When Can We Trust AI Coding of Student-Generated Text?

A Committee-Based Approach to Diagnosing Agreement & Uncertainty at Scale

Fanjie Li, Madison Lee Mason, Daniel T. Levin, Alyssa Friend Wise




VANDERBILT UNIVERSITY

LIVE VANDERBILT Learning Innovation Incubator

A New Challenge: Monitoring AI Assessment of Student-Generated Text

- **Deployment challenge:** As AI assessment of student work (e.g., essays, open-ended responses) moves into deployed educational systems, a critical question emerges:

  How do we know it continues to interpret student work as intended?

- **Limits of standard test-set evaluation:** Initial offline evaluation establishes baseline performance, but cannot guarantee stable behavior as models encounter new learners, contexts, and forms of expression.

- **Need for ongoing monitoring:** This calls for interpretable diagnostics that support ongoing monitoring of AI coding on unlabeled data, once deployed.

A Committee-Based Diagnostic Framework to Support Interpretable Monitoring of Automated Coding

Uncertainty Quantification

support human decisions about when to **trust**, **re-evaluate**, or **reject** automated categorization

Including this AI Coding System Element:	Allows us to Assess the **Uncertainty** :
Guided Reasoning with Uncertainty Flags	In Individual LLM Judgement
Repeated Inference for Decision Stability	Across Multiple Runs by an LLM
Committee-Based Coding with Multiple LLMs	Across Multiple LLMs

LIVE

Guided Reasoning with Uncertainty Flags

Input

Prompt Template
(+ codebook for one category)

Data Instance

Prompt Template

Domain expert persona

Goal, context & task + Coding Manual

Guided chain-of-thought

You follow the four steps below for each reflection you code:

Step 1: Immerse and Read
"… read the excerpt holistically, don't try to code yet …"

Step 2: Document Your Analytical Memo
" … analytical scratchpad to think through the data … document reasoning, questions and uncertainties …"

Step 3: Synthesize a Formal Rationale
< Detailed instructions for:
evidence analysis, ambiguity assessment, conclusive judgement, and ***borderline flag*** *setting >*

Step 4: Make the Final Coding Decision
"Based on the above rationale, select the formal code …"

Structured output

Guided Reasoning with Uncertainty Flags

Input

Prompt Template
(+ codebook for one category)

Data Instance

AI Coder

Output
(One LLM Coding Pass)

Coding Rationale (including ambiguity assessment)

Coding Decision (0 or 1)

Borderline Flag (0 or 1)

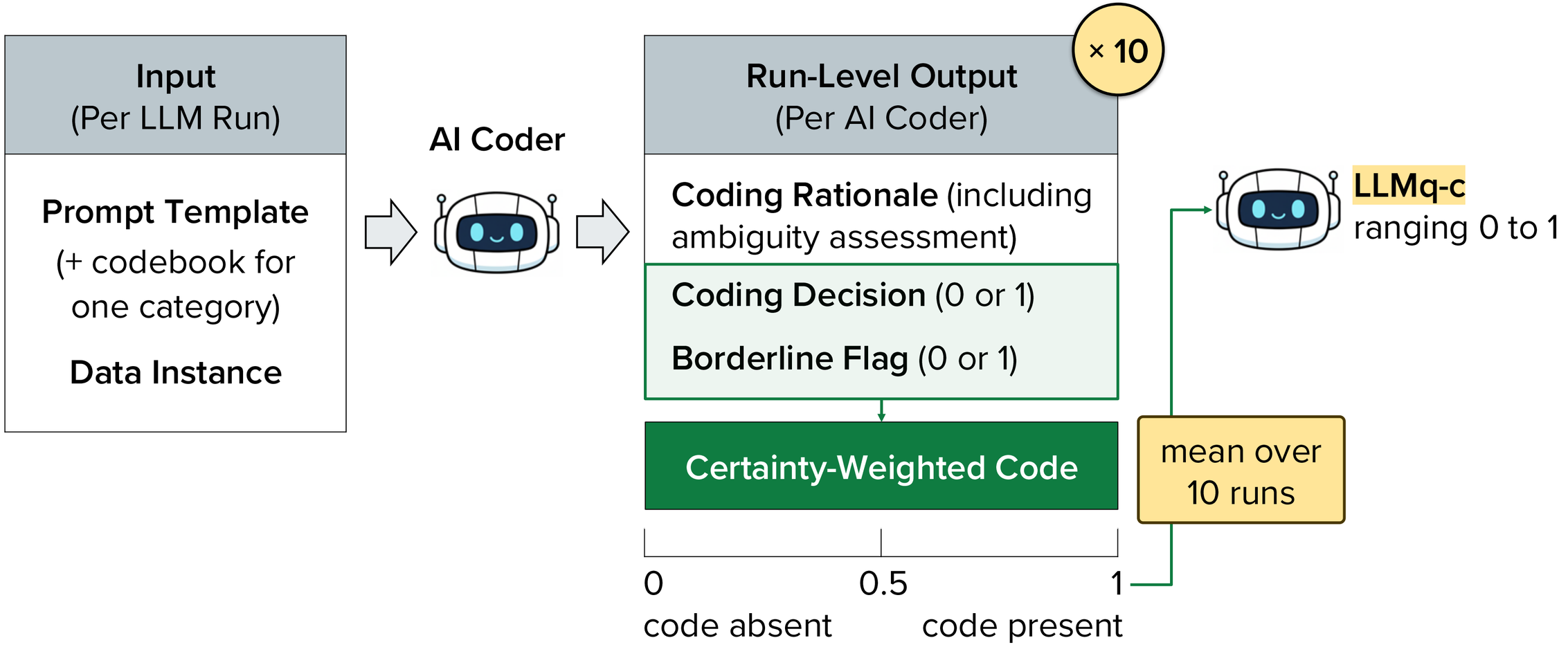
Certainty-Weighted Code

0 — 0.5 — 1

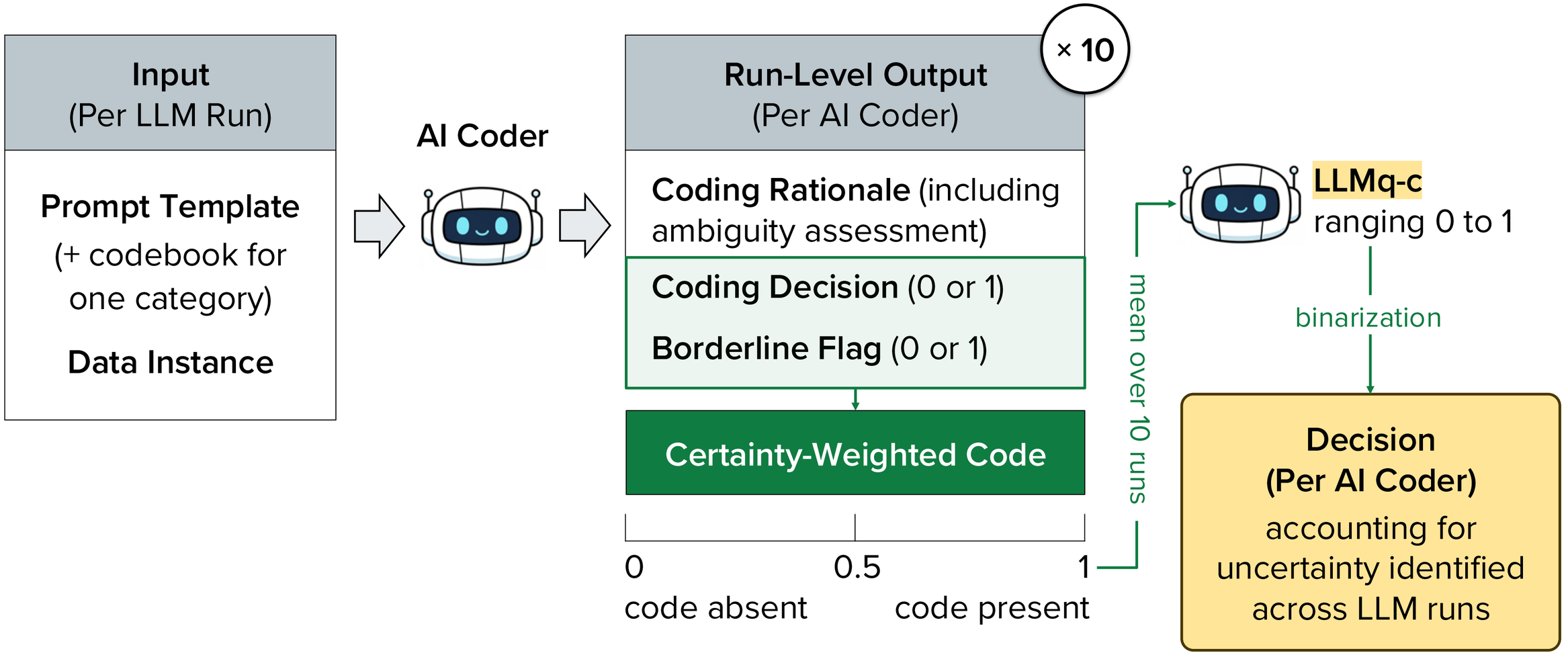
code absent — code present

Decision
(Per Coding Pass)
down-weight those w/ borderline flag to 0.5

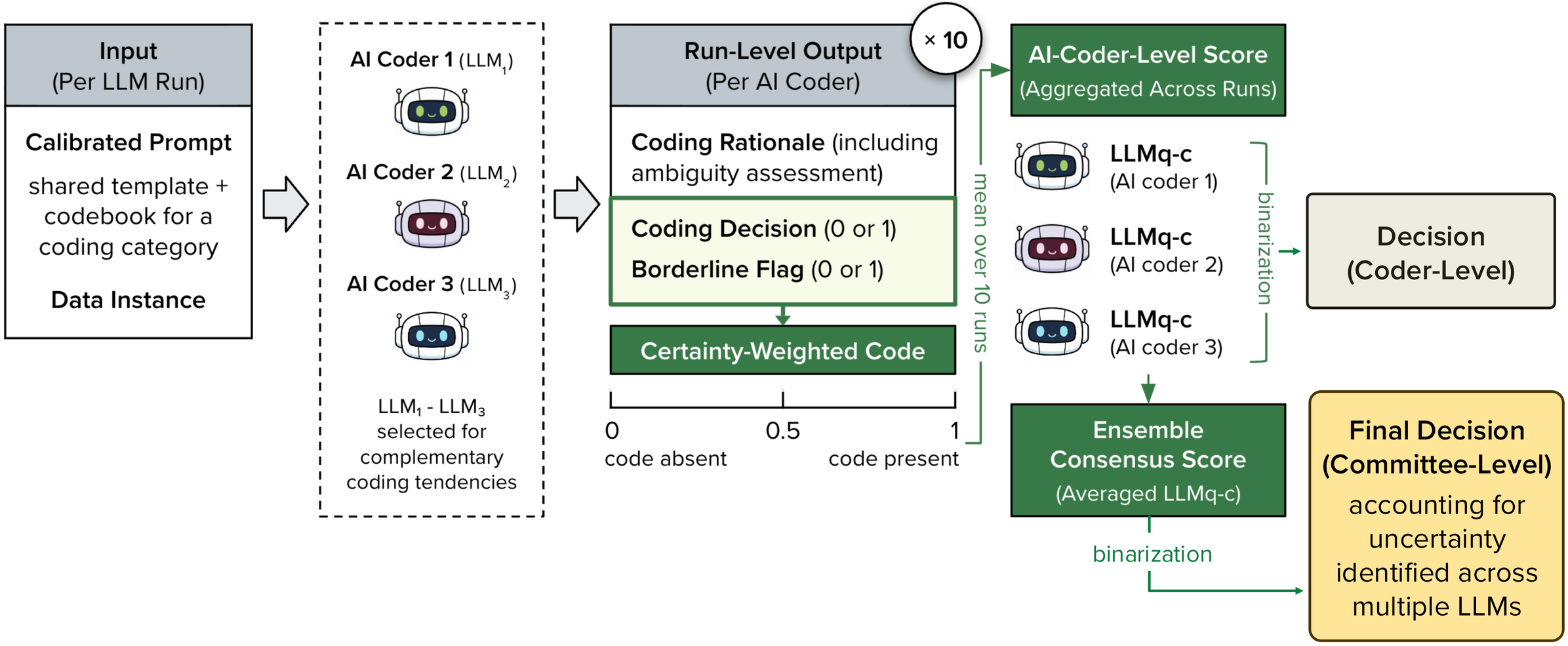
Repeated Inference for Decision Stability



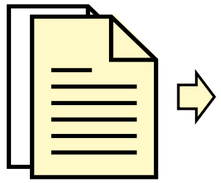
Repeated Inference for Decision Stability



Committee-Based Coding with Multiple LLMs



Monitoring Metrics (Available on Unlabeled Data)



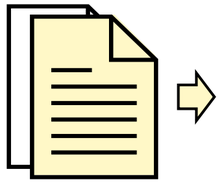
(LLM $_1$)	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1	1_{flag}	1	1	→ LLMq-c = 0.79
(LLM $_2$)	0	0	0_{flag}	0	0	0_{flag}	0	0	0	0_{flag}	→ LLMq-c = 0.09
(LLM $_3$)	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	→ LLMq-c = 0.3

Ensemble Consensus Score **0.39**

Set-Level Monitoring

- Inter-Model Agreement (α_{AI})
- Consensus Entropy (H)
- Borderline Rate (B)

Monitoring Metrics (Available on Unlabeled Data)



(LLM $_1$)	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1	1_{flag}	1	1	→ **LLMq-c = 0.79**
(LLM $_2$)	0	0	0_{flag}	0	0	0_{flag}	0	0	0	0_{flag}	→ **LLMq-c = 0.09**
(LLM $_3$)	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	→ **LLMq-c = 0.3**

Ensemble Consensus Score **0.39**

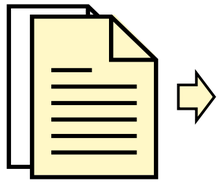
Set-Level Monitoring

- **Inter-Model Agreement (α_{AI})**
- **Consensus Entropy (*H*)**
- **Borderline Rate (*B*)**

$\alpha_{AI} > 0.67$ ------ Moderate agreement

$\alpha_{AI} > 0.80$ ------ Strong agreement

Monitoring Metrics (Available on Unlabeled Data)



(LLM $_1$)	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1	1_{flag}	1	1	→ LLMq-c = 0.79
(LLM $_2$)	0	0	0_{flag}	0	0	0_{flag}	0	0	0	0_{flag}	→ LLMq-c = 0.09
(LLM $_3$)	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	→ LLMq-c = 0.3

$\hat{p}$

Ensemble Consensus Score

0.39

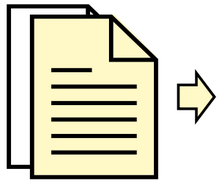
Set-Level Monitoring

- **Inter-Model Agreement (α_{AI})**
- **Consensus Entropy (*H*)**
- **Borderline Rate (*B*)**

Lower *H* --- more **decisive** judgments ($\hat{p}$ near 0 or 1)

Higher *H* --- greater **indecisiveness** ($\hat{p}$ near 0.5)

Monitoring Metrics (Available on Unlabeled Data)



(LLM $_1$)	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1_{flag}	1	1_{flag}	1	1	→ **LLMq-c = 0.79**
(LLM $_2$)	0	0	0_{flag}	0	0	0_{flag}	0	0	0	0_{flag}	→ **LLMq-c = 0.09**
(LLM $_3$)	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	0_{flag}	→ **LLMq-c = 0.3**

$\hat{p}$

Ensemble Consensus Score

0.39

Set-Level Monitoring

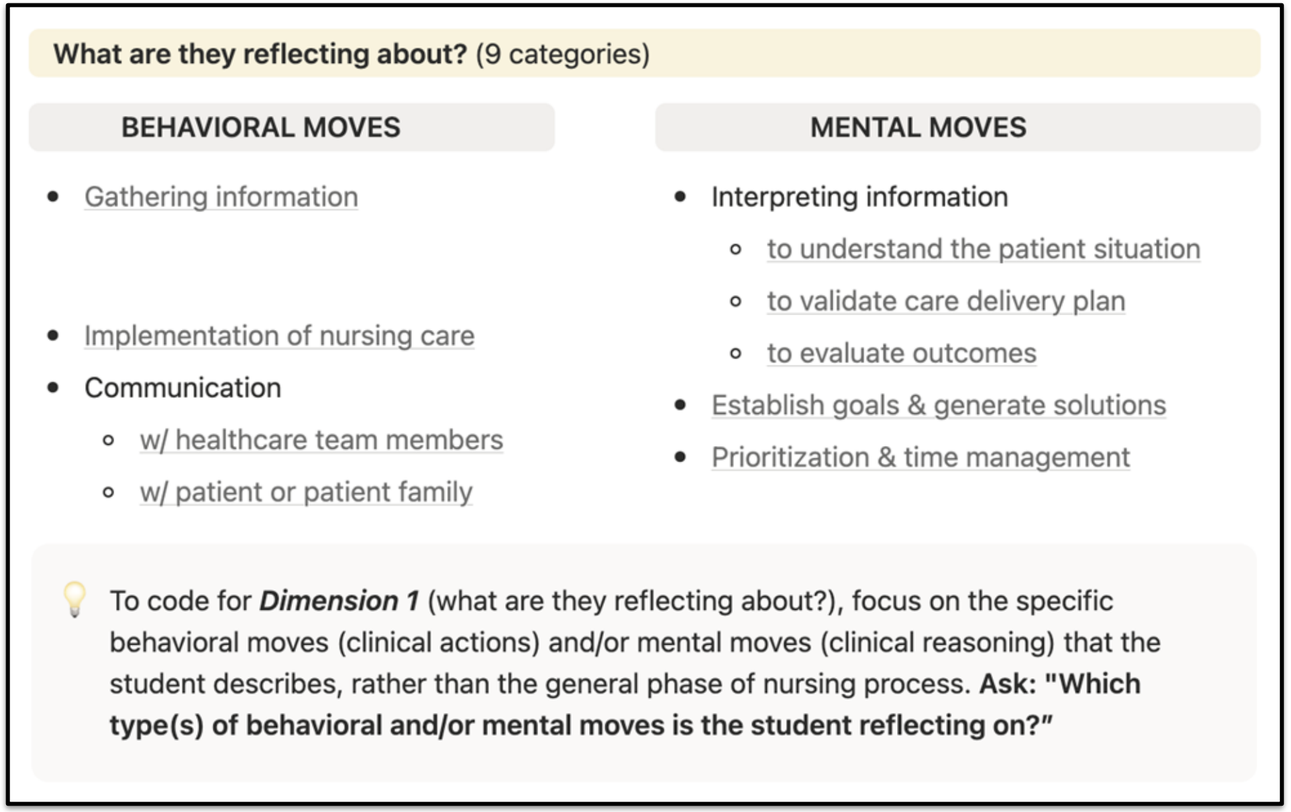
- **Inter-Model Agreement (α_{AI})**
- **Consensus Entropy (*H*)**
- **Borderline Rate (*B*)**

Lower *H* --- more **decisive** judgments ($\hat{p}$ near 0 or 1)
Higher *H* --- greater **indecisiveness** ($\hat{p}$ near 0.5)

Percentage of items for which $L \leq \hat{p} \leq U$
Distinguish between **pervasive uncertainty** vs **uncertainty concentrated on a few items**

Proof-of-Concept Pilot | Context and Data

- **Nursing students' post-simulation written reflection** (213 event-level reflections, 16 students); Human coders double-coded 60 reflections (IRR: 0.63 < α < 0.91)
- **Coding scheme:**
  9 simulation-general nursing activity codes for student reflections
- **AI committee set up:**
  3 LLMs × 10 runs each
  (o3, claude-sonnet-4, deepseek-r1)



What are they reflecting about? (9 categories)

BEHAVIORAL MOVES

- Gathering information
- Implementation of nursing care
- Communication
  - w/ healthcare team members
  - w/ patient or patient family

MENTAL MOVES

- Interpreting information
  - to understand the patient situation
  - to validate care delivery plan
  - to evaluate outcomes
- Establish goals & generate solutions
- Prioritization & time management

To code for ***Dimension 1*** (what are they reflecting about?), focus on the specific behavioral moves (clinical actions) and/or mental moves (clinical reasoning) that the student describes, rather than the general phase of nursing process. **Ask: "Which type(s) of behavioral and/or mental moves is the student reflecting on?"**

Proof-of-Concept Pilot | Results

Behaviorally explicit codes

Need inference about **cognitive** activity

	Test Data				Unlabeled Data		
Activity Code	F1	α_{AI}	H	B	α_{AI}	H	B
GATHER	0.966	0.942	0.295	4.17%	0.885	0.275	7.69%
IMPLMT	0.941	1.000	0.322	0.00%	0.951	0.218	5.13%
T-COMM	0.923	0.877	0.295	8.33%	0.946	0.230	2.56%
P-COMM	0.923	0.944	0.214	8.33%	0.917	0.131	3.42%
INTRPT-SIT	0.770	0.683	0.344	8.33%	0.756	0.276	9.40%
INTRPT-VLDT	0.947	1.000	0.321	0.00%	0.837	0.343	9.40%
INTRPT-EVAL	1.000	0.844	0.195	0.00%	0.532	0.163	1.71%
GOALS	0.824	0.698	0.506	25.00%	0.719	0.456	12.82%
TIME	1.000	0.735	0.379	4.17%	0.644	0.309	11.11%

Proof-of-Concept Pilot | Results

Behaviorally explicit codes

	Test Data				Unlabeled Data		
Activity Code	F1	α_{AI}	H	B	α_{AI}	H	B
GATHER	0.966	0.942	0.295	4.17%	0.885	0.275	7.69%
IMPLMT	0.941	1.000	0.322	0.00%	0.951	0.218	5.13%
T-COMM	0.923	0.877	0.295	8.33%	0.946	0.230	2.56%
P-COMM	0.923	0.944	0.214	8.33%	0.917	0.131	3.42%
INTRPT-SIT	0.770	0.683	0.344	8.33%	0.756	0.276	9.40%
INTRPT-VLDT	0.947	1.000	0.321	0.00%	0.837	0.343	9.40%
INTRPT-EVAL	1.000	0.844	0.195	0.00%	0.532	0.163	1.71%
GOALS	0.824	0.698	0.506	25.00%	0.719	0.456	12.82%
TIME	1.000	0.735	0.379	4.17%	0.644	0.309	11.11%

GATHER (gather info)
IMPLMT (implementation of nursing care)
T-COMM (communication w/ healthcare team)
P-COMM (communication w/ patient family)

Stable consensus profile

- High test-set performance (F1 > .9)
- Stable α_{AI} across test and unlabeled data (all α_{AI} > .8)
- Consensus entropy (*H*) remain relatively stable (low uncertainty)
- Borderline rates stayed below 10%

LIVE

Proof-of-Concept Pilot | Results

	Test Data				Unlabeled Data		
Activity Code	F1	α_{AI}	H	B	α_{AI}	H	B
GATHER	0.966	0.942	0.295	4.17%	0.885	0.275	7.69%
IMPLMT	0.941	1.000	0.322	0.00%	0.951	0.218	5.13%
T-COMM	0.923	0.877	0.295	8.33%	0.946	0.230	2.56%
P-COMM	0.923	0.944	0.214	8.33%	0.917	0.131	3.42%
INTRPT-SIT	0.770	0.683	0.344	8.33%	0.756	0.276	9.40%
INTRPT-VLDT	0.947	1.000	0.321	0.00%	0.837	0.343	9.40%
INTRPT-EVAL	1.000	0.844	0.195	0.00%	0.532	0.163	1.71%
GOALS	0.824	0.698	0.506	25.00%	0.719	0.456	12.82%
TIME	1.000	0.735	0.379	4.17%	0.644	0.309	11.11%

Need inference about **cognitive** activity

INTER-MODEL AGREEMENT

LOW HIGH

DECISIVENESS

LOW

HIGH

Codes need inference about cognitive activity

Behaviorally explicit codes

Proof-of-Concept Pilot | Results

	Test Data				Unlabeled Data		
Activity Code	F1	α_{AI}	H	B	α_{AI}	H	B
INTRPT-SIT	0.770	0.683	0.344	8.33%	0.756	0.276	9.40%

Example Reflection | Coding Category: **INTRPT-SIT** (Interpret Info to Understand Patient Situation)

What aspects of the simulation or your own actions led to you achieving your goals during this event?

The mother informing me that the patient has dysphagia influenced my nursing actions before administering med.

(o3)	1	1	1	1	1	1	1	1_{flag}	1	1	→ LLMq-c $_{(o3)}$	= **0.95**
(claude)	1	0_{flag}	1	0	0_{flag}	0_{flag}	1_{flag}	1	0_{flag}	1_{flag}	→ LLMq-c $_{(claude)}$	= **0.6**
(deepseek)	0	0	0	0	0_{flag}	0_{flag}	0	0_{flag}	0	0_{flag}	→ LLMq-c $_{(deepseek)}$	= **0.2**

Ensemble Consensus Score

0.58

Sample Rationale (Code = 0):

"While the reflection could be interpreted as implicitly containing interpretation work (understanding that dysphagia requires modified medication administration), they do not explicitly describe interpreting this information to understand the patient's condition."

Sample Rationale (Code = 1):

"Some uncertainty: the reflection is brief and most of the emphasis is on the subsequent action (elevating the patient). However, the student does explicitly note the significance of the dysphagia information for understanding the patient's current condition, providing enough evidence for the code."

Proof-of-Concept Pilot | Results

	Test Data				Unlabeled Data		
Activity Code	F1	α_{AI}	H	B	α_{AI}	H	B
GATHER	0.966	0.942	0.295	4.17%	0.885	0.275	7.69%
IMPLMT	0.941	1.000	0.322	0.00%	0.951	0.218	5.13%
T-COMM	0.923	0.877	0.295	8.33%	0.946	0.230	2.56%
P-COMM	0.923	0.944	0.214	8.33%	0.917	0.131	3.42%
INTRPT-SIT	0.770	0.683	0.344	8.33%	0.756	0.276	9.40%
INTRPT-VLDT	0.947	1.000	0.321	0.00%	0.837	0.343	9.40%
INTRPT-EVAL	1.000	0.844	0.195	0.00%	0.532	0.163	1.71%
GOALS	0.824	0.698	0.506	25.00%	0.719	0.456	12.82%
TIME	1.000	0.735	0.379	4.17%	0.644	0.309	11.11%

Illustrate how α_{AI} , ***H***, and ***B*** can be jointly interpreted to distinguish **stable vs problematic ensemble behavior** and guide actionable next steps, such as **targeted data review** or **construct refinement**

Limitations & Summary of Contribution

Contribution

- **Interpretable uncertainty for LLM coding** that move beyond token-level confidence to diagnostics grounded in coding task semantics
- **Complementary diagnostics on agreement, decisiveness, and certainty** to characterize AI coding behavior on unlabeled data
- **Continuous deployment-time monitoring** that support ongoing human oversight of AI coding beyond standard test-set evaluation.

Limitations and Future Work

- **Pilot-scale validation** (one coding scheme, one educational domain)
- **Computational cost of committee-based inference** is non-trivial; future work need to explore sparse and adaptive monitoring strategies.
- Diagnostics indicate ***when* further investigation is needed**, not yet ***what* intervention should follow**.

Thank You! Questions?

Calibrate AI coder committee via iterative codebook refinement

bit.ly/lak26-clarify

Support human oversight of AI coding after deployment

bit.ly/aied26-monitor

Fanjie Li, Vanderbilt University, LIVE Learning Innovation Incubator