Bridging the Novice-Expert Gap via Models of Decision-Making

Rose E. Wang* Qingyang Zhang Carly Robinson Susanna Loeb Dorottya Demszky

*rewang@cs.stanford.edu, Stanford University

Background

- **Motivation:** The pandemic has caused a historical low in K-12 mathematics performance.
- **Challenge of scaling high-quality tutoring:** Due to growing demand, many tutoring platforms employ novice tutors who, unlike experienced educators, struggle to address student mistakes and thus fail to seize prime learning opportunities.
- **Novice tutors** have content knowledge (+), but struggle with writing pedagogically aligned responses (-).
- **Large language models (LLMs)** generate coherent text at scale (+) but have questionable content & pedagogical knowledge (-).
- **Experienced math teachers** have content & pedagogical knowledge (+) but are hard to scale (-).

Key Question: Can we model how experts *think* to improve LLM performance and scale high-quality tutoring?

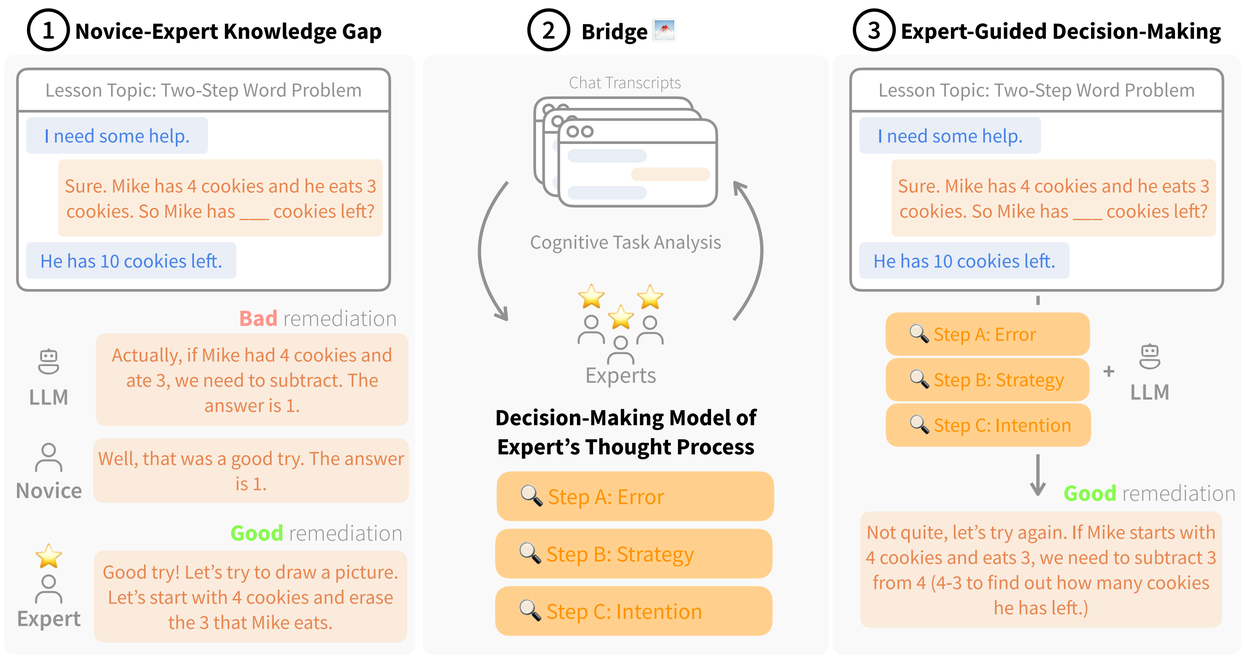
Contributions

Method. Bridge, a framework that breaks down experts' hidding decision processs in remediating student math mistakes (a key learning opportunity).

Dataset. 700 examples with expert decisions and responses, across 120 different math topics.

Evaluations. Bridge improves LLM performance on remediation!

Bridge: Method for Modeling Expert Decision-Making



Human expert decisions paths are extremely *diverse*.

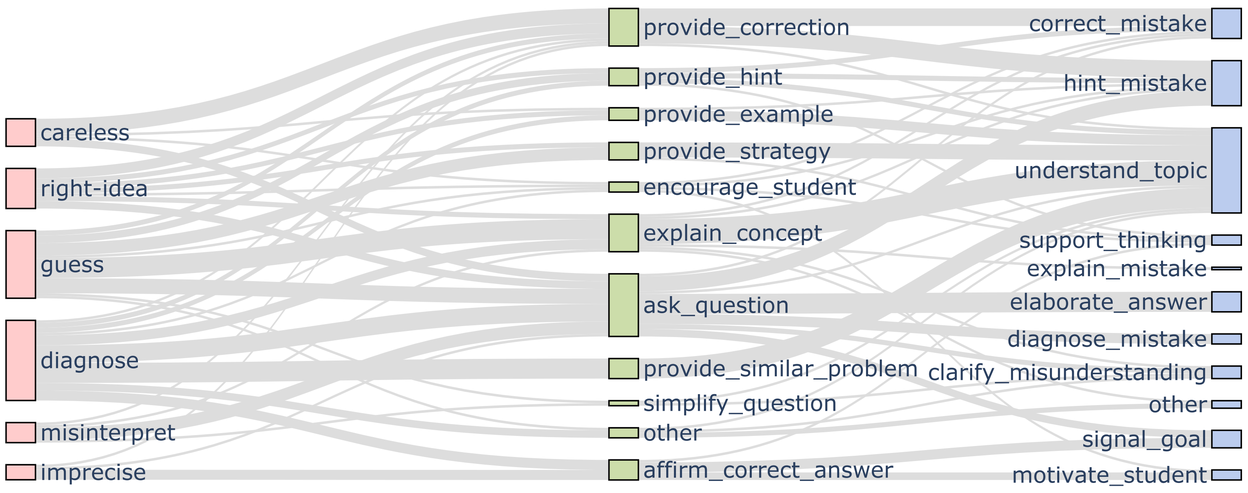


Figure 1. Human Expert Decision Paths of student error, strategy and intention.

Domain Experts

- 4 certified math teachers from diverse demographics in terms of gender and race;
- Each with 8+ years of teaching experience including public schools, Title 1 schools, and charter schools;
- Paid $50/hr for framework; $40/hr for annotation.

Data Sources

- Tutoring chat transcripts with elementary school students from Southern school district serving serving > 30k students;
- 3rd-8th grade students;
- 120 different math topics, including "Word Problems", "Order of Operations", and "Graphing on a Coordinate Grid";
- Majority of schools classified as Title I and $\frac{3}{4}$ students identify as Hispanic/Latinx.

Results

LLMs benefit from Bridge decision-making.

Method		Prefer	Useful	Care	Not Robot	Overall
Bridge	Model c_r					
	Expert	**1.26**	**1.19**	**0.86**	**0.78**	**1.02**
-	GPT-3.5	0.47	0.47	−0.04	0.23	0.28
-	GPT-4	0.54	0.54	0.50	0.47	0.51
Expert	GPT-3.5	0.65	0.58	−0.04	0.59	0.45
Expert	GPT-4	**0.95**	**0.97**	**0.70**	**0.70**	**0.83**
Self	GPT-3.5	0.36	0.33	−0.17	0.15	0.16
Self	GPT-4	**1.02**	**1.05**	**0.62**	**0.68**	**0.84**
Random	GPT-3.5	0.20	0.12	0.10	0.28	0.17
Random	GPT-4	0.32	0.36	−0.13	0.51	0.26

Table 1. **Abbreviated Human evaluations.** The expert-written responses are grayed as a reference. The highest column values are **bolded**; and highest values amongst LLMs are highlighted. Two rows are highlighted if they are not statistically different.

LLMs do not make diverse decisions.

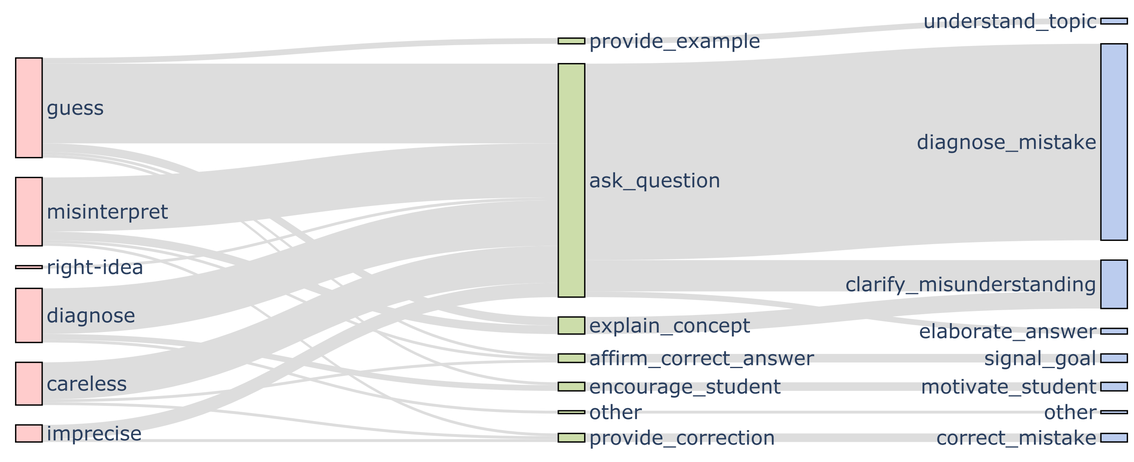


Figure 2. GPT-4 Decision-Making Paths.

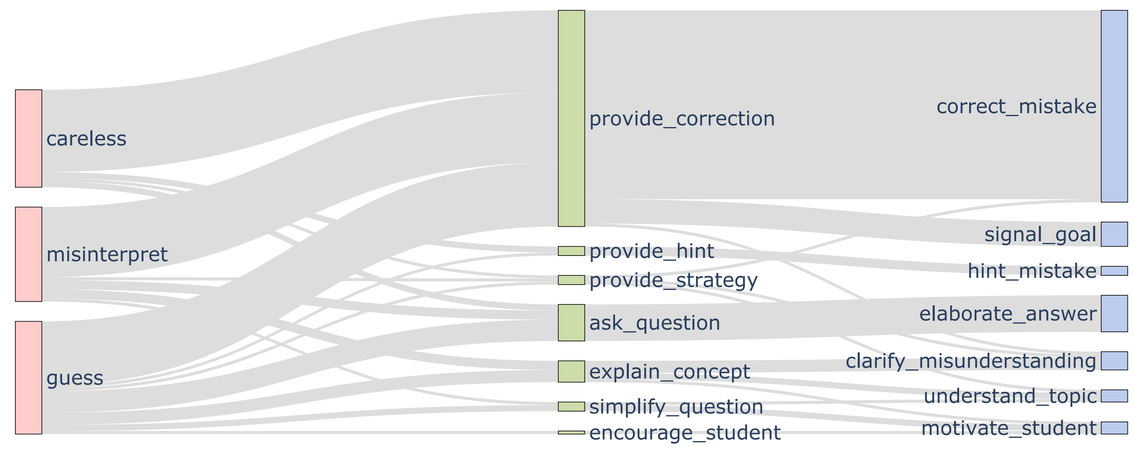


Figure 3. GPT-3.5 Decision-Making Paths.

Bridge language centers the student's problem-solving *process*.

GPT4		Expert + GPT4		Self + GPT4		Random +GPT4	
bigram	log odds	bigram	log odds	bigram	log odds	bigram	log odds
lets_closer	2.76	steps_took	2.04	can_explain	4.98	good_try	1.82
closer_look	2.68	review_concept	1.66	explain_arrived	4.78	start_remember	1.58
effort_lets	2.55	understand_concept	1.56	arrived_answer	4.2	thats_right	1.57
appreciate_effort	2.29	help_understand	1.56	arrived_number	2.19	try_again	1.54
correct_solution	2.19	explain_steps	1.56	are_sure	2.19	thats _good	1.43
look_problem	2.18	took_arrive	1.56	sure_that	2.19	lets_break	1.37
great_effort	1.62	lets_step	1.51	correct_remember	1.38	glasses_water	1.3
lets_steps	1.55	better_understand	1.31	and_long	1.38	for_example	1.3
need_help	1.55	ones_place	1.31	digit_answer	1.38	times_equal	1.3
let_know	1.55	number_sides	1.31	answer_step	1.38	represents_glasses	1.29

Table 2. Top 10 bigrams. `GPT4` with expert- or self decision-making engages more with the student's problem-solving process. `GPT4` with no and random decision-making engages superficially with the student's answer.

Summary and Next Steps

Teaching is hard. Challenge is hidden in their internal, pedagogical decisions.

This work's insight: We need to explicitly model the internal decisions of real experts with Bridge.

Can real novice tutors benefit Bridge? Ongoing Randomized Controlled Trial with Tutor CoPilot `https://osf.io/8d6ha`.