Inter-rater Reliability Open Response Scoring

Inter-rater reliability (IRR) was established by comparing AI scores against a "human truth" gold standard. Two expert coders manually graded a subset of 40 responses per lesson (20 *predict* and 20 *explain*), reconciling disagreements to establish the ground truth. Cohen's Kappa was then calculated to measure the agreement between the AI and truth. Reliability results were heterogeneous, with Kappa values ranging from 0.31 to 1.0 across different tutor moves and question types.

Table A. Inter-rater Reliability for Open Response Scoring (Cohen's Kappa)

Lesson	Human-to-human IRR (κ)		Human-to-LLM IRR (κ)	
	Predict	Explain	Predict	Explain
GUIDE_THINKING	0.78	0.57	0.78	0.31
AFFIRM_CORRECT	1.0	0.77	1.0	1.0
DETERMINE_KNOW	0.69	0.69	0.89	0.74
GIVE_PRAISE	0.64	1.0	0.41	0.76
PROMPT_EXPLAIN	1.0	0.67	1.0	0.53
REACT_ERRORS	1.0	0.89	0.69	0.67

Inter-rater Reliability of Transcript Scoring - Evaluation

To validate the LLM-based assessment method two experienced researchers annotated a subset of 10 tutoring transcriptions across all six tutor moves. IRR between the human raters was calculated, and all disagreements were reconciled to establish a "ground truth" score. The model was then evaluated on these same 10 transcriptions to determine its alignment with expert judgment. Cohen's Kappa scores for the Gemini-2.5-pro compared against the ground truth across the six lessons ranged from 0.0 (when the transcripts nearly all lacked the given tutor move) to 0.8. However, these scores are rendered less representative of LLM ability due to the fact that the set of transcripts analyzed was so small.

Table B. Inter-rater Reliability of Transcript Scoring - Evaluation

Lesson	Human-to-Human IRR (κ)	Human-to-LLM IRR (κ)
GUIDE_THINKING	1.0	0.38
AFFIRM_CORRECT	1.0	0.38
DETERMINE_KNOW	0.58	0.35
GIVE_PRAISE	1.0	0.2
PROMPT_EXPLAIN	0.58	0.8
REACT_ERRORS	0.0*	0.0*

*Low kappa value despite high agreement due to data imbalance.

Inter-rater Reliability of Transcript Scoring - Opportunity

Table C. Inter-rater Reliability of Transcript Scoring - Opportunity

Lesson	Human-to-Human IRR (κ)	Human-to-LLM IRR (κ)
GUIDE_THINKING	0.74	0.52
AFFIRM_CORRECT	1.0	0.74
DETERMINE_KNOW	0.52	0.38
GIVE_PRAISE	1.0	0.4
PROMPT_EXPLAIN	0.55	0.0*
REACT_ERRORS	1.0	1.0

***Low kappa value despite high agreement due to data imbalance.