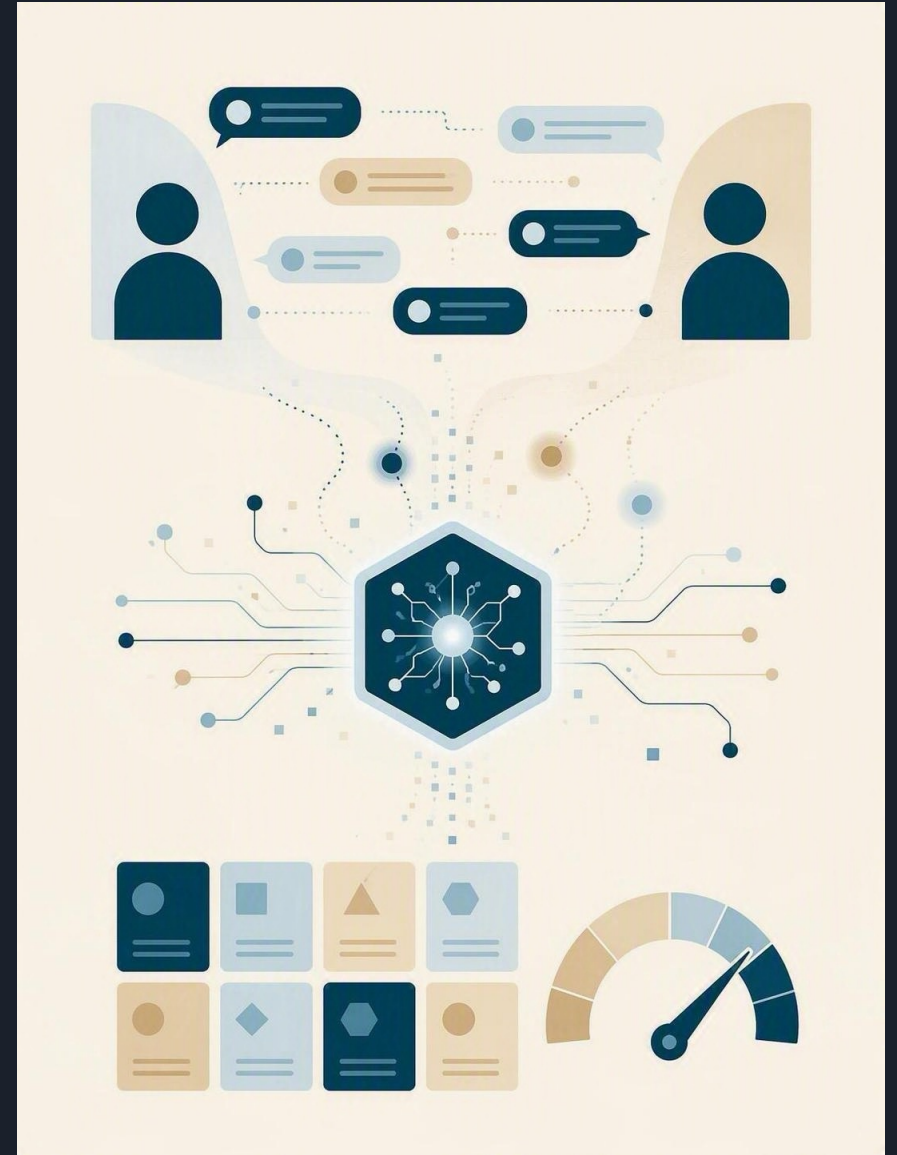


Scaling Evidence Identification from Communication Process Data with AI

Jiangang Hao

ETS Research Institute, ETS

Beyond Results Workshop - 2026



AI Transforms Assessments – an ECD Perspective



Competency Model

- Emergent new constructs when humans interact with AI
- Existing constructs changed as AI reshapes cognitive process



Task Model

- Automated generation of multimodal tasks and content
- Interactive, adaptive tasks for hard-to-measure skills
- Agentic coding to develop assessment instruments



Evidence Model

- Automated item and task characterization
- Scaling up multimodal data scoring and aggregation

The core measurement principles must still be respected

Validity

Reliability

Fairness

Comparability

Generalizability

Process Data in the AI Age

WHY PROCESS DATA MATTERS

- Providing direct evidence for measuring constructs
- Informing designs of items and delivery platforms
- Expanding knowledge of psychology and cognitive science
- Creating more informative reporting



How AI changed process data analysis?

AI agents now handle logging, transformation, and structuring, plus feature extraction very well, the data engineering bottleneck is largely alleviated

Hao & Mislevy, 2022

- Providing direct support to psychometric operations
- Advancing new psychometric and statistical models
- Supporting test security and remote testing
- Shedding new light on group differences and fairness



Not yet solved

Evidence identification, deciding which behaviors count as valid evidence of the target construct, still needs human-designed, validated models.

Communication Process Data as Evidence of Complex Skills

Social Emotional Skill



Listening Skill Step 1: Look
There are three steps for listening to others. Click NEXT to see the next skill step.

0:05/0:15

Task Resources

TELL

Who is listening and who is not?

Who in these pictures is listening?

Who in the pictures is not likely listening?

Chat with your group or speak up in class!

Who in the pictures is making eye contact? Smiling? Showing good position?

Unit 1 | Lesson 2
Listen To Others

Copyright © 2023 SAS. All rights reserved.

14

epcal

Exit Task NEXT 1/18

Sync Team Raise Hand

Dog Cat AI Agent

Cat Thursday 10:16am

Thanks everyone! Almost there.

You Thursday 11:41am

Awesome! Thanks.

You Thursday 11:41am

AI: What's next?

AI Agent Thursday 11:44am

I think we've finished discussion on this slide.

You Thursday 11:41am

Let's move forward.

Message

Communication Skill

*ets research institute

00:25

STEP 2 OF 3 KICKOFF MEETING - EVENT PLANNING

Back Next

Kickoff Meeting - Event Planning

INSTRUCTIONS
Join the virtual planning meeting and follow the discussion carefully. As you listen, take notes to identify:

- important event details
- communication suggestions
- possible next steps

When Alex asks, use your notes to give a short summary during the meeting.

GOAL
Listen carefully to understand and summarize key information communicated throughout a conversation.

Alex K.

Maya T.

Jiangang

Click to Talk

Submit Notes

REAL-TIME TRANSCRIPT

ALEX K.
Jiangang, can you take notes for us? Thanks for joining. We need a basic plan for the school car wash fundraiser—date, location, fundraising goal, and next steps. I'm thinking Saturday, April 18, in the school parking lot.

MAYA T.
That makes sense. I can make a list of what we'd need and what might be donated. We can check with local businesses to see if they can help.

ALEX K.

MEETING NOTES
Use this space to record the most important information from the meeting.

fund raising for \$500

*ets research institute

00:28

STEP 3 OF 3 EVENT FLYER - EXPLAIN

Back Next

Event Flyer - Explain

INSTRUCTIONS
After reviewing your revised flyer, Alex, the committee chair, asked you to explain the reasoning behind your changes before the flyer is finalized.

Chat with Alex and explain:

- your choice of image for the flyer
- the wording changes you made to the draft
- how these choices help the flyer communicate the message more effectively to students

GOAL
Explain how communication choices can make a message clearer, more engaging, and more appropriate for the audience.

ALEX K.
I looked at your revised flyer and the image you picked. Can you explain why you made those text changes and why that image is a good fit for students?

NI, how are you

ALEX K.
Hey, I'm good. What made you change the wording on the flyer?

I think that align better with the pic.

ALEX K.
Got it—how does that picture help students quickly understand what the car wash is about?

EXPLAIN YOUR DESIGN RATIONALE TO ALEX:

Type your response...

You have a max of 10 turns.

WHY IT MATTERS

- 1 Communication process data contain direct evidence of the targeted skills
- 2 The “core” step in any scoring strategy is coding these communication process data into categories according to some framework
- 3 Scaling that coding with LLMs is the core step to unleash the full potential of AI-enabled assessment

Traditional Coding Process

Coding maps turns of communication into predefined categories from an established framework — embedding communication into a target vector space.



HUMAN CODING PLATFORM

CPS-RATER: AUTOMATED CODING FOR CHATS

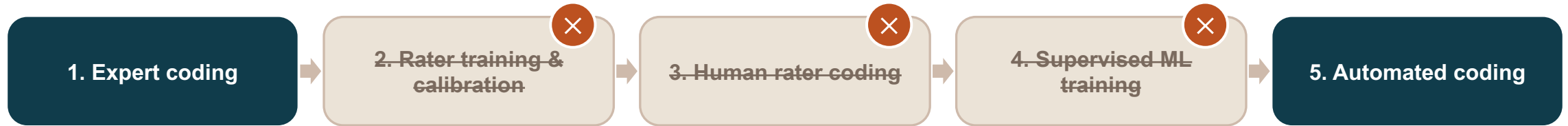
Methods	Accuracy (%)	Kappa (unweighted)
Human-Human	75.8	0.675
Baseline (the most frequent category)	29.0	NA
Maximum Entropy	66.9	0.551
Random Forest	69.8	0.589
Naiver Bayes	70.4	0.596
Linear SVM	71.9	0.619
Linear Chain CRF	73.2	0.636

Hao et al., ETS RR-17-07, 2017

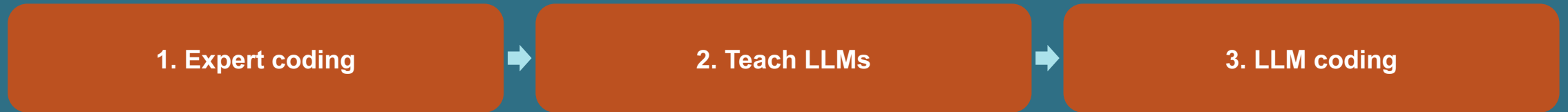
LLM-based Coding

LLMs COLLAPSE FIVE STEPS INTO THREE

Traditional pipeline — 5 steps



LLM-based pipeline — 3 steps



FASTER ITERATION

Trying a different coding framework now costs a prompt revision, not a new rater-training and model-training cycle.

BETTER ACCURACY

LLM-based coding also improves agreement with expert codes over the supervised ML baseline.

Coding Is a Context-Based Reasoning Task

LLM-based coding is one instance of a much broader AI challenge: reasoning over a given context.

Part of a bigger problem

Coding open-ended process data = grounding an answer in a specific, supplied context.

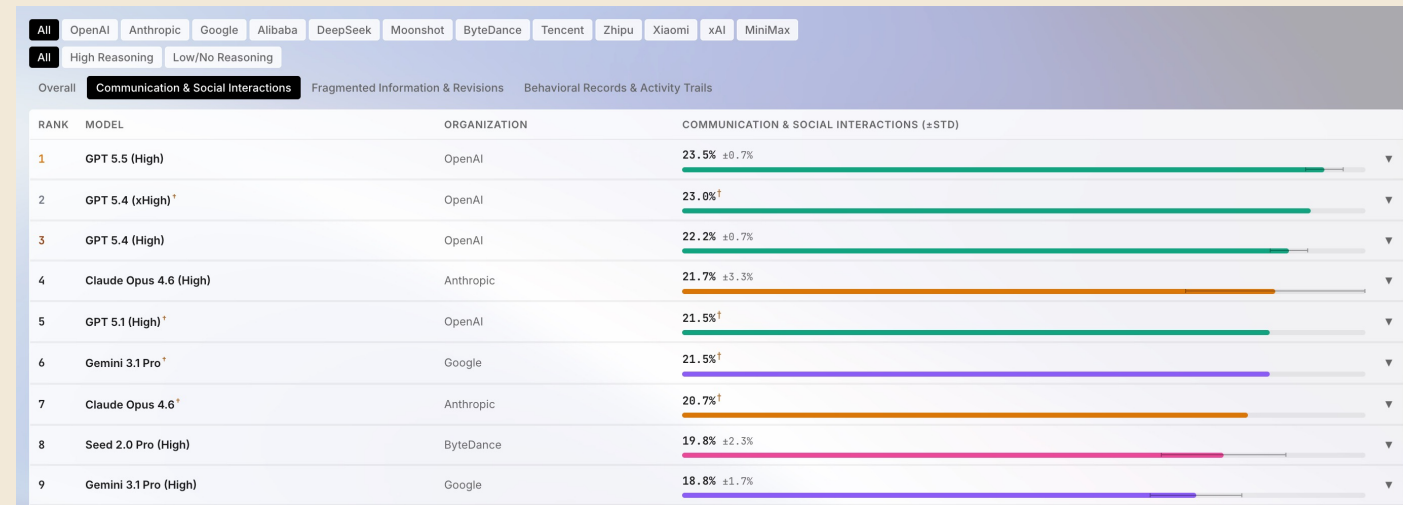
Not solved by scale

Frontier models still cluster far below human reliability on context-bound benchmarks; a bigger model does not close the gap.

It is a long-tail problem

Performance breaks on the rare, idiosyncratic cases: local rubrics, unusual phrasing, domain conventions, often thinly covered in training data.

EVIDENCE: CONTEXT-BOUND REASONING BENCHMARK

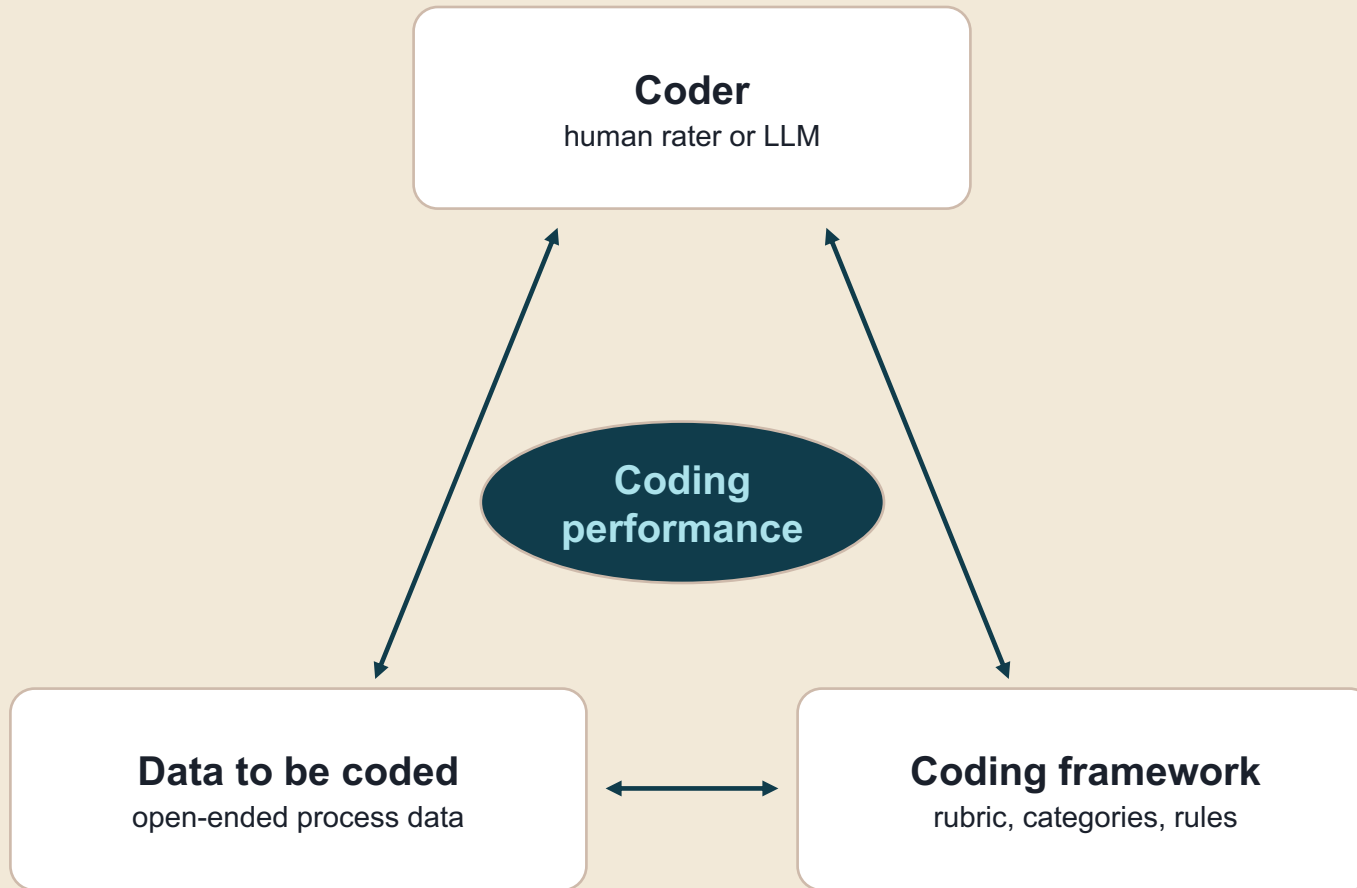


Best models still score in the 20% range

<https://www.clbench.com>

Coding Performance Depends on Three Interacting Components

THREE INTERACTING COMPONENTS



WHAT WE SEE IN PRACTICE

Easy, well-defined tasks

LLMs beat humans, e.g., spotting typos or surface features in text.

Framework–data misalignment

When the rubric does not map cleanly onto the data, Both human and LLM struggle.

Case-by-case diagnosis

What really matters is figuring out what works, and what does not, at each stage.

Some Important Questions

1 How accurately can LLMs code communication relative to trained human raters?

2 Do newer reasoning and agentic LLMs meaningfully improve coding accuracy?

3 Can we improve coding performance with feedback from mis-coding?

4 Do hierarchical prompting strategies improve coding?

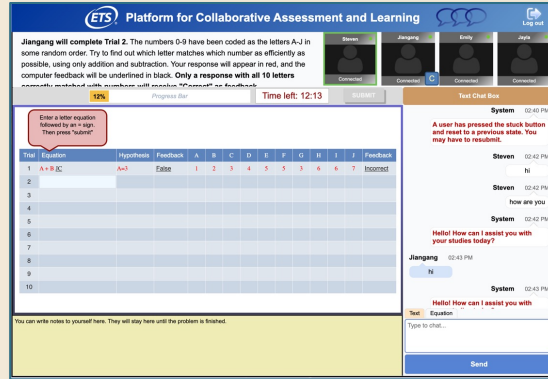
5 Is LLM-based coding consistent across demographic groups?

6 Do LLMs exhibit the same miscoding patterns as human raters?

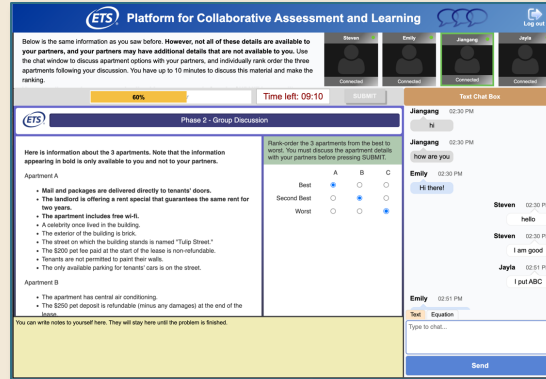
Empirical Study

STUDY DESIGN One typical framework applied to all three datasets | two raters coded independently, differences adjudicated | ~8,000 chat turns

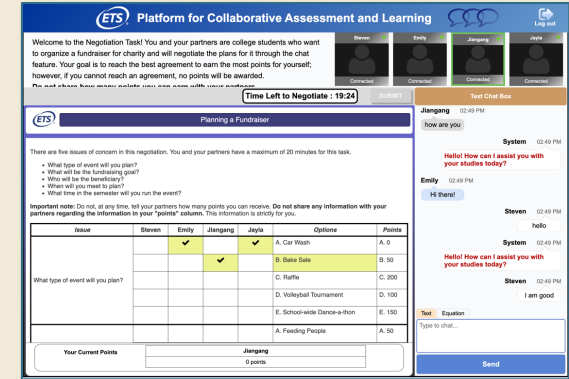
Problem Solving



Decision Making



Negotiation



EVOLUTION OF CPS CODING FRAMEWORKS

2012

ATC21S

Participation · Perspective taking · Social and task regulation · Knowledge building

2015

PISA

Establish/maintain shared understanding · Take action to solve the problem · Maintain team organization

2015

ETS (Liu)

Share idea · Negotiate idea · Regulate problem solving · Maintain communication

2017

ETS (Andrews-Todd)

Collaborative · Cooperative · Dominant-dominant · Dominant-passive · Expert-novice

2023

ETS (Kyllonen) ← USED IN THIS STUDY

Maintain communication · Stay on track · Elicit information · Sharing information · Acknowledging

2026

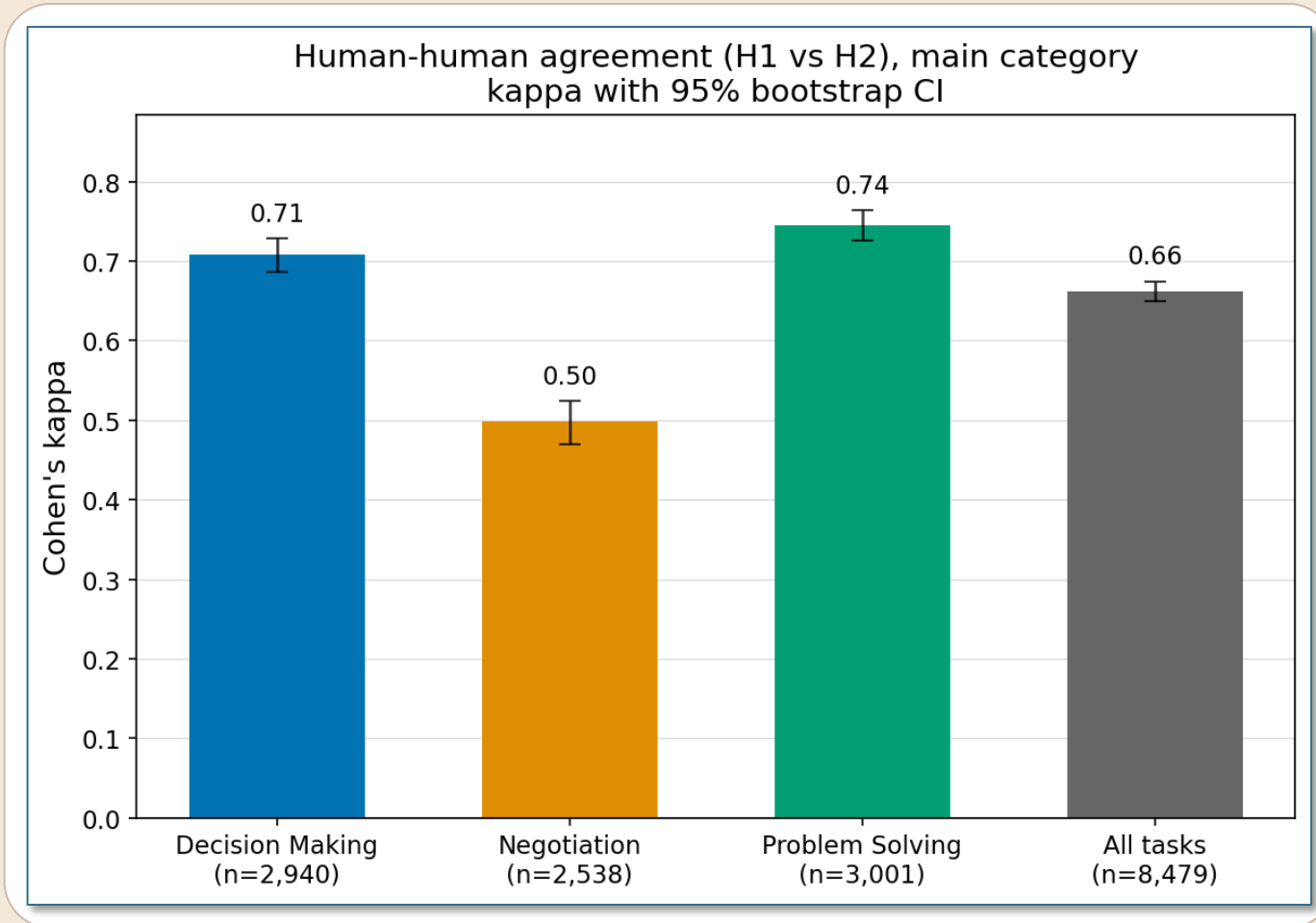
NAEP

Attending to others' contributions · Evaluating merit of ideas · Responding to ideas

The CPS Coding Framework Used in this Study

- 1. Maintaining communication (MC):** chats that involve greetings, emotional responses (including emojis), technical discussions, and other communications that cannot be classified elsewhere.
- 2. Staying on task (OT):** chats that keep things moving, that involve monitoring time, and steering team effort.
- 3. Eliciting information (EI):** chats that elicit information from another about the task, including strategies, goals, and opinions.
- 4. Sharing information (SI):** chats that share information, strategies, goals, or opinions.
- 5. Acknowledging (AK):** chats involving acknowledging partners' input, stating agreement or accepting a tradeoff, stating disagreement or rejecting a tradeoff, building off one's own or a teammate's idea, and proposing a negotiation tradeoff or suggesting a compromise.

Alignment between Coding Framework and Data Matters



Cohen's kappa, human-human agreement (H1 vs H2), 95% bootstrap CI

THE FINDING

Human raters agree far less on Negotiation than on the other tasks - the same framework does not fit every dataset equally.

0.71

Decision Making

n = 2,940

0.50

Negotiation

n = 2,538

0.74

Problem Solving

n = 3,001

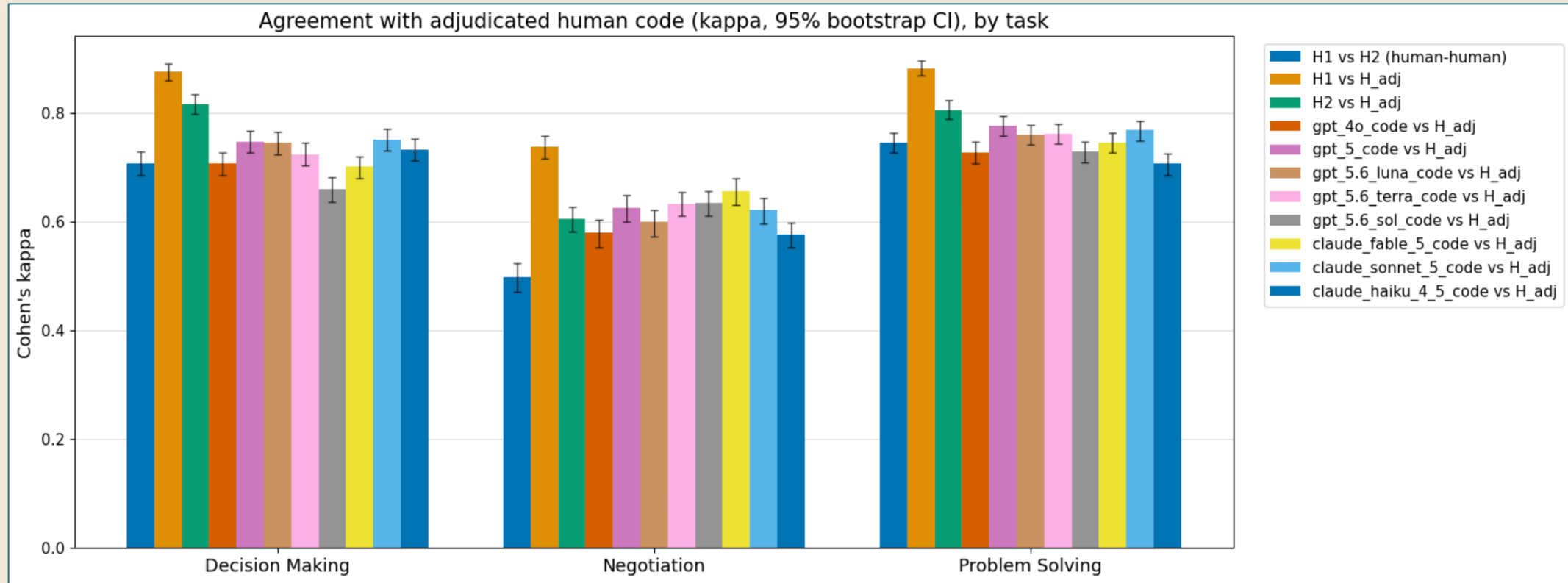
0.66

All tasks

n = 8,479

1 LLMs Match Human Raters 2 Stronger LLMs Add Little

AGREEMENT WITH THE ADJUDICATED HUMAN CODE (COHEN'S KAPPA, 95% CI), BY TASK AND LLMs



MATCHES HUMAN RATERS

Top LLMs reach kappa 0.71-0.78 against the adjudicated human code - on par with two trained human raters.

NEWER IS NOT BETTER

GPT-5.x, Claude and GPT-4o land within a few hundredths of each other; no newer model breaks away.

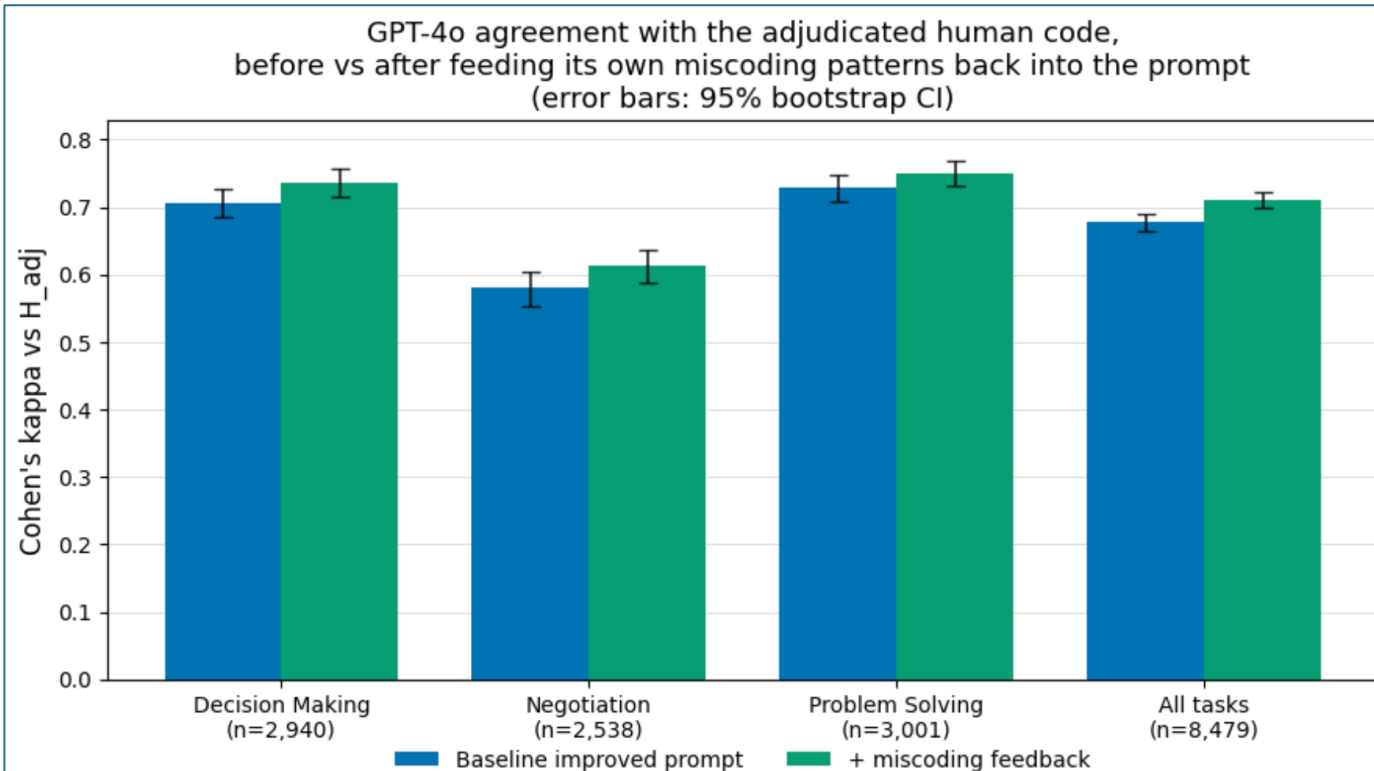
TASK BEATS MODEL

Every model dips on Negotiation and peaks on Problem Solving - task fit drives performance more than model choice.

3 Feedback from Mis-Coding Brings No Significant Gain

MOTIVATION

Can we incorporate mis-coding information into the prompt to improve coding?



THE FINDING

Feeding the model its own mis-coding patterns nudges agreement up on every task, but none of the gains are statistically significant - the 95% confidence intervals overlap throughout.

0.68 → 0.71

All tasks: kappa vs. adjudicated human code (not significant)

Not significant

Gains of +0.02 to +0.03 with overlapping 95% CIs on every task

4 Hierarchical Prompting Makes it Worse

THE APPEAL: ANY COMPLEX CODING FRAMEWORK CAN BE DECOMPOSED INTO A HIERARCHY OF SIMPLER ONES - code the main category first, then the subcategory within it

MAIN AND SUBCATEGORY

Main category	Subcategory
1. Maintaining communication (MC): chats that involve greetings, emotional responses (including emojis), technical discussions, and other communications that cannot be classified elsewhere.	1. Greeting (GRT): chats that greet partner and engage in small talk. 2. Emotional Response (EMO): chats that emphatically express an emotion or feeling. 3. Technical Discussion (TD): chats that deal with technical issues. 4. Other (NA): chats that do not fit into other coding categories; generally, off-topic or typos, or incomplete or partial statements.
2. Staying on task (OT): chats that keep things moving, that involve monitoring time, and steering team effort.	5. Making Things Move or Monitoring Time (MTM): chats that help teams stay on track but are not related to the task content. 6. Steering Team Effort (STE): chats that steer the conversation back to the task.
3. Eliciting information (EI): chats that elicit information from another about the task, including strategies, goals, and opinions.	7. Elicits Information (EIF): chats that ask for information related to the task. 8. Elicits Strategies (EST): chats that ask about task strategies but do not propose any. 9. Elicits Goals or Opinions (EOP): chats that ask for task-related goals or opinions.
4. Sharing information (SI): chats that share information, strategies, goals, or opinions.	10. Shares Information (SIF): chats that share information related to the task. 11. Shares Strategies (SST): chats that share strategies for solving the task. 12. Shares Goals or Opinions (SOP): chats that share task-related goals or opinions.
5. Acknowledging (AK): chats involving acknowledging partners' input, stating agreement or accepting a tradeoff, stating disagreement or rejecting a tradeoff, building off one's own or a teammate's idea, and proposing a negotiation tradeoff or suggesting a compromise.	13. Acknowledge Partner Input (API): chats that neutrally acknowledge a partner's statement. 14. State Agreement or Accept Trade-off (AGR): chats that agree with or support a partner's statement. 15. State Disagreement or Reject Trade-off (DAG): chats that disagree with a partner's statement. 16. Build-off (BOF): chats that add details to a previously made chat turn, their own or another player. 17. Propose Trade-off (PTO): chats that suggest a solution of give and take; may simultaneously reject an option on the table and propose a compromise.

HIERARCHICAL APPROACH

TWO-STEP

Output the main category, then assign the subcategory in a separate step.

VS. DIRECT APPROACH

No hierarchy: assign subcategories directly, never mentioning the main categories.

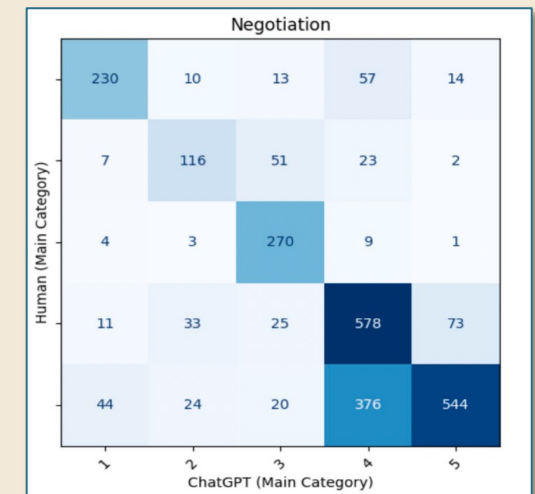
STATISTICAL SIGNIFICANCE: MIXED-EFFECTS LOGISTIC MODEL

Agreement ~ prompt strategy + task, with random effects for person and turn.

Direct vs. 2-step hierarchical: $\beta = 0.162$, $z = 3.41$, $p < .001$ — direct is significantly better.

Decomposing the framework into a hierarchy of simpler ones does not help

Miscoding Propagation



5

LLM Coding Is Consistent across Demographic Groups

AGREEMENT VARIABLE Y

Human and AI disagree: $Y = 0$

Human and AI agree: $Y = 1$

DEMOGRAPHIC VARIABLES

Gender: male (ref) / female

Race/ethnicity: White (ref) / Black /
Hispanic / Asian

GENERALIZED LINEAR MIXED-EFFECTS MODEL (GLMM)

Model 1: $\text{Logit}(P(Y = 1)) = \beta_0 + \beta_1 \cdot \text{demographic} + u + v$

Model 2: $\text{Logit}(P(Y = 1)) = \beta_0 + \beta_1 \cdot \text{demographic} + \beta_2 \cdot \text{task} + \beta_3 (\text{demographic} \cdot \text{task}) + u + v$

Person effect

Team effect

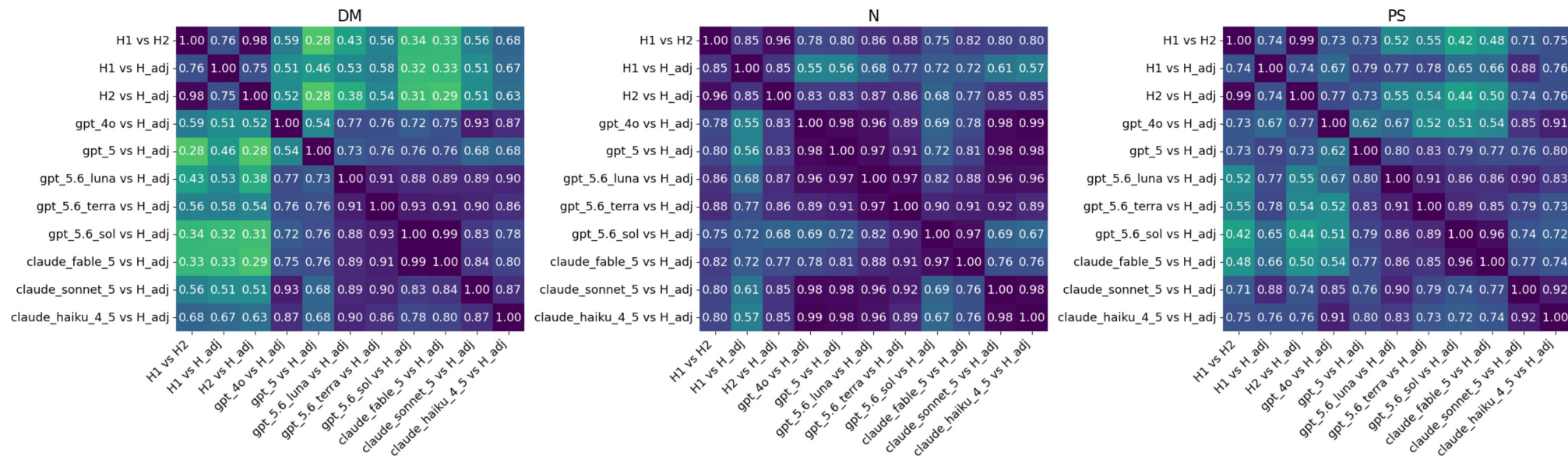
WHAT THE MODEL SHOWS

Overall, demographic effects are not significant predictors of human–AI agreement.

6

LLMs Exhibit Different Miscoding Patterns

HOW SIMILAR ARE THE MIS-CODING PATTERNS? CORRELATION BETWEEN CONFUSION MATRICES, BY TASK



Each cell is the correlation between two raters' confusion matrices — 1.00 means they make the same mistakes on the same categories. Rows and columns are human raters and LLMs; the three panels are the DM, N and PS tasks.

Humans err similarly

Human–human confusion patterns correlate well (because not independent from the adjudicated code)

Some LLMs err similarly but not all

Some model–model pairs correlate highly but some do not.

Human and LLMs do not correlate well

Model–human correlations drop to ~0.3–0.6, so inter-model agreement reflects shared model bias, not valid coding.

Summary

We systematically investigated six key questions about coding communication data with LLMs - on ~8,000 chat turns from three collaborative tasks, coded with a typical CPS framework.

KEY TAKEAWAYS

- 1 **LLMs code communication as accurately as trained human raters.**
- 2 **Newer reasoning and agentic models add little.**
- 3 **Feeding mis-coding info back into the prompt brings no significant gain.**
- 4 **Hierarchical prompting makes coding worse, not better.**
- 5 **Human–AI agreement is consistent across demographic groups.**
- 6 **LLMs do not miscode the way human raters do.**

LIMITATION

Findings are bound to coding tasks similar to this framework and data; generalizing to other coding schemes or contexts needs caution.

ONGOING AND FUTURE WORK

Several efforts are underway - more is coming.

Thank You

Contact: jhao@ets.org
for additional details



References

- Hao, J., Cui, W., Kyllonen, P., & Kerzabi, E. (2026). Automated coding of communication data using LLM: Consistency across subgroups. *Journal of Educational Measurement*, 63(2), e70049.
- Cui, W., Hao, J., Jiang, Y., Kyllonen, P. C., & Kerzabi, E. (2026). Automated coding of communication data using large language models: A comparison of hierarchical and direct prompting strategies. *Frontiers in Education*, 11, 1764154.
- Hao, J., Cui, W., Kyllonen, P., Kerzabi, E., Liu, L., & Flor, M. (2025). Automated coding of communications in collaborative problem-solving tasks using ChatGPT. *Journal of Educational Measurement*, 62(4), 809–837.
- Hao, J. (2025). AI for educational assessment: Opportunities and challenges through the lens of evidence-centered design and measurement values. In *AI and VR for Higher Education*, Springer Nature.
- Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16–29.
- Hao, J., & Mislevy, R. J. (2022). A data science perspective on computational psychometrics. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational Psychometrics* (pp. 133–158). Springer.
- Hao, J., Chen, L., Flor, M., Liu, L., & von Davier, A. A. (2017). *CPS-Rater: Automated sequential annotation for conversations in collaborative problem-solving activities* (Research Report No. RR-17-07). Educational Testing Service.