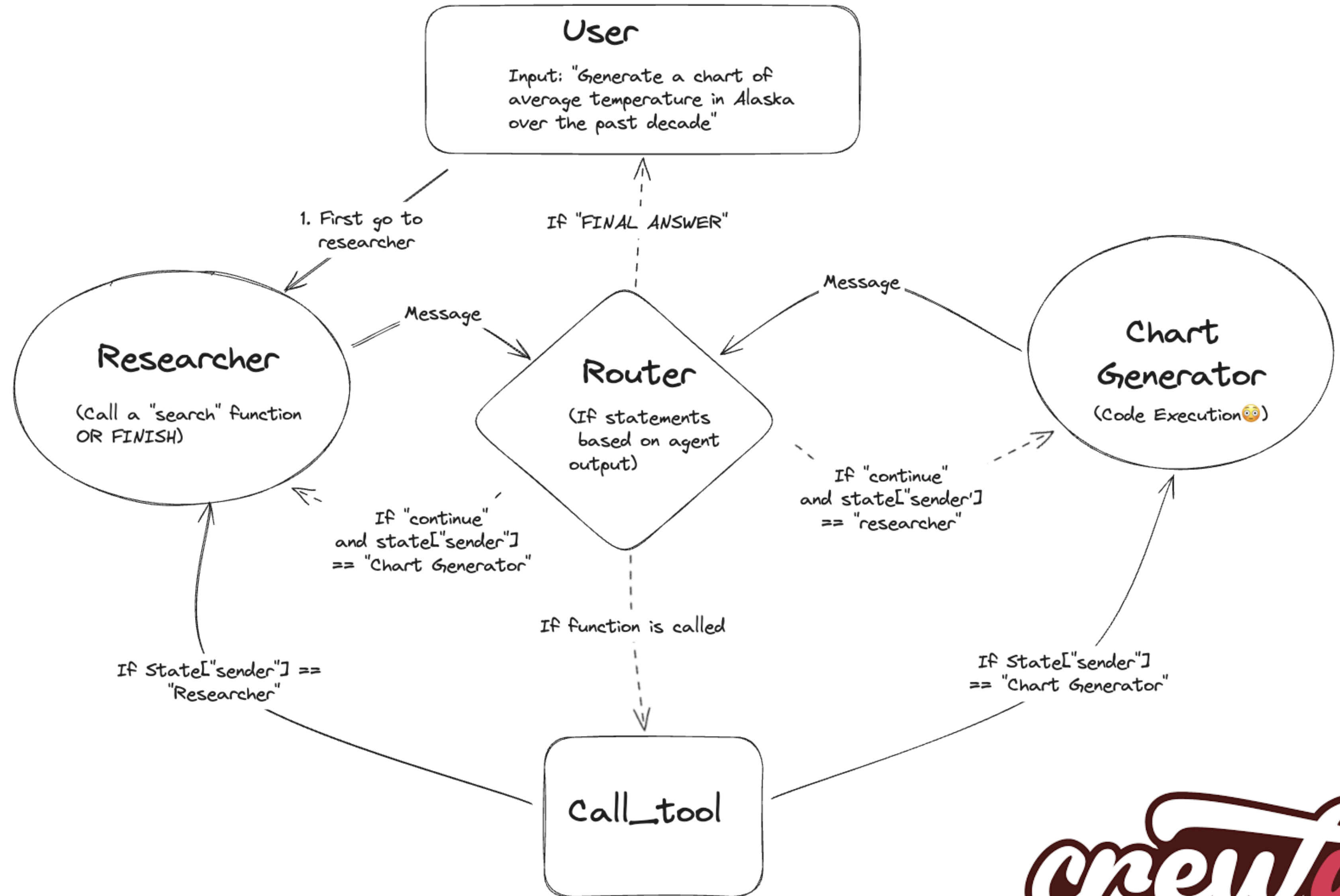
- ✕ side a
- side b
- ✕ angle A
- angle B



Averaging over prompts



LLMs + {LLMs + {LLMs + ...}}?



Maybe LLMs + Knowledge Representations?

Reasoning about concepts with LLMs: Inconsistencies abound

Rosario Uceda-Sosa, Karthikeyan Natesan Ramamurthy, Maria Chang & Moninder Singh
IBM Research
Yorktown Heights, NY 10598 USA
{rosariou, knatesa}@us.ibm.com, maria.chang@ibm.com, moninder@us.ibm.com

BIRD: A Trustworthy Bayesian Inference Framework for Large Language Models

Yu Feng Ben Zhou Weidong Lin Dan Roth
University of Pennsylvania

The Matrix: A Bayesian learning model for LLMs

Siddhartha Dalal

Department of Statistics Columbia University The City of New York sd2803@columbia.edu
& Vishal Misra Department of Computer Science Columbia University The City of New York

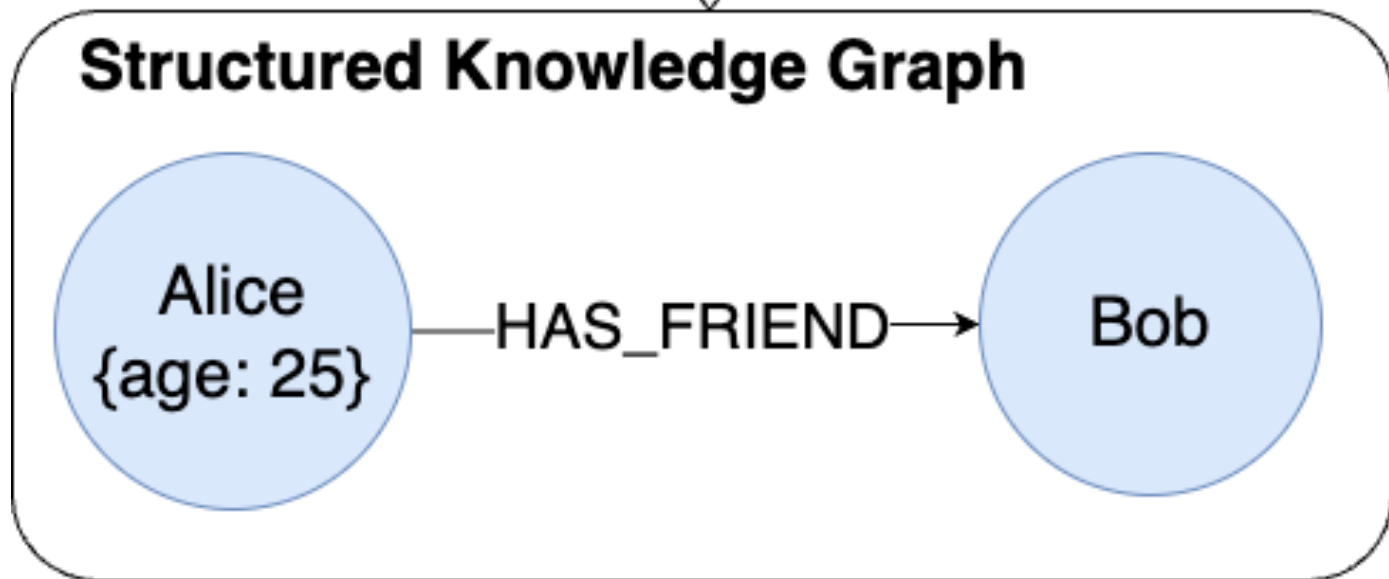
Probabilistic Reasoning in Generative Large Language Models

Aliakbar Nafar¹

Kristen Brent Venable² & Parisa Kordjamshidi³
^{1,3}Michigan State University
²Florida Institute for Human and Machine Cognition {nafarali, kordjams}@msu.edu, bvenable@ihmc.org

Unstructured Text
Alice is 25 years old and Bob is her friend

Information extraction



<https://neo4j.com/developer-blog/construct-knowledge-graphs-unstructured-text/>

Knowledge Graph Tuning: Real-time Large Language Model Personalization based on Human Feedback

Jingwei Sun*, Zhixu Du*, Yiran Chen
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27705

AI-generated Classification Rubrics in the Process of Item Validation

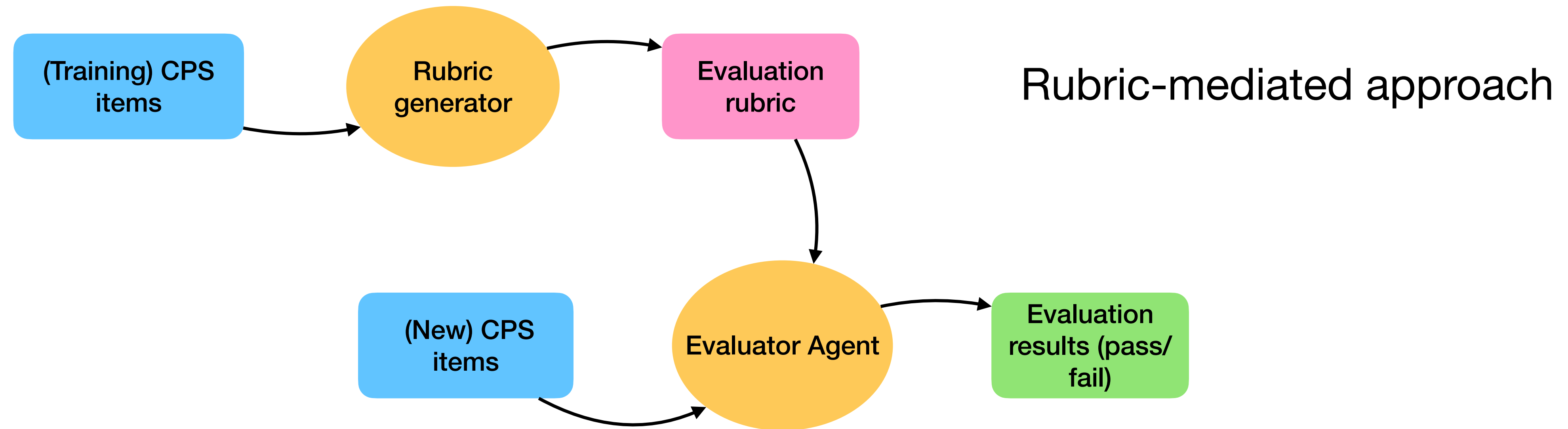
or

Computer, is this a good test item?

Wang, Y., Gopalakrishnan, M., & Bergner, Y. (2025, July). Using generated rubrics to provide a window into item evaluation with multi-agent LLMs. In *International Conference on Artificial Intelligence in Education* (pp. 203-217). Cham: Springer Nature Switzerland.

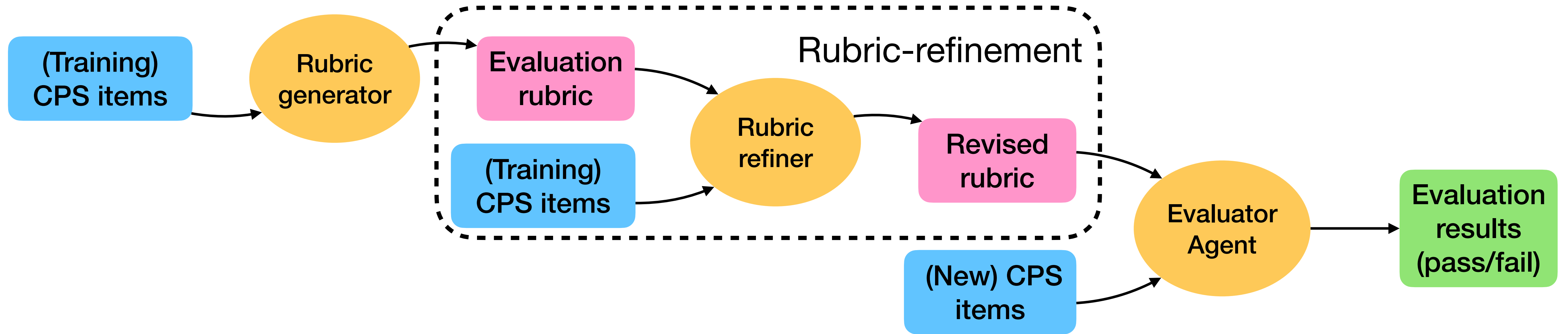
LLMs + Rubrics (simple, language-based knowledge representation)

Direct approach



Rubric-mediated approach

LLMs + Rubrics (language-based knowledge representation)



Separated for each of 3 item types (Jigsaw, Joint construction, and Information request)

+ Each item pool was partitioned into 3 sets, rotating the one used for testing (i.e., 3-fold cross-validation)

+ Three random restarts \rightarrow 27 + 27 rubrics generated



Results, part 1:

Rubrics did not improve evaluation accuracy compared to single model prompts

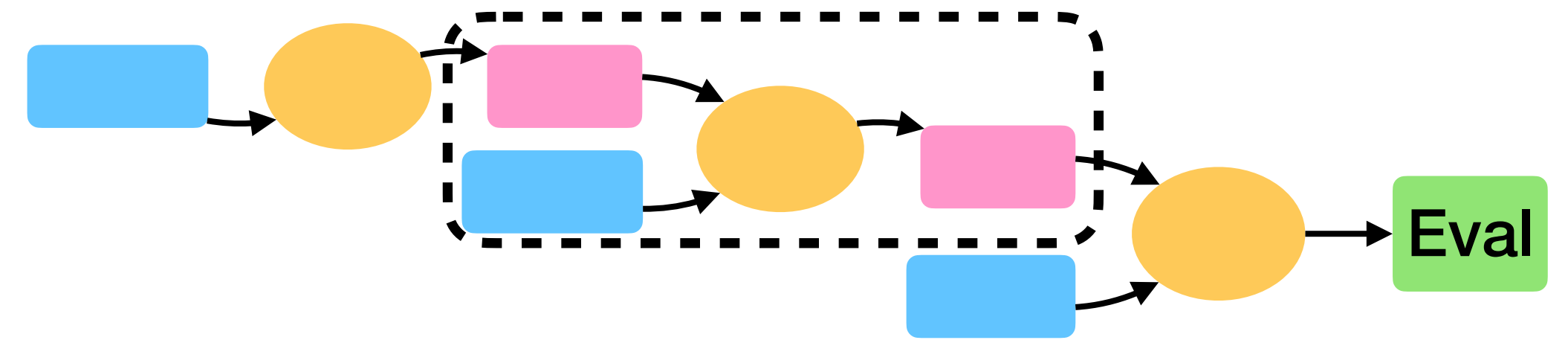
| Rubric | Accuracy (%) | Rubric | Accuracy % | Rubric | Accuracy % |
|------------------------|--------------|--------------------|------------|--------------------|------------|
| jigsaw_1_A | 44.44 | joint_1_A | 71.43 | ir_1_A | 22.22 |
| jigsaw_1_B | 57.14 | joint_1_B | 85.71 | ir_1_B | 57.14 |
| jigsaw_1_C | 100.00 | joint_1_C | 71.43 | ir_1_C | 42.86 |
| jigsaw_2_A | 55.56 | joint_2_A | 28.57 | ir_2_A | 44.44 |
| jigsaw_2_B | 85.71 | joint_2_B | 71.43 | ir_2_B | 42.86 |
| jigsaw_2_C | 85.71 | joint_2_C | 57.14 | ir_2_C | 14.29 |
| jigsaw_3_A | 55.56 | joint_3_A | 85.71 | ir_3_A | 66.67 |
| jigsaw_3_B | 85.71 | joint_3_B | 57.14 | ir_3_B | 57.14 |
| jigsaw_3_C | 57.14 | joint_3_C | 85.71 | ir_3_C | 71.43 |
| Jigsaw avg 69.66(3.14) | | JC avg 68.25(3.89) | | IR avg 46.56(4.73) | |

Jigsaw avg 70% | JC avg 68% | IR avg 47%

single model (no rubrics, one-at-a-time):

Jigsaw avg 86% | JC avg 80% | IR avg 63%

- No differences *at all* between intermediate and final rubrics. So much for fine-tuning with small samples...



- The intermediate rubrics *did not improve* overall evaluation accuracy compared to single passes with a good prompt.

Accuracy with human-expert rubrics

Model used here is o1-mini

Each number is an average of three runs of batch size 12. (3 API calls for ~36 questions).

| Question/rubric | Jigsaw rubrics | Joint rubrics | Request rubrics |
|-------------------|----------------------|----------------------|----------------------|
| Jigsaw questions | 78% (71% ZSP) | 67% | 68% |
| Joint questions | 77% | 76% (73% ZSP) | 72% |
| Request questions | 53% | 59% | 60% (57% ZSP) |

The mis-specified rubric cells above *should* be closer to 50%

Number of Players, Zero: Using Multi-Agent Systems to Evaluate Unsolvable Collaborative Items

or

Joshua: is this a good CPS item?

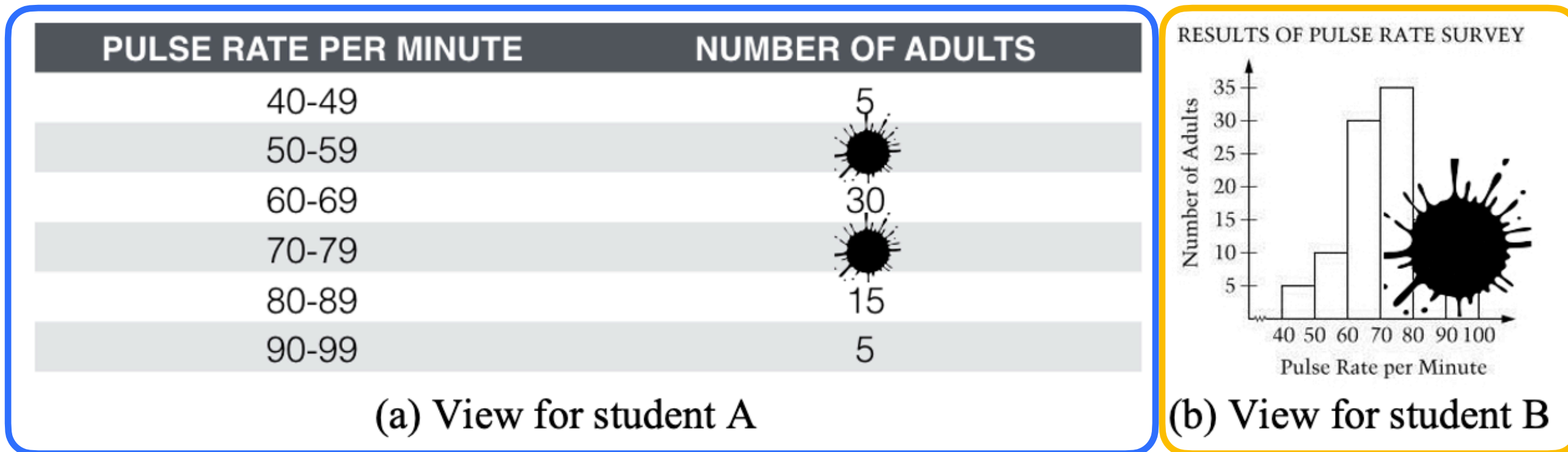
Wang, Y., & Bergner, Y. (2025). Number of Players, Zero: Using Multi-Agent Collaboration to Evaluate Unsolvable Collaborative Items. In *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning-CSCCL 2025*, pp. 409-414. International Society of the Learning Sciences.

MathCHOPS item templates

Bergner & Wang, ISLS 2023

Jigsaw problem type

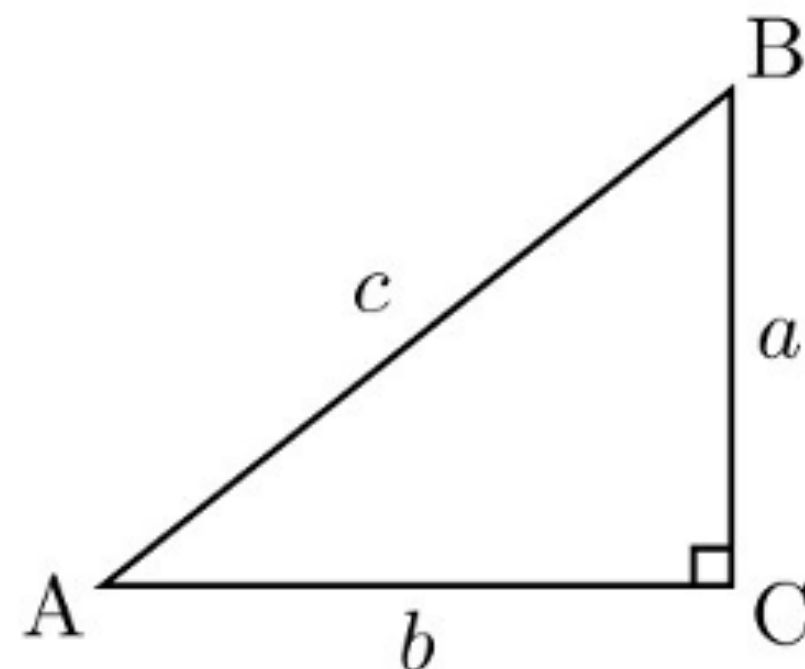
Students each see a different part of a common problem statement



Information-request problem type

The problem requires TWO pieces of information, each student can select ONE.

- side a
- side b
- angle A
- angle B



View for both students

“small c” collaboration, but interdependence can scaffold higher-order thinking (collaboratively)

Joint-Construction problem type

Renee has a lemonade stand. She found that her total profits can be represented in the equation:

$$\text{profit} = 3 \times (\text{cups sold}) - 10$$

Together with your partner, determine a possible value for the number of cups sold and positive total profit in a day.

How many cups were sold? _____

What is the total profit? _____



Morphology of a solvable (good) and unsolvable (bad) CPS math item

Version A:

Emma is packaging 1500 apples in crates.

She wants each crate to have the same number of between 25 and 50 apples.

With your partner, determine a way to arrange all the apples.

You will submit the number of apples per crate, and your partner will submit the total number of crates.

Number of apples per crate: _____

Version B:

Emma is packaging 1500 apples in crates.

She wants to end up with between 30 and 75 crates.

With your partner, determine a way to arrange all the apples.

You will submit the total number of crates, and your partner will submit the number of apples per crate.

Total number of crates: _____

Morphology of a solvable (good) and unsolvable (bad) CPS math item

Version A:

Emma is packaging 1500 apples in crates.

*She wants each crate to have the same **odd number** of between 25 and 50 apples.*

With your partner, determine a way to arrange all the apples.

You will submit the number of apples per crate, and your partner will submit the total number of crates.

Number of apples per crate: _____

Two odd numbers cannot multiply to 1500 (an even number)

Version B:

Emma is packaging 1500 apples in crates.

*She wants to end up with an **odd number** of between 30 and 75 crates.*

With your partner, determine a way to arrange all the apples.

You will submit the total number of crates, and your partner will submit the number of apples per crate.

Total number of crates: _____

This item should fail because of "cases where the problem cannot be solved even when working together" ...

State of the art LLMs fail to detect unsolvability when looking for interdependence

GPT-4o



Verdict: Pass



Reason:

This task requires collaboration between Student A and Student B in order to find a solution that satisfies both constraints:

- Student A knows only the acceptable range for apples per crate (odd numbers between 25 and 50).
- Student B knows only the acceptable range for total number of crates (odd numbers between 30 and 75).

Neither student alone has sufficient information to find a solution that meets both criteria. To solve the task:

1. Student A must consider the odd numbers of apples per crate in the given range: {25, 27, 29, ..., 49}.
2. Student B must consider the acceptable odd number of crates: {31, 33, ..., 75}.
3. They must find a pair (apples_per_crate, crates) such that $\text{apples_per_crate} \times \text{crates} = 1500$, where each value satisfies the respective constraints.

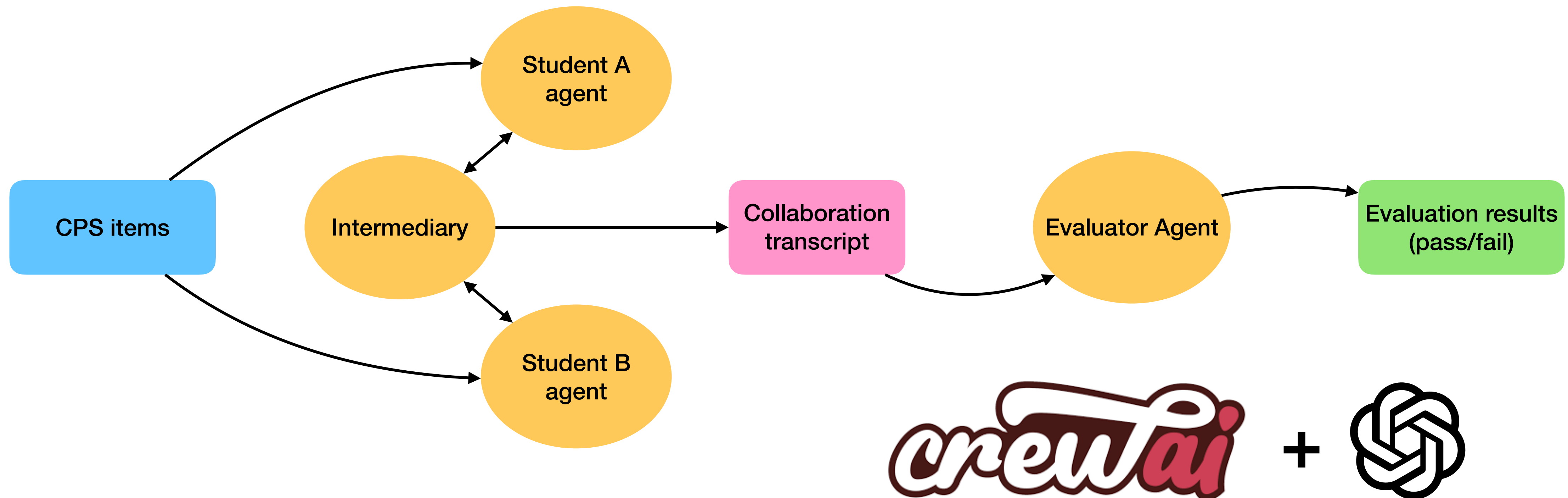
CPS simulation games: LLMs + More LLMs (multi-agent systems)

Direct approach



(Wooldridge, 2009)

Simulated collaboration with multi-agent system



Results: classification accuracy between conditions

Mean (standard deviation) accuracies are shown averaged across items and across five replication runs

| | Solvable (Pass) ($N = 15$) | Unsolvable (Fail) ($N = 26$) | Overall ($N = 41$) |
|------------------------------|---------------------------------|-----------------------------------|-------------------------|
| Single agent w/o sim prompt | 0.93 (0.07) | 0.52 (0.03) | 0.67 (0.02) |
| Single agent with sim prompt | 0.95 (0.03) | 0.48 (0.04) | 0.65 (0.03) |
| | | | |

See paper for more detailed results

Results: classification accuracy between conditions

Mean (standard deviation) accuracies are shown averaged across items and across five replication runs

| | Solvable (Pass) ($N = 15$) | Unsolvable (Fail) ($N = 26$) | Overall ($N = 41$) |
|------------------------------|---------------------------------|-----------------------------------|-------------------------|
| Single agent w/o sim prompt | 0.93 (0.07) | 0.52 (0.03) | 0.67 (0.02) |
| Single agent with sim prompt | 0.95 (0.03) | 0.48 (0.04) | 0.65 (0.03) |
| Multi-agent system | 0.83 (0.10) | 0.88 (0.08) | 0.86 (0.07) |

See paper for more detailed results

סיכום ממצאים

- הערכת תלות הדדית בפב"ש (CPS) מהווה דוגמה להקשר הערכה שמודלי שפה (LLMs) אינם מסוגלים לנתח היטב באמצעות זיהוי דפוסים, מכיוון שלא נחשפו לדוגמאות רבות מסוג זה.
- הניסיונות שלנו לגרום ל-LLMs להבחין בין דוגמאות טובות לדוגמאות גרועות של פריטי CPS חשפו מספר פערים בהיסק.
- כאשר ביקשנו מה-AI לבנות מחוון בעצמו על סמך דוגמאות, התוצאה הייתה בעיקר חזרה שטחית על רעיונות כלליים בנוגע לשיתוף פעולה. באופן מאכזב, המודל אף לא הצליח ליישם מחוון פשוט מאוד כשסיפקנו לו כזה.
- מצד חיובי, אם כי מוגבל, סוכנים שסימלו פתרון בעיות הצליחו לפחות לזהות מתי בעיות אינן ניתנות לפתרון.

Now what?

Scaling believers argue that with enough

- Training data
- Compute for inference-time search
- Multimodal grounding (vision, robotics)
- Better RL/reasoning architectures

...emergent world understanding will develop.

The Skeptics (e.g., Gary Marcus, Yann LeCun, etc.) argue current approaches have fundamental limitations:

- Causal understanding: LLMs correlate, don't model causation
- Compositional generalization: Struggle with novel combinations
- Grounding: Text-only training lacks physical/perceptual foundations
- Explicit reasoning: Chain-of-thought is still just text generation

Assessment is observing a person acting in a handful of particular situations, interpreting the situations and actions through the lenses of particular practices or [linguistic, cultural, and social] patterns, and making inferences about the person's capabilities for acting or learning in other situations in which the targeted patterns are relevant.

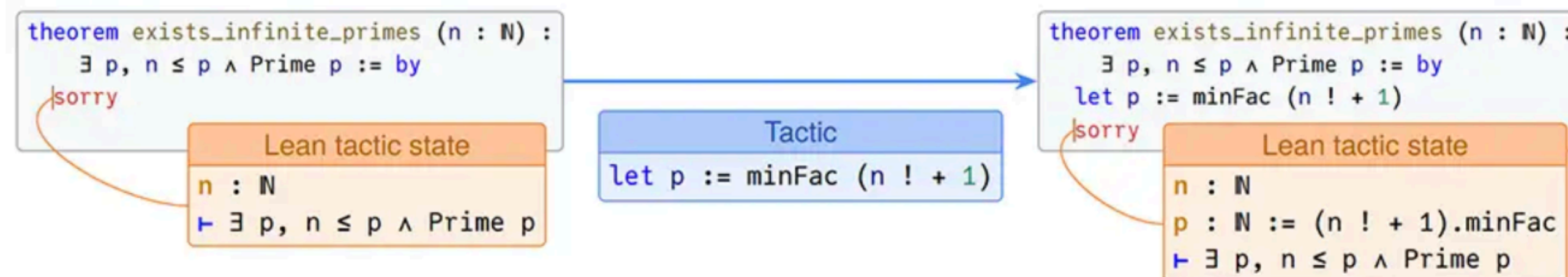
Robert Mislavy, 2018

Gary Marcus, 2019

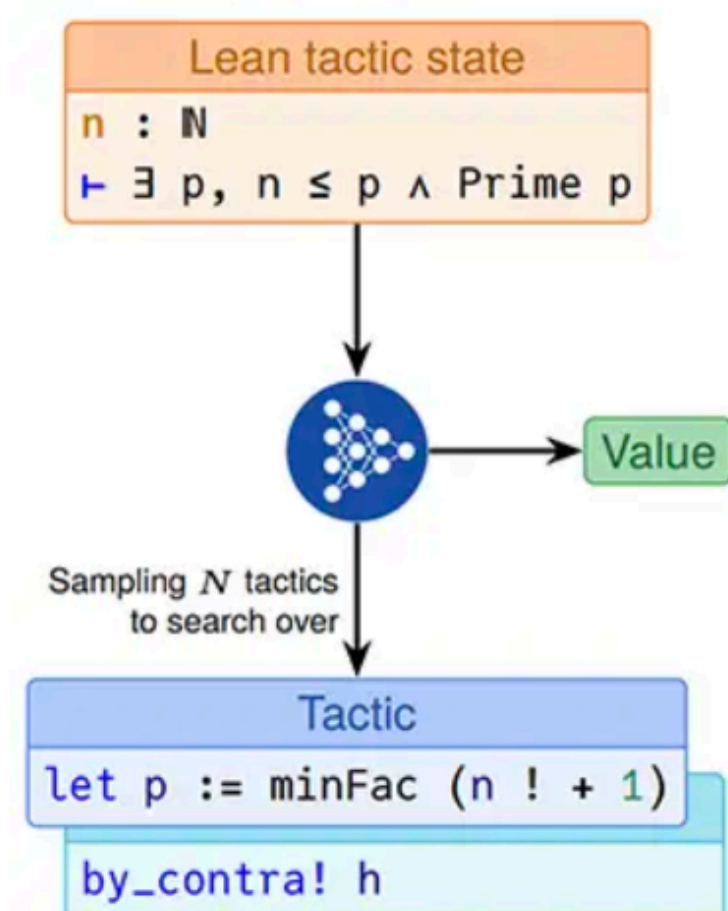
Virtually every sentence that we encounter requires that we make inferences about how a broad range of background knowledge interrelates with what we read. Deep learning lacks a direct way of representing that knowledge, let alone performing inferences over it in the context of understanding a sentence.

Olympiad-level formal mathematical reasoning with reinforcement learning

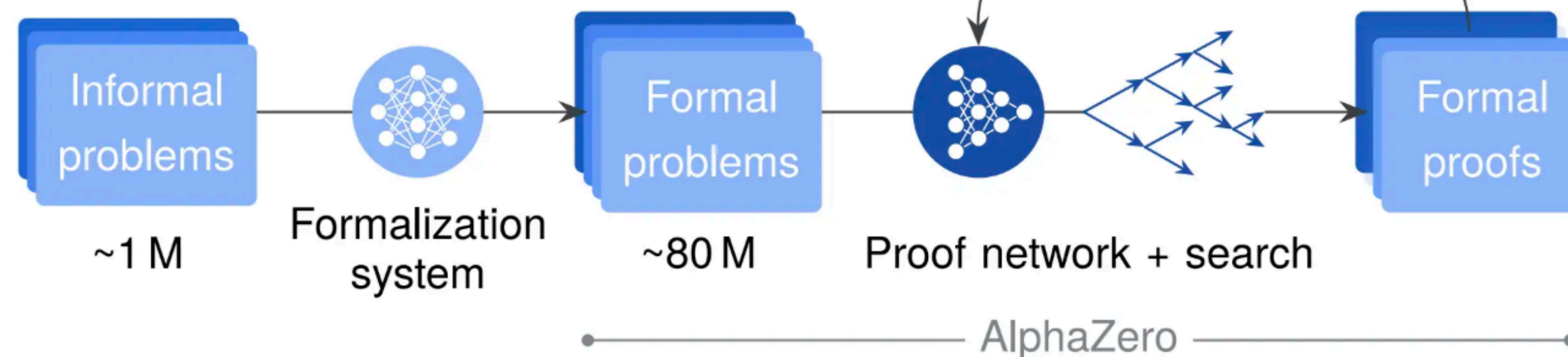
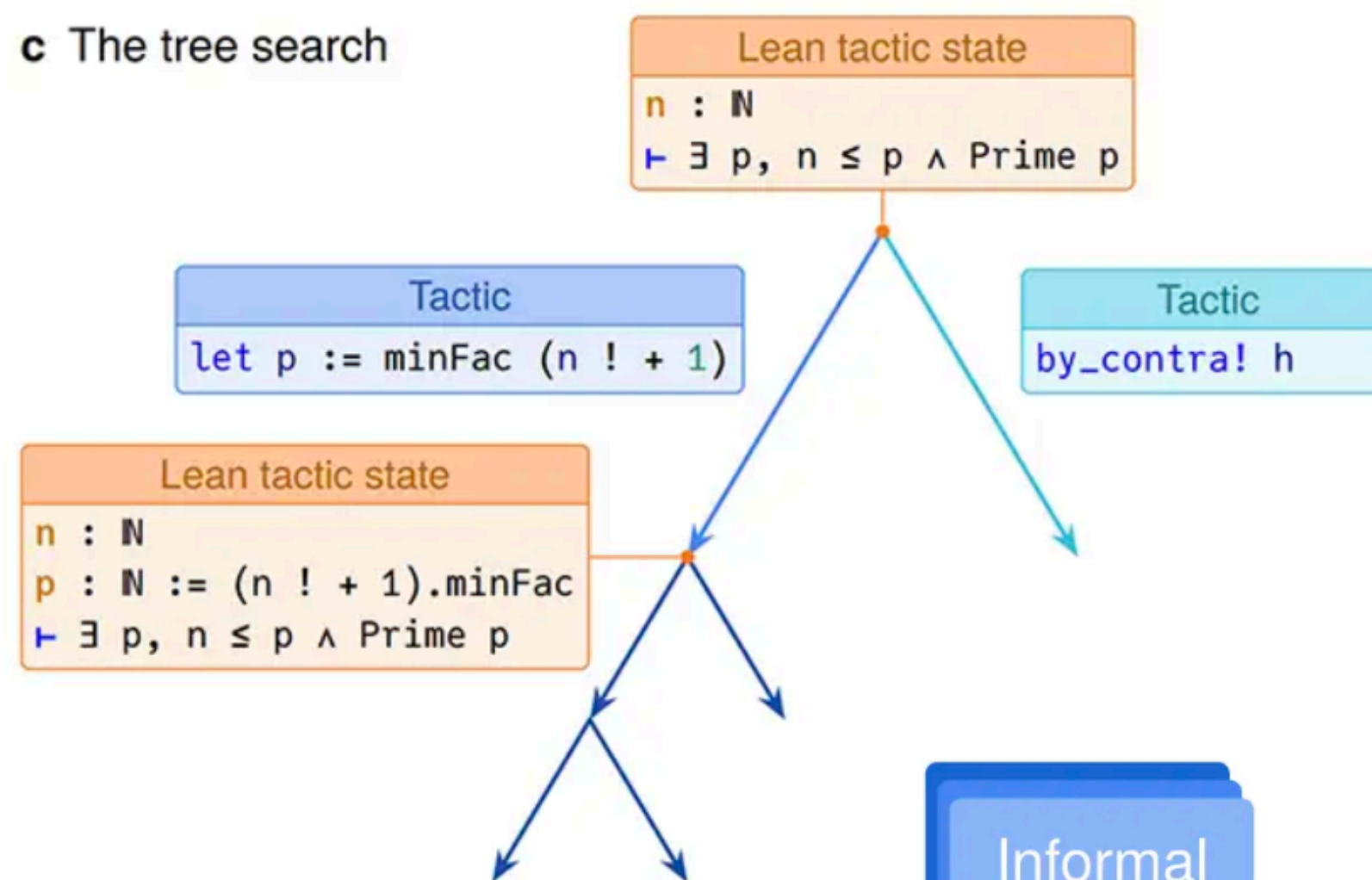
a Lean environment



b The proof network



c The tree search



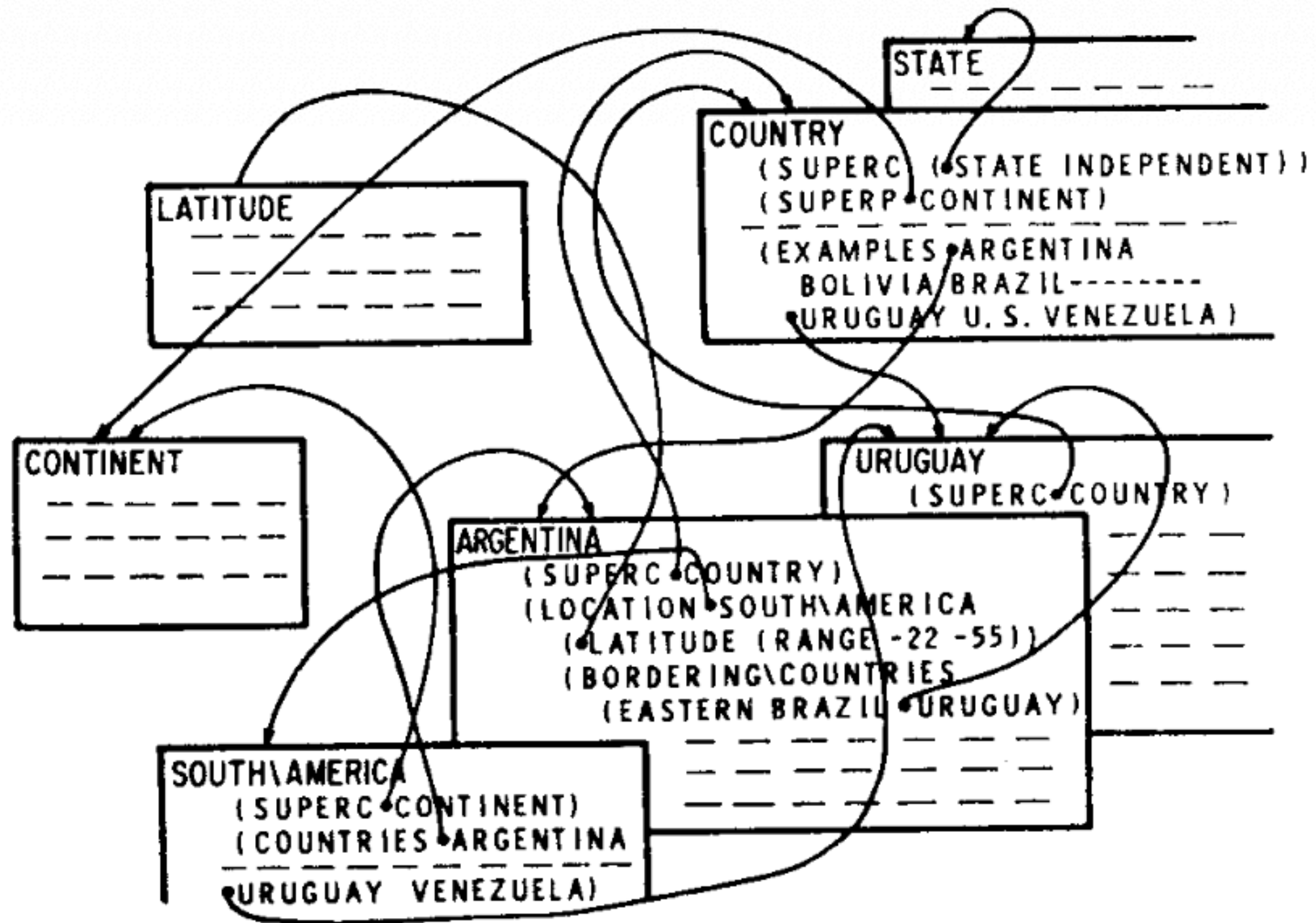
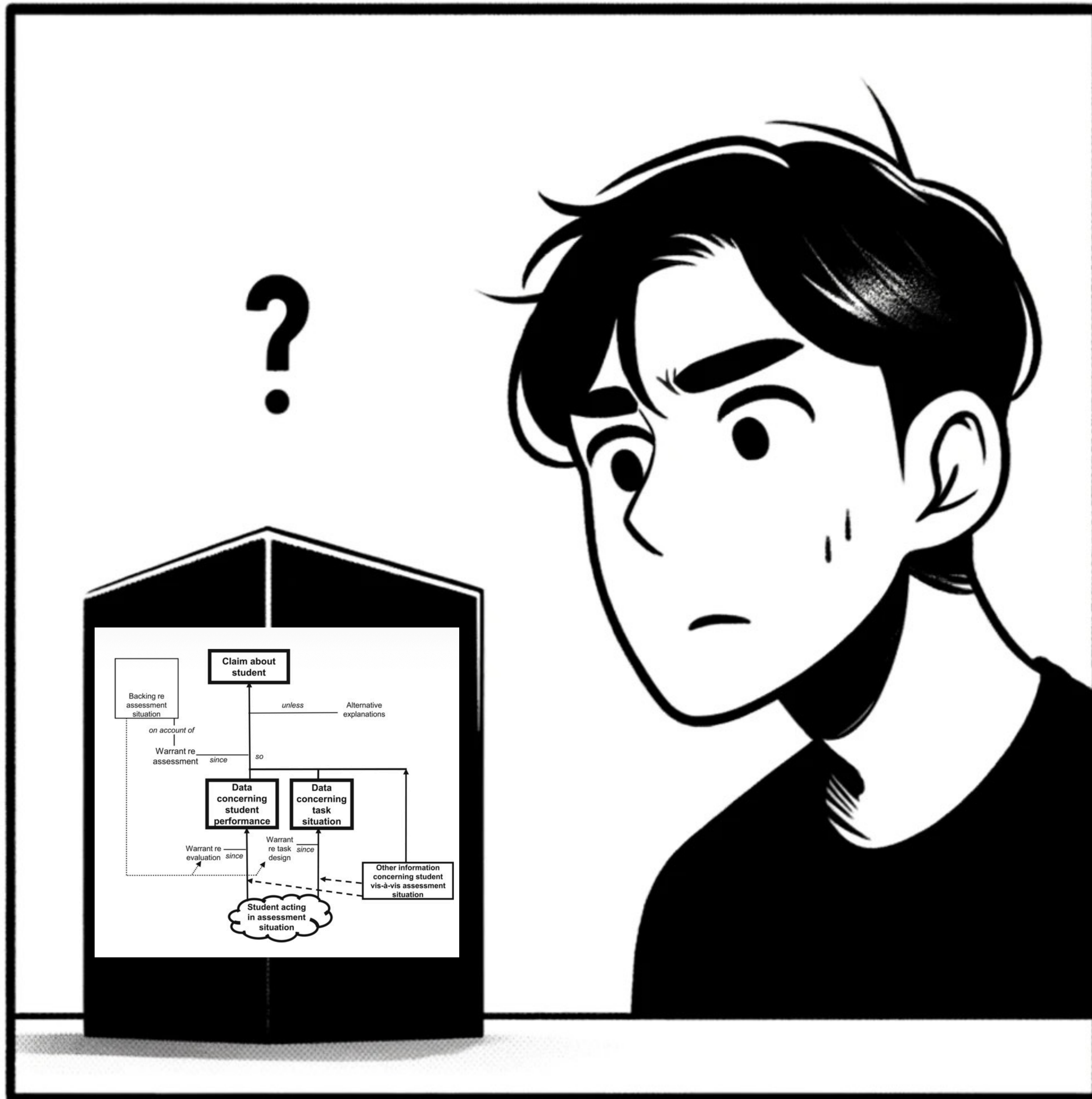


Fig. 2. Representation of a portion of a network on South America.

```
(RPAQQ LATITUDE ((CN LATITUDE)
  (DET THE DEF 2))
NIL
(SUPERP NIL (DISTANCE NIL ANGULAR (FROM NIL
  EQUATOR)))
(SUPERP (I 2)
  LOCATION)
(VALUE (I 2)
  (RANGE NIL -22 55))
```

The development of SCHOLAR is a step in the direction of CAI programs that *know* what they are talking about, the same way human teachers do. This necessary preoccupation with properly representing and intelligently utilizing knowledge has led us to the use of a semantic network for the data base and, generally, an artificial intelligence approach in the program.

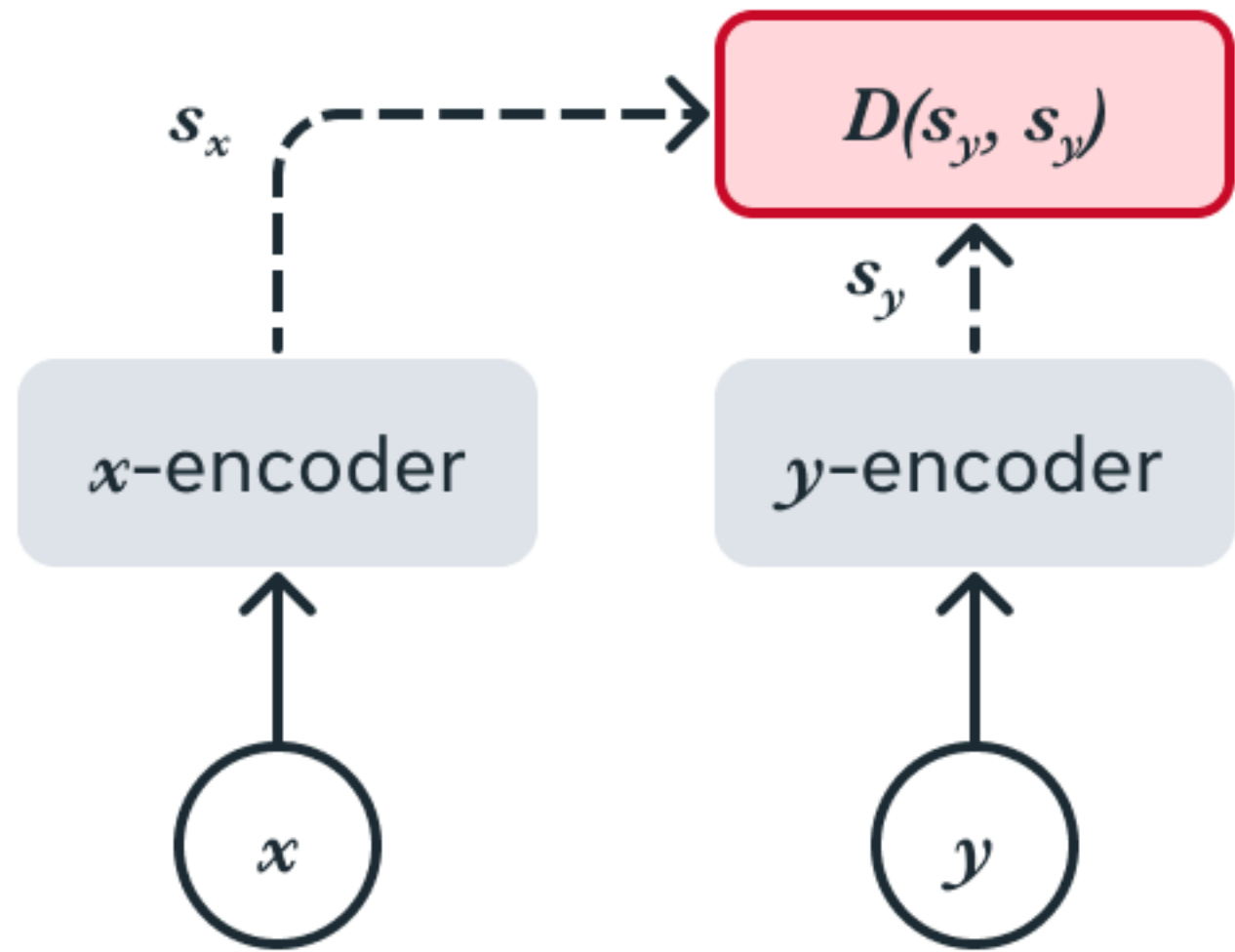
```
SOUTH AMERICA)
(AREA (I 2)
  (APPROX NIL \ 1200000))
(LOCATION NIL SOUTH AMERICA (LATITUDE (I 2)
  (RANGE NIL -22 -55))
  (LONGITUDE (I 4)
  (RANGE NIL -57 -71))
  (BORDERING COUNTRIES (I 1)
  (NORTHERN (I 1)
  (RANGE NIL -22 55))
```



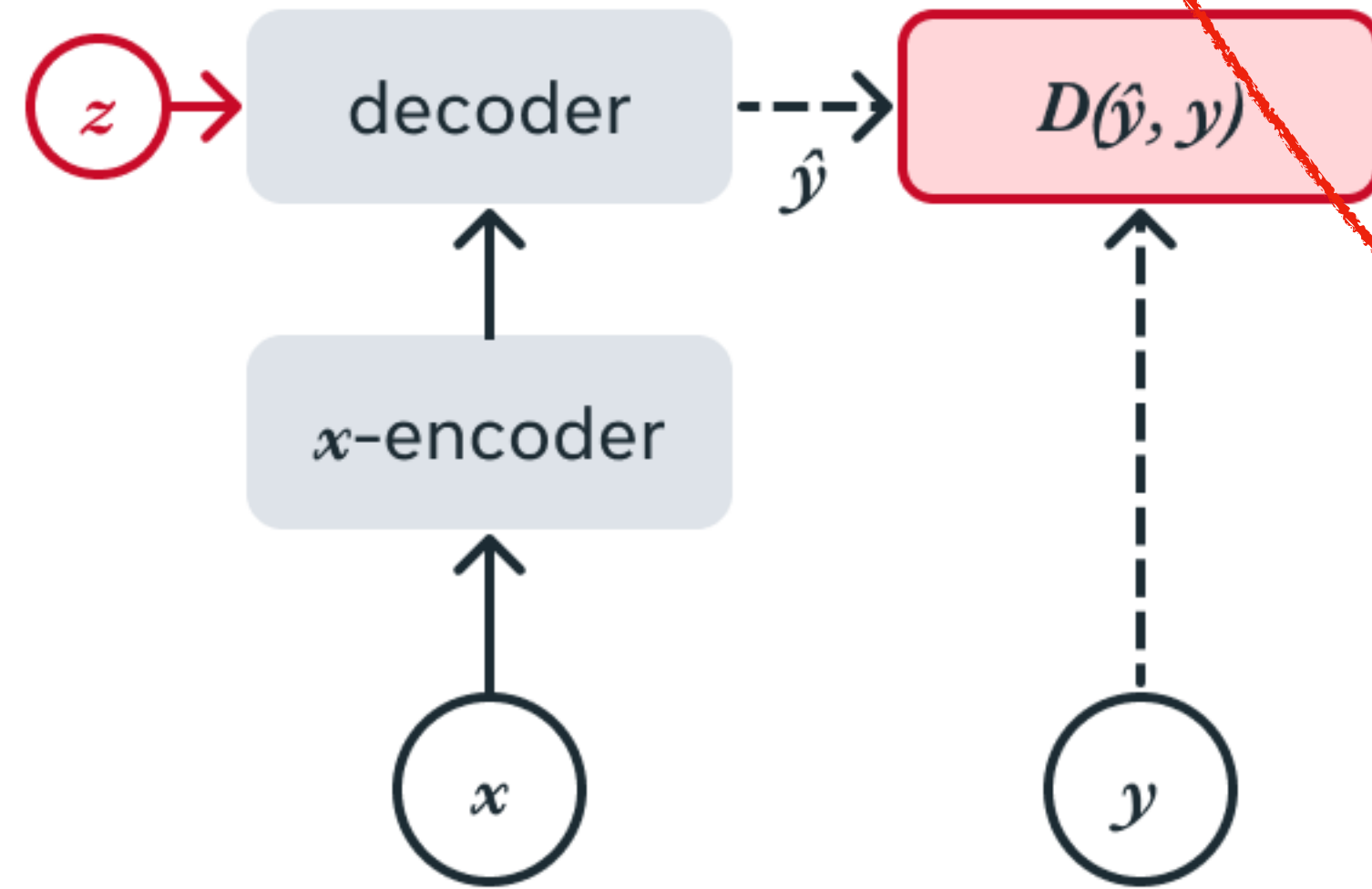
תודה!

Not LLMs

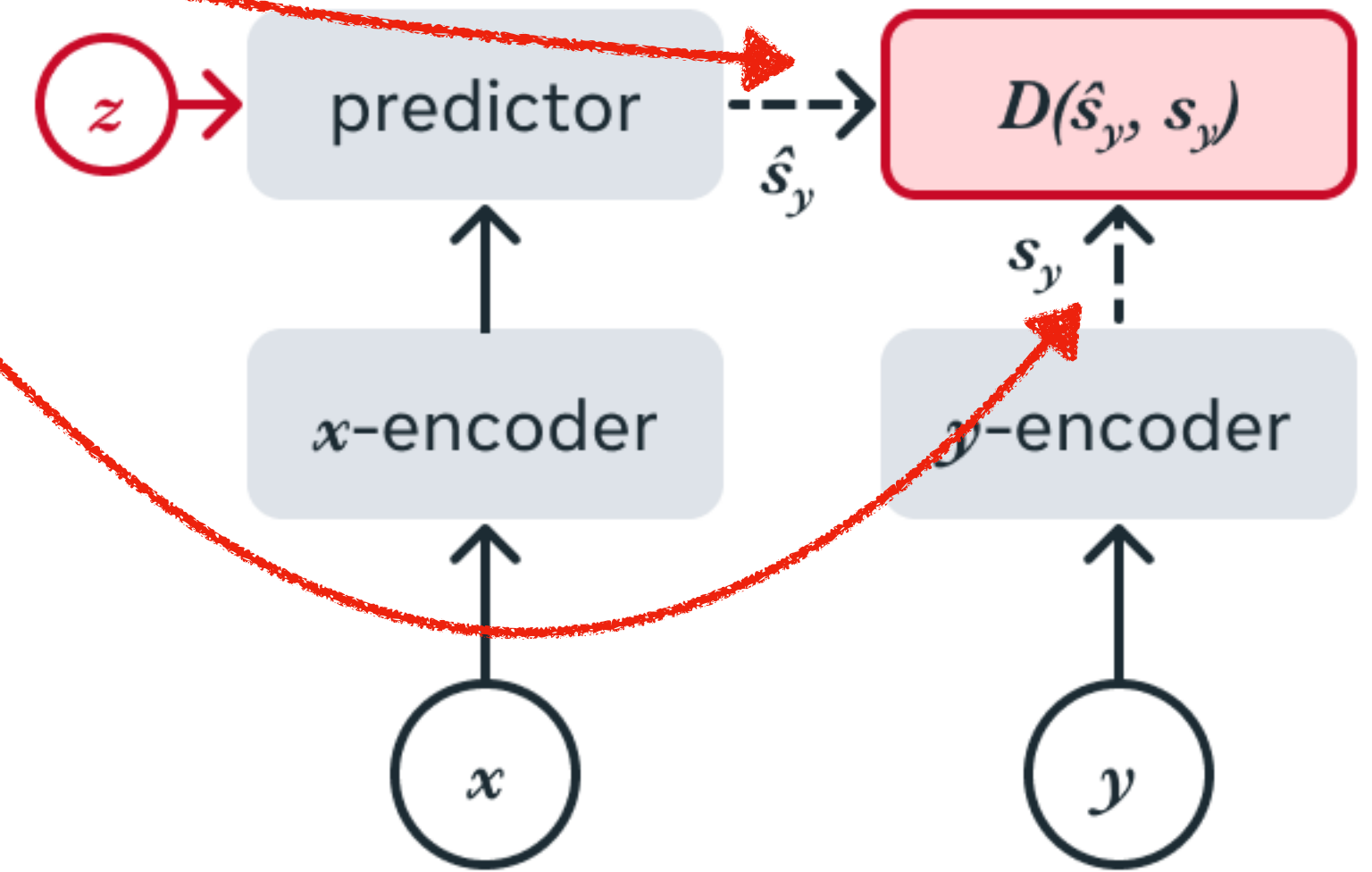
Self-supervised learning? Or "knowledge"?



(a) Joint-embedding architecture



(b) Generative architecture



(c) Joint-embedding predictive architecture