

# Cognitive Engagement in GenAI Tutor Conversations: At-scale Measurement and Impact on Learning

Kodi Weatherholtz<sup>1</sup>, Kelli Millwood Hill<sup>1</sup>, Kristen DiCerbo<sup>1</sup>, Walt Wells<sup>1</sup>,  
Phillip Grimaldi<sup>1</sup>, Maya Miller-Vedam<sup>1</sup>, Charles Hogg<sup>1</sup>, and Bogdan Yamkovenko<sup>1</sup>

<sup>1</sup>Khan Academy,

Correspondence: [kodi@khanacademy.org](mailto:kodi@khanacademy.org)

## Abstract

We developed and validated a scalable LLM-based labeler for classifying student cognitive engagement in GenAI tutoring conversations. Higher engagement levels predicted improved next-item performance, though further research is needed to assess distal transfer and to disentangle effects of continued tutor use from true learning transfer.

## 1 Introduction

Student engagement is a key predictor of learning outcomes, but not all engagement is equally beneficial. Behavioral engagement (e.g., time-on-task, number of tasks completed) captures participation, whereas cognitive engagement refers to the quality of mental effort devoted to sense-making. Theory and evidence consistently link cognitively productive behaviors—such as self-explaining, elaborating, and drawing inferences—to superior learning, beyond surface-level participation alone (Sinatra et al., 2015; Greene and Azevedo, 2010).

Measuring cognitive engagement at scale remains challenging. Conventional learning-analytics behavioral measures (e.g., time on task, retry rates) are convenient and can correlate with learning outcomes. But they provide limited insight into the metacognitive and inferential processes that underpin knowledge abstraction, integration and transfer (Winne, 2020). By contrast, text-based GenAI tutoring produces rich discourse traces of students' questions, explanations, and reasoning. These traces enable rubric-guided text mining—including LLM-as-judge approaches—to operationalize cognitive engagement directly within authentic tutoring conversations.

The present research has two main aims:

1. **At-scale labeling.** To develop and validate a reliable and scalable procedure for labeling

student cognitive engagement in GenAI math tutoring conversations.

2. **Impact evaluation.** To measure the impact of different levels of *sustained* cognitive engagement on learning outcomes, focusing on near transfer to subsequent performance on the same skill. (Distal transfer to mixed-skill practice or assessments is of interest but out of scope.)

## 2 Background

### The ICAP framework for cognitive engagement.

The ICAP framework is a central account of how the depth of cognitive engagement shapes learning. This framework outlines four levels of engagement—Interactive » Constructive » Active » Passive—and predicts deeper learning as learners progress from *passive* activities (e.g., listening, reading) to *active* behaviors (e.g., highlighting/underlining, copying steps, selecting answers), to *constructive* generation (e.g., self-explaining, making inferences, creating concept maps/diagrams), and ultimately to *interactive* co-construction of knowledge with a peer (e.g., dialogue where partners question, explain, and build on each other's contributions) (Chi and Wylie, 2014).

A central prediction of the ICAP framework is that *constructive* engagement facilitates robust knowledge abstraction and transfer. The *interactive* level refers to dyads in which both peers engage constructively, generating and amplifying learning benefits for both participants. This contrasts with, for instance, a *passive-constructive* dyad, in which one partner generates constructive ideas while the other primarily listens or affirms without reciprocally elaborating or extending the reasoning. In the latter case, the constructive participant might benefit from the dialogue, but the passive participant is predicted to benefit minimally, if at all (Chi and Menekse, 2015).

The ICAP framework was initially introduced to characterize the affordances of learning activities (Chi, 2009) and has since been used extensively to analyze engagement in *peer-to-peer* contexts (e.g., (Vosniadou et al., 2023; Stump et al., 2017; Chi et al., 2017; Hsiao et al., 2022; Chi et al., 2018)). Empirical applications typically (a) compare activity types hypothesized to induce different engagement levels, or (b) use labor-intensive, human coding of classroom observations/transcripts to identify ICAP behaviors in situ. These approaches clarified which activities tend to promote deeper learning, but they pose two limitations for large-scale analytics: the unit of analysis is often coarse (activity episodes rather than specific turns of talk), and expert-level human annotation does not scale.

**Adapting ICAP to student–tutor GenAI conversations.** Text-based GenAI tutoring changes both the data available for analysis and the interactional context. On the data side, it produces detailed discourse traces—student turns and tutor responses—that can be examined at the level of individual contributions rather than entire activities. This granularity makes it possible to apply rubric-guided text mining approaches to authentic tutoring conversations, enabling scalable labeling methods (e.g., LLM-as-judge).

On the interactional side, GenAI tutoring differs fundamentally from peer collaboration. Whereas the ICAP framework’s notion of *interactive* engagement emphasizes co-construction among peers of similar ability, GenAI tutoring is structured as a novice–expert dialogue. The GenAI tutor is, by design, a constructive partner with expert-level domain knowledge. What varies is how the learner contributes to the dialogue.

Although much attention has been given to tutor behaviors—such as eliciting reasoning, correcting misconceptions, and calibrating when to provide or withhold information (VanLehn, 2011; Koedinger et al., 2012; Graesser et al., 2012; Chi et al., 2001)—the more fundamental issue for whether and how much learners benefit from intelligent tutoring may be how learners engage in these dialogues: whether they participate passively, actively, or constructively.

**Present approach and study roadmap.** Study 1 adapts the ICAP framework for the student–GenAI tutor setting by developing and validating a rubric-guided, LLM-as-judge procedure for labeling student cognitive engagement at scale. Study 2 then

examines how within-learner variation in cognitive engagement relates to learning, as measured by near transfer to subsequent performance on the next item attempt on the same skill (knowledge component). These analyses use student panel data (multiple tutoring conversations and corresponding item attempts per learner) and employ models with rich fixed effects to remove confounds due to time-invariant aspects of student ability, variation in skill difficulty, and local practice context.

### 3 Study 1: At-scale labeling

#### 3.1 Methods

**Operationalization.** We extended the ICAP framework to student–tutor (novice–expert) dialogues by classifying student contributions as passive, active, or constructive (see Table 1 for a summary of operational definitions of these levels). In brief, *passive* turns reflect perfunctory acknowledgments of the tutor’s information (e.g., yes/no answers) or answer-mining behavior. *Active* turns typically involve the learner mechanically following procedural guidance from the tutor without contributing new information. And *constructive* turns involve the learner adding new information by asking targeted questions, explaining reasoning, making inferences, or engaging in other behaviors that deepen the conversation and show ownership of learning.

**Expert coding.** Two human experts applied this rubric to a non-random coverage sample of conversations ( $n = 106$ ), selected to reflect a diverse range of student and tutor discourse behaviors. While each student turn in a conversation can be classified independently, the expert-labeling effort focused on holistically classifying the dominant pattern of student cognitive engagement across their turns in a conversation (see examples in Table A1). After the experts labeled the conversations, we measured inter-rater reliability, resolved disagreements, and refined the rubric until high agreement was achieved.

**LLM-as-judge labeling.** We then converted the final rubric to an LLM-as-judge prompt (Table A2) and used OpenAI’s Batch API (model snapshot = gpt-4o-2024-05-13) to machine-label both the human-labeled threads and a larger random set. We iterated on the prompt (e.g., few-shot examples, instruction refinements) until suitable human/LLM agreement was reached.

PASSIVE	ACTIVE	CONSTRUCTIVE	INTERACTIVE
The Tutor fully drives the conversation, with the Learner passively following along, <b>guessing, missing for the answer, or providing perfunctory inputs</b> (e.g., ‘yes’, ‘no’, ‘idk’, ‘help’)	The Tutor drives the conversation, but the Learner is actively involved. This often looks like the Learner <b>mechanically following the Tutor’s procedural guidance but without contributing new information</b> or without demonstrating deeper exploration / questioning.	The Learner takes ownership of the problem solving by <b>asking targeted questions, explaining their thinking, making inferences</b> , generalizations or connections to other material, or generally building on the Tutor’s guidance.	Only applies in peer-to-peer situations where learners are <b>co-constructing</b> knowledge. Does not apply when learners interact with an (ostensibly) “all knowing” other / GenAI tutor.

Table 1: Operational definitions of student cognitive engagement in GenAI tutoring conversations (adapted from ICAP framework).

Inter-rater agreement							
<i>For human-human and human-LLM rater pairs</i>							
	N threads	Both raters agree	Counts				Krippendorff’s $\alpha$
			TP	TN	FP	FN	
<b>Human expert #1 vs. #2</b>							
Passive	107	93%	16	83	2	6	0.755
Active	107	85%	28	63	12	4	0.667
Constructive	107	88%	40	54	4	9	0.754
<b>Human expert #1 vs. LLM-judge</b>							
Passive	344	90%	142	167	23	12	0.796
Active	344	87%	84	214	25	21	0.688
Constructive	344	94%	56	266	6	16	0.796

Table 2: Summary of human/human and human/LLM-judge agreement.

### 3.2 Results

Inter-rater agreement is shown in Table 2. Overall agreement exceeded 84% for both human/human and human/LLM pairs. Krippendorff’s alpha ranged from 0.67 to 0.80, where values above 0.6 indicate good agreement and above 0.8 indicate very strong agreement. Human/LLM agreement was slightly higher than human/human agreement, suggesting the LLM judge performed comparably to human experts, if not slightly better.

### 3.3 Discussion

The LLM-as-judge approach for labeling student cognitive engagement from tutoring discourse transcripts achieved expert-level agreement while outputting structured data and rationales. This enables both high-throughput labeling and downstream analytics (e.g., linking cognitive engagement to outcomes), complementing traditional

learning-analytics pipelines that rely on click-stream data and behavioral measures.

## 4 Study 2: Impact evaluation

### 4.1 Data and samples

The global sample from which we derived our causal impact analysis samples is summarized in Table 3. Specifically, we curated Khan Academy platform telemetry data and school district administrative data for students in six US public school districts who had access to Khanmigo – Khan Academy’s GenAI tutor – for the 2024-2025 SY. The platform data were limited to students ( $n = 9,129$ ) with non-zero use of Khanmigo as a math tutor during the Fall and early Spring semesters (2024-08-01 to 2025-03-01). For these students, we curated question performance for all Khan Academy math items attempted during the eligibility window, totalling nearly 9 million question

attempts, along with the conversation text of all Khanmigo threads that took place on those items (about 200K threads). We ran all of the tutoring conversations through a series of machine labelers – including the cognitive engagement labeler described above – to extract discourse features.<sup>1</sup>

## 4.2 Within-learner panel creation

To estimate near transfer while controlling for prior skill knowledge, we required sustained (multi-turn) threads and at least one prior and one subsequent attempt on the same knowledge component. We created two learner-level panels: Panel 1 includes learners with at least one qualifying thread in each of Passive/Active/Constructive; Panel 2 includes learners with qualifying threads in at least two of the three levels. This yields roughly 200 learners/5k threads (Panel 1) and 2k learners/35k threads (Panel 2).

Table 4 presents a descriptive summary of the panels. The rate of Constructive threads is low – an inherent property of these data. The analysis threads are by definition ‘sustained’ / multi-turn conversations: on average, there are 5 back-and-forth turns between the learner and the GenAI Tutor. The vast majority (>90%) of threads started before the learner attempted the math question (i.e., non-punitive pre-submission help); hence we expect answer correctness to be biased upwards on these items. Further, there is a non-zero rate (variable by student cognitive engagement) at which the Tutor (undesirably) reveals the final answer to the problem, which further biases correctness upwards. For knowledge transfer analysis, we examine learner performance on the next skill item. Ideally, that next item would not involve use of the Tutor. Deriving a sample with that restriction is not feasible with these data (given that about two-thirds of treatment items are followed by Tutor use). We adjust for this property of the data in the causal impact analyses; but there remains a limitation that what appears to be a treatment effect that generalizes to future performance could be explained (in part, or in full) by the future Tutor use.

<sup>1</sup>Other labelers included deterministic labelers for linguistic complexity, reading ease, and sentiment; an LLM-judge labeler for whether the tutor revealed the final answer without the student doing the work first; and an embedding-based machine labeler for the Tutor’s degree of repetitiveness during a dialogue.

## 4.3 Statistical methods

To examine whether deeper cognitive engagement is associated with improved learning, we modeled student performance on the next item practiced within the same skill. Specifically, we fit a fixed effects logistic regression:

$$\begin{aligned} \text{logit Pr}(Y_{s,k,i+1} = 1) = & \beta_1 \text{Engagement}_{s,k,i} + \\ & \alpha_{\text{student}(s)} + \\ & \alpha_{\text{skill}(k)} + \\ & \alpha_{\text{next\_item\_had\_tutor}(s,k,i+1)} + \\ & \alpha_{\text{prev\_item\_had\_tutor}(s,k,i-1)} + \quad (1) \\ & \alpha_{\text{prev\_item\_credited}(s,k,i-1)} + \\ & \alpha_{\text{incoming\_skill\_mastery}(s,k)} + \\ & \alpha_{\text{tutoring\_conversation\_length}(i)} + \\ & \varepsilon_{s,i} \end{aligned}$$

This specification models the probability that learner ( $s$ ) answers the subsequent item ( $i + 1$ ) on the same skill ( $k$ ) correctly as a function of their observed engagement with the GenAI tutor on the current item ( $i$ ). The key treatment variable is the level of cognitive engagement (Passive, Active, or Constructive) exhibited by the student during the tutoring exchange. By estimating a logistic regression with an extensive set of fixed effects ( $\alpha$  terms), the model seeks to isolate within-student, within-skill variation in performance that is plausibly attributable to differences in cognitive engagement level.

The inclusion of student fixed effects ( $\alpha_{\text{student}}$ ) accounts for all stable (time-invariant) learner characteristics that might otherwise confound the estimates, such as baseline ability, motivation, or demographic background. Skill fixed effects ( $\alpha_{\text{skill}}$ ) absorb persistent differences in the inherent difficulty of skills, ensuring that comparisons are not biased by heterogeneity in task complexity. Additional item-level controls capture contextual features of practice: whether the previous or subsequent item involved tutor use, whether the prior item was credited as correct, the learner’s incoming mastery estimate for the skill, and the number of conversational interactions within the thread. Together, these fixed effects substantially reduce between-student and between-skill variation, narrowing the identifying variation to comparisons within the same learner–skill trajectory.

The implicit assumption of this modeling approach is that, conditional on the fixed effects and included controls, variation in observed engagement level is orthogonal to other unobserved deter-

	N students	N question attempts	N question attempts with tutor use	% attempts with tutor use	% off topic	% passive	% active	% constructive
<b>Total</b>	9,129	8,853,983	211,229	2.4	5.5	72.2	21.6	0.7
<b>By district</b>								
District 1	5,834	7,971,201	173,806	2.2	5.9	74.5	19.1	0.5
District 2	1,987	268,099	12,516	4.7	4.3	62.5	32.1	1.2
District 3	257	128,332	7,403	5.8	2.5	60.7	35.5	1.3
District 4	500	108,028	6,743	6.2	1.2	63.4	33.9	1.4
District 5	216	211,523	6,202	2.9	7.1	61.5	30.0	1.3
District 6	335	166,800	4,559	2.7	1.7	58.0	35.4	5.0

Table 3: Summary of the global sample and by-district breakdown.

ICAP Level	Learners	Threads	Thread Characteristics								Surrounding Context					
			% of Total Threads	Skills	In a Completed Task	Median Position in Task	Median Thread Length	Median User Word Count	Thread Started Before Question Attempt	Tutor Revealed Final Answer	% Correct			Tutor Use		
											Prev Item	Current Item	Next Item	Prev Item	Current Item	Next Item
<b>Panel 1: Each student has all three P/A/C levels</b>																
Passive	191	2,257	47%	705	70%	2	5	1.0	94%	20%	71%	74%	74%	57%	100%	64%
Active	191	2,215	46%	689	80%	2	5	4.0	94%	13%	76%	83%	79%	59%	100%	66%
Constructive	191	375	8%	271	82%	3	5	10.5	95%	6%	80%	87%	82%	61%	100%	62%
Total	191	4,847	100%	954	75%	2	5	2.0	94%	16%	74%	79%	77%	58%	100%	65%
<b>Panel 2: Each student has at least two of the three P/A/C levels</b>																
Passive	2,116	23,278	68%	1,221	65%	2	5	1.0	93%	18%	69%	70%	69%	51%	100%	61%
Active	2,158	10,548	31%	1,158	74%	2	5	4.0	93%	11%	72%	81%	75%	53%	100%	61%
Constructive	253	503	1%	343	81%	3	5	9.2	92%	7%	79%	82%	81%	56%	100%	57%
Total	2,168	34,329	100%	1,362	68%	2	5	1.5	93%	15%	70%	73%	71%	52%	100%	61%

Table 4: Descriptive summary of analysis samples (Panel 1: learners with all three engagement levels; Panel 2: learners with at least two engagement levels). Includes thread characteristics, tutor interaction patterns, correctness, and tutor use on adjacent items.

minants of performance on the next item. In other words, after holding constant learner ability, skill difficulty, prior correctness, and local tutoring context, remaining differences in next-item correctness are attributed to differences in engagement mode. While this strengthens the rigor of inference relative to pooled or random-effects models (Angrist & Pischke, 2009), the design remains observational: unmeasured time-varying factors such as transient motivation, fatigue, or classroom context could still bias estimates.

#### 4.4 Results

The regression results are reported in Table 5. We show cumulative stepwise addition of fixed effects to highlight how the treatment estimates changed as the amount of within-variation was controlled. Figure 1 shows the estimated marginal contrasts for the full/saturated regression model for each sample after converting the log-odds estimates to probabilities. We observe a significant stair-step effect:

learners are more likely to answer the next skill item correctly (by 4-5 percentage points) after a sustained Active tutoring conversation, relative to sustained Passive, and even more likely (10ppt) after sustained Constructive.

Note that while the results trended toward Constructive > Active, the estimate for this contrast was not statistically significant. The base rate of constructive threads is very low, resulting in an imprecise measure, as seen in the wide confidence intervals in Figure 1.

#### 4.5 Discussion

Our analyses indicate that constructive cognitive engagement with a GenAI tutor is associated with substantively better performance on subsequent skill practice, compared to passive engagement, highlighting the promise of scalable, log-based evaluation of tutor effects. Still, given the nonrandom treatment, low base rates, and outcome scope, we temper causal claims and detail limitations and



Table 5: Fixed-effect logistic regression results for the impact of cognitive engagement on knowledge transfer (i.e., correctness on next item within skill).

	Student Panel 1							Student Panel 2						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
<b>Sustained Active</b>	0.29 *** [0.14, 0.44]	0.30 ** [0.07, 0.52]	0.27 * [0.05, 0.49]	0.27 * [0.05, 0.49]	0.24 * [0.02, 0.46]	0.20 + [-0.05, 0.43]	0.20 [-0.05, 0.45]	0.30 *** [0.23, 0.37]	0.32 *** [0.25, 0.40]	0.29 *** [0.22, 0.36]	0.28 *** [0.21, 0.35]	0.27 *** [0.20, 0.35]	0.24 *** [0.17, 0.32]	0.27 *** [0.19, 0.34]
<b>Sustained Constructive</b>	0.51 ** [0.18, 0.84]	0.67 * [0.15, 1.21]	0.65 * [0.11, 1.18]	0.63 * [0.10, 1.17]	0.61 * [0.08, 1.15]	0.55 + [-0.01, 1.10]	0.72 + [-0.02, 1.46]	0.49 *** [0.21, 0.77]	0.71 *** [0.35, 1.08]	0.67 *** [0.31, 1.04]	0.65 *** [0.28, 1.02]	0.63 *** [0.26, 1.00]	0.57 ** [0.19, 0.94]	0.61 ** [0.21, 1.01]
<i>Model statistics</i>														
Num. Obs.	4,706	3,408	3,408	3,408	3,408	3,408	3,403	33,098	32,022	32,022	32,022	32,022	32,022	32,015
R <sup>2</sup>	0.069	0.178	0.199	0.202	0.232	0.232	0.240	0.087	0.145	0.163	0.164	0.168	0.198	0.199
R <sup>2</sup> Adj.	0.000	-0.105	-0.085	-0.083	-0.054	-0.054	-0.058	-0.007	-0.001	0.017	0.018	0.021	0.051	0.050
R <sup>2</sup> Within	0.004	0.004	0.004	0.004	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.002	0.002	0.002
R <sup>2</sup> Within Adj.	0.003	0.003	0.003	0.003	0.002	0.001	0.001	0.002	0.003	0.003	0.002	0.002	0.002	0.002
RMSE	0.41	0.41	0.40	0.40	0.40	0.39	0.39	0.43	0.42	0.41	0.41	0.41	0.40	0.40
<i>Fixed effects included (stepwise)</i>														
FE: student (kaid)		X	X	X	X	X	X	X	X	X	X	X	X	X
FE: skill_id			X	X	X	X	X	X	X	X	X	X	X	X
FE: next_item_on_skill_used_km				X	X	X	X	X	X	X	X	X	X	X
FE: prev_item_on_skill_used_km					X	X	X	X	X	X	X	X	X	X
FE: prev_item_on_skill_is_credited						X	X	X	X	X	X	X	X	X
FE: skill_fpm_level_before						X	X	X	X	X	X	X	X	X
FE: n_interactions							X							X

Notes: +  $p < .10$ , \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ . Each cell shows the coefficient (top) and 95% CI (bottom). Highlighted columns are fully saturated specifications, per Equation 1.

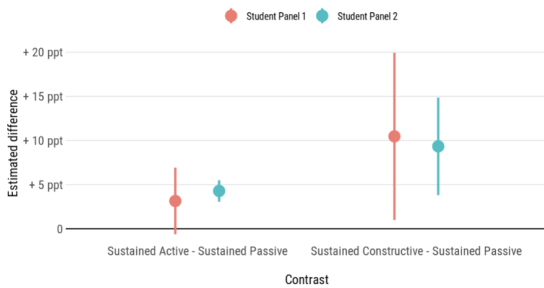


Figure 1: Estimated marginal contrasts from fixed effects regressions for the probability of answering the next skill item correct following 'sustained' Active or Constructive cognitive engagement relative to Passive engagement.

threats to validity.

**Observational design.** The treatment is learners' observed engagement level, not random assignment. Despite rich fixed effects that account for time-invariant within-learner and within-skill properties, unobserved confounds remain possible (e.g., transient motivation).

**Subsequent tutor use.** About two-thirds of treated items are followed by tutor use on the next item. Models adjust for this, but residual bias is plausible; what looks like transfer may partly reflect continued support.

**Base-rate imbalance.** Constructive threads are rare (<1%), limiting precision and external validity across subgroups/skills.

**Outcome scope.** We evaluated near transfer (next item on the same skill). Whether gains persist to mixed-skill practice or distal assessments is left for future work.

## 5 Conclusions

We developed a rubric-grounded LLM-as-judge that can validly classify cognitive engagement in GenAI tutor conversations at scale, enabling new forms of learning analytics that attend to the quality of student thinking. In large-scale field data, sustained Active and especially Constructive engagement were associated with higher odds of success on the next item in the same skill (near transfer), even after rigorous controls. Future work should test whether these effects extend to distal outcomes and persist without continued tutor support. Improving tutor designs to increase constructive moves is a promising path to better learning with GenAI tutors.

## 5.1 References

### Acknowledgments

We are grateful to Michelene Chi and Kurt Van Lehn for feedback on the rubric we developed to classify student cognitive engagement in text-based tutoring interactions.

### References

- Michelene T. H. Chi. 2009. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1):73-105.
- Michelene T. H. Chi, Joshua Adams, Elizabeth B. Bogusch, Colleen Bruchok, Seokmin Kang, Matthew Lancaster, Roger Levy, Katherine McEldoon, Glenda S. Stump, Ruth Wylie, Dongchen Xu, and David L. Yaghmourian. 2018. [Translating the icap theory of cognitive engagement into practice](#). *Cognitive Science*, 42(6):1777-1832.

- Michelene T. H. Chi, Seokmin Kang, and David L. Yaghmourian. 2017. [Why students learn more from dialogue- than monologue-videos: Analyses of peer interactions.](#) *Journal of the Learning Sciences*, 26(1):10–50.
- Michelene T. H. Chi and Muhsin Menekse. 2015. Dialogue patterns that promote learning. In Lauren B. Resnick, Christa S. Asterhan, and Sherice N. Clarke, editors, *Socializing Intelligence through Academic Talk and Dialogue*, chapter 21, pages 263–274. American Educational Research Association, Washington, DC.
- Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. 2001. [Learning from human tutoring.](#) *Cognitive Science*, 25(4):471–533.
- Michelene T. H. Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243.
- Arthur C. Graesser, Mark W. Conley, and Andrew Olney. 2012. [Intelligent tutoring systems.](#) In Karen R. Harris, Steve Graham, and Tim Urdan, editors, *APA Educational Psychology Handbook, Vol. 3: Application to Learning and Teaching*, pages 451–473. American Psychological Association.
- Jeffrey A. Greene and Roger Azevedo. 2010. The measurement of learners’ self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist*, 45(4):203–209.
- Jo-Chi Hsiao, Ssu-Kuang Chen, Wei Chen, and Sunny S. J. Lin. 2022. [Developing a plugged-in class observation protocol in high-school blended stem classes: Student engagement, teacher behaviors and student–teacher interaction patterns.](#) *Computers & Education*, 178:104403.
- Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2012. [The knowledge-learning-instruction framework: Bridging the science–practice chasm to enhance robust student learning.](#) *Cognitive Science*, 36(5):757–798.
- Gale M. Sinatra, Benjamin C. Heddy, and Doug Lombardi. 2015. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1):1–13.
- Glenda S. Stump, Na Li, Seokmin Kang, David Yaghmourian, Dongchen Xu, Joshua Adams, Katherine L. McEldoon, Matthew Lancaster, and Michelene T. H. Chi. 2017. [Coding dosage of teachers’ implementation of activities using icap: A video analysis.](#) In *Promoting Spontaneous Use of Learning and Reasoning Strategies*. Routledge.
- Kurt VanLehn. 2011. [The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems.](#) *Educational Psychologist*, 46(4):197–221.
- Stella Vosniadou, Michael J. Lawson, Erin Bodner, Helen Stephenson, David Jeffries, and I. Gusti Ngurah Darmawan. 2023. [Using an extended icap-based coding guide as a framework for the analysis of classroom observations.](#) *Teaching and Teacher Education*, 128:104133.
- Philip H. Winne. 2020. Learning analytics for self-regulated learning. In *Handbook of Learning Analytics*, 2 edition. Society for Learning Analytics Research (SoLAR), Vancouver, BC.

## A Supplementary Tables

The table shows a proportional relationship between the number of pounds of bananas purchased and the total cost of bananas.

Bananas	
Number of pounds	Total cost (dollars)
3	1.47
5	2.45
9	4.41
?	?

A row of values is missing in the table.

Which numbers of pounds of bananas and total costs of the bananas could be used as the missing values in the table?

Choose 3 answers:

- A Pounds of bananas: 2  
Total cost: \$0.98
- B Pounds of bananas: 7  
Total cost: \$4.45
- C Pounds of bananas: 6  
Total cost: \$2.94
- D Pounds of bananas: 1  
Total cost: \$0.54
- E Pounds of bananas: 8  
Total cost: \$3.92

Target math problem for conversations below.

Correct answers = A, C, and E.

Student cognitive engagement	Role	Text
<b>Active</b>  The learner makes a generic help request (passive) and then mechanically follows the tutor's procedural guidance.	Student	help
	Tutor	Let's solve this together! First, what do you notice about the relationship between the pounds of bananas and their total cost in the table? How can we use this to find the missing values?
	Student	find the total cost of 1 pound
	Tutor	Great idea! Using the table, can you calculate how much 1 pound of bananas costs?
	Student	\$0.49
	Tutor	That's correct! Each pound costs \$0.49. How can you use the cost per pound to find the total cost for different numbers of pounds?
	Student	multiply
	Tutor	Exactly! By multiplying the cost per pound (\$0.49) by the number of pounds, you can find the total cost. Can you try this for the options given and see which ones fit the pattern?
	Student	2 x \$0.49 is 0.98
	Tutor	Great job! You found that 2 pounds cost \$0.98. Now, try using the same method to calculate the total cost for the other options. Which ones match the pattern?
<b>Constructive</b>  The learner summarizes the problem, follows the tutor's guidance, errs on the first substep, gets feedback, retries while explaining their reasoning, and then solves the problem.	Student	The table says 3 pounds would be \$1.47. 5 pounds cost \$2.45 and 9 pounds cost \$4.41. But we have to find the proportional relationship.
	Tutor	You're doing well! To find the proportional relationship, calculate the cost per pound by dividing the total cost by the number of pounds for each given data point. What do you find when you do this?
	Student	So if I divide all of the pounds by the cost it would be 2.04 but we have to find the actual cost every banana.
	Tutor	It looks like there might be a small mix-up. To find the cost per pound, you should divide the total cost by the number of pounds, not the other way around. Try calculating it again using this method. What do you find?
	Student	If I divide each of the costs by the number of pounds, it is exactly 0.49
	Tutor	Great! Now that you know the cost per pound is \$0.49 per pound, you can use this to check which options maintain this consistent cost. Which options do you think match this cost per pound?
	Student	So A C E would be exactly like \$0.49 because when I divide all of them it would be \$0.49
Tutor	You've got it! Options A, C, and E all maintain the consistent cost of \$0.49 per pound. Great work on verifying the proportional relationship! If you have any more questions, feel free to ask.	

Table A1: Example student-tutor conversations.



---

**Instructions**

You are a Critic evaluating a Student’s cognitive engagement during a back-and-forth conversation with a Tutor. Your task is to assess the Student’s cognitive engagement in the conversation, using the rubric provided. **Think before you respond.** Read the entire conversation first; then use the rubric to provide a holistic judgment of the student’s contributions to the conversation. You must output your evaluation as a **nested JSON object** as defined below.

**Cognitive Engagement Rubric**

**Level: Constructive** – The Student is actively and constructively involved in the conversation. The Student takes ownership of problem-solving by asking targeted questions about the content, building on the Tutor’s guidance, explaining their thinking, and/or making inferences, generalizations or connections to other material.

**Rating:** 8–10

**Indicators:**

- **High-specificity help requests:** The Student asks targeted questions that deepen conceptual understanding of the problem (e.g., “Why do we use a negative sign in this equation?”, “How is a diamond different from a square?”).
- **Substantive responses:** The Student builds on the Tutor’s responses with further explanation, reasoning, or problem-solving steps (e.g., “I see, we multiply here because it’s a percentage increase, not an absolute number.”).
- **Deep engagement with feedback:** The Student revises their approach, while also asking follow-up questions or explaining how they modified their thinking (e.g., “I see why my approach didn’t work—so I should have used the distributive property?”).
- **Deep ownership of learning:** The Student actively guides the conversation by asking insightful questions, proposing alternative methods, or pointing out potential mistakes (e.g., “Could we also solve this using substitution?”).

**Level: Active** ...

**Level: Passive** ...

**Level: Off-topic** ...

**How to Use the Rubric:**

The rubric defines behavioral indicators for four levels of cognitive engagement (off topic, passive, active, and constructive) along a numeric scale from 1–10. Determine which level best describes the Student’s contributions to the conversation. And then determine a numeric rating based on how well the conversation fits that level. The numeric ratings are not 1:1 with the levels: for example, the Passive range is from 2–4 and Active is from 5–7. These numeric ranges allow room to distinguish conversations where the Student is, for instance, exclusively Passive (rating=2) from conversations where the Student demonstrates a mixture of Passive and Active contributions (which could be rated 3–5 depending on the mixture).

**Output Format:**

You must output a **nested JSON object** where the key is “icap” for the ICAP Framework, and the value is an object containing:

1. **rating:** An integer rating (1–10).
2. **evidence:** A numbered list of instances from the Student’s behavior in the conversation that justify the rating.
3. **off\_topic\_count:** An integer denoting the number of Student turns that are best classified as “off topic”.
4. **passive\_count:** An integer denoting the number of Student turns that are best classified as “passive”.
5. **active\_count:** An integer denoting the number of Student turns that are best classified as “active”.
6. **constructive\_count:** An integer denoting the number of Student that are best classified as “constructive”.

If a type of engagement doesn’t occur, still include its object: e.g., “constructive\_count: 0”. The sum total of passive\_count, active\_count, and constructive\_count values must equal (NEVER EXCEED) the total number of Student turns in the conversation.

**Output Example:**

```
{
  "icap": {
    "rating": 5,
    "evidence": "(1) Student provides surface-level answers. (2) Student follows
      steps mechanically with no elaboration or reflection.",
    "off_topic_count": 0,
    "passive_count": 1,
    "active_count": 3,
    "constructive_count": 0
  }
}
```

**Example Scenarios** [redacted few-shot prompting]