HILAK 2026 Workshop, Bergen, Norway

When AI Questions, Human CLARIFY:

Using RE-LLM Coding Uncertainty to Resolve Codebook Ambiguities

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




VANDERBILT UNIVERSITY

LIVE VANDERBILT Learning Innovation Incubator

Rethinking AI's Roles in Codebook-Based Qual Coding

1

Human + AI

code better,
not just faster

2

Uncertainty as Signal

AI uncertainty reveals
ambiguity in the codebook

3

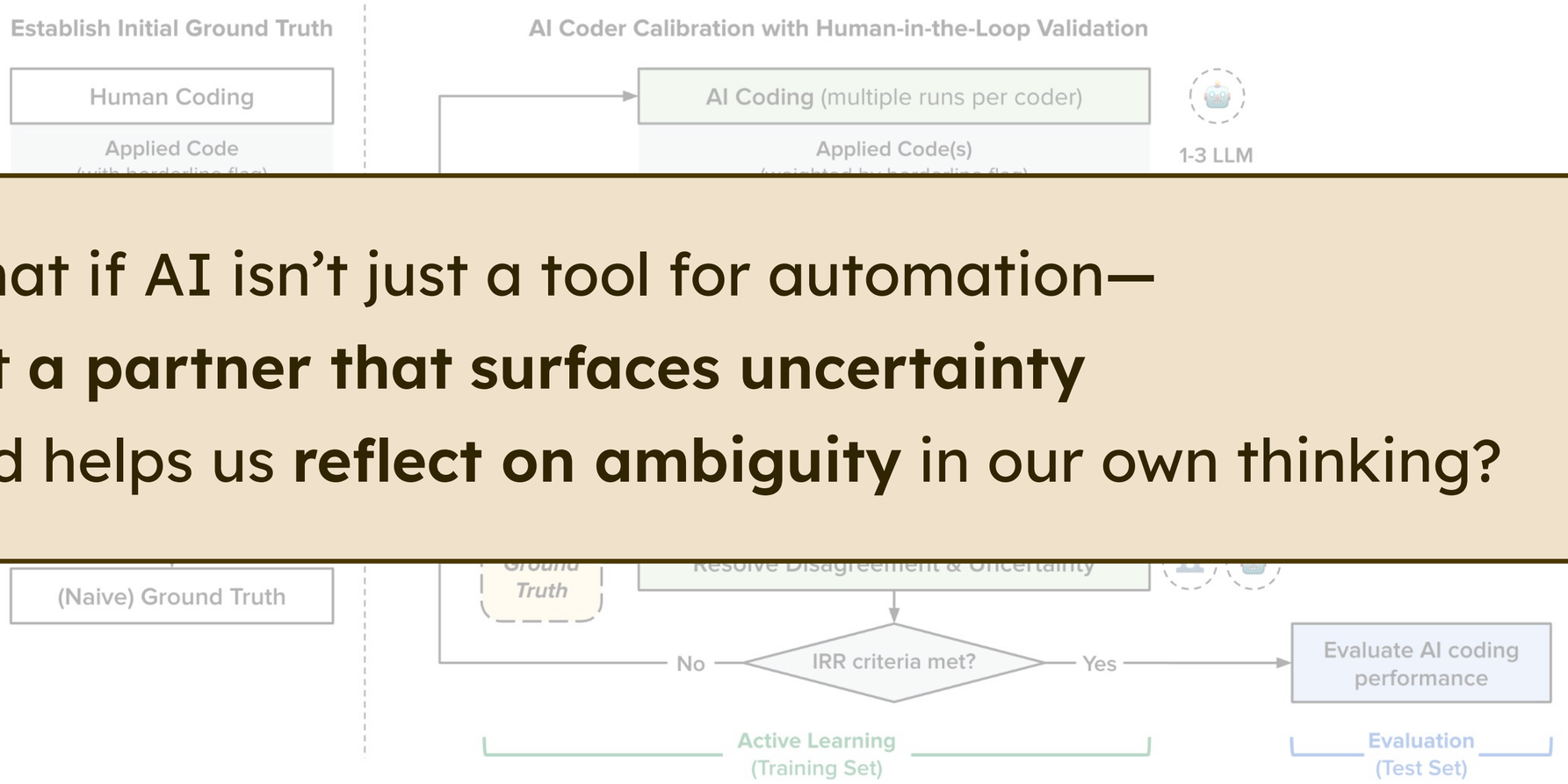
CLARIFY Workflow

illustrated w/ nursing
reflection analytics

Human-AI Partnership in LLM-Assisted Content Analytics

Traditional Approach	Hybrid Intelligence Approach
Code Faster	Code Better
AI for Automation	AI as Thinking Partner
Replicates Human Coding	Stress-Tests Human Coding
Obscures Ambiguity	Surfaces Uncertainty, Diagnoses Ambiguity

CLARIFY Toolset and Workflow



What if AI isn't just a tool for automation—
but **a partner that surfaces uncertainty**
and helps us **reflect on ambiguity** in our own thinking?

Establish Initial Ground Truth

Human Coding

Applied Code
(with borderline flag)

Coding Memo
(rationale & uncertainties)

Calculate Human-Human IRR

Reconcile Disagreements

(Naive) Ground Truth

AI Coder Calibration with Human-in-the-Loop Validation

AI Coding (multiple runs per coder)

1-3 LLM coder(s)

Applied Code(s)
(weighted by borderline flag)

Coding Memo / Reasoning Template
(rationale & uncertainties)

Self Consistency
(certainty-weighted LLMq)

Calculate Human-AI IRR

Uncertainty / Disagreement Sampling
(rank & select informative cases for expert review)

Resolve Disagreement & Uncertainty

Informed Ground Truth

Iterative codebook refinement & ground truth assessment

IRR criteria met?

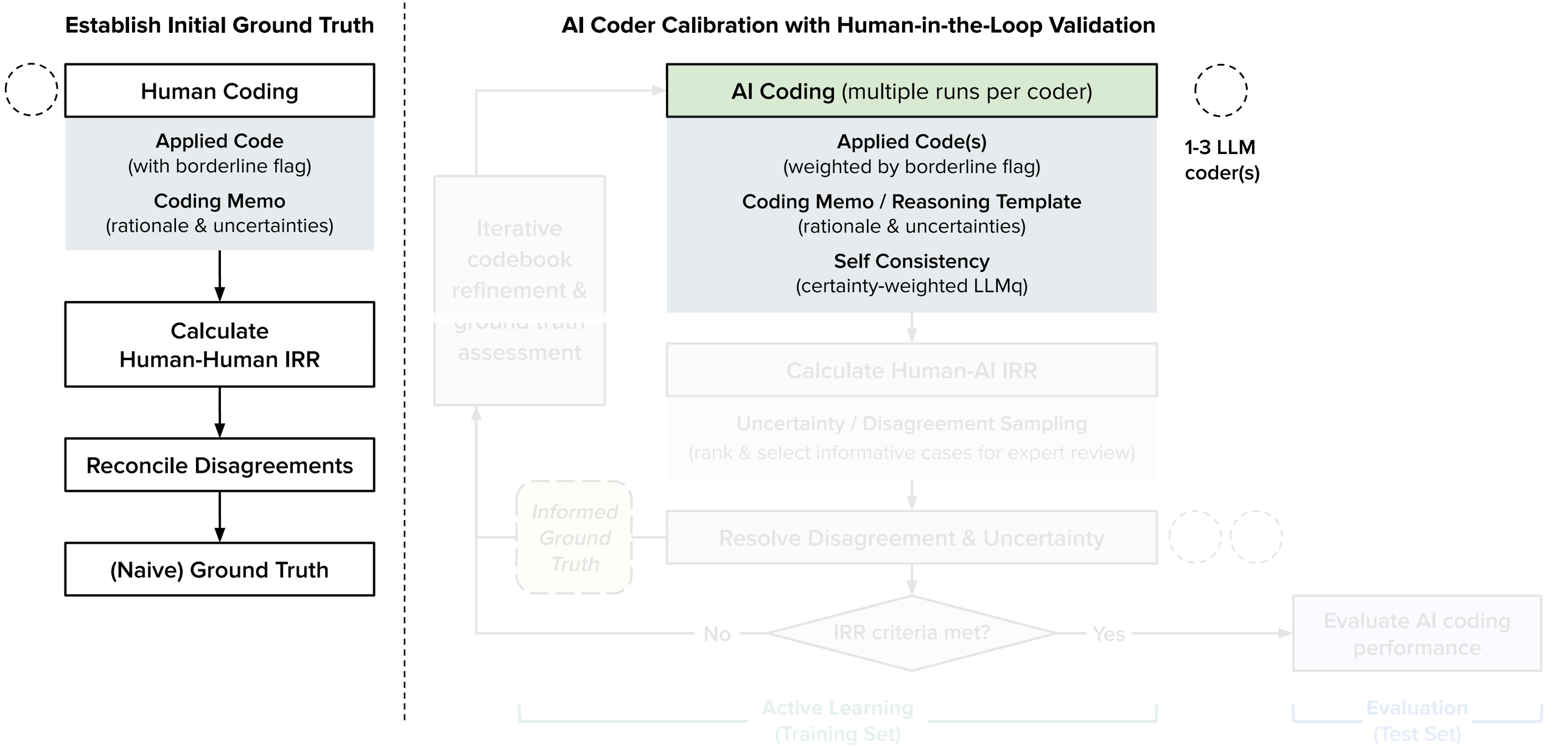
No

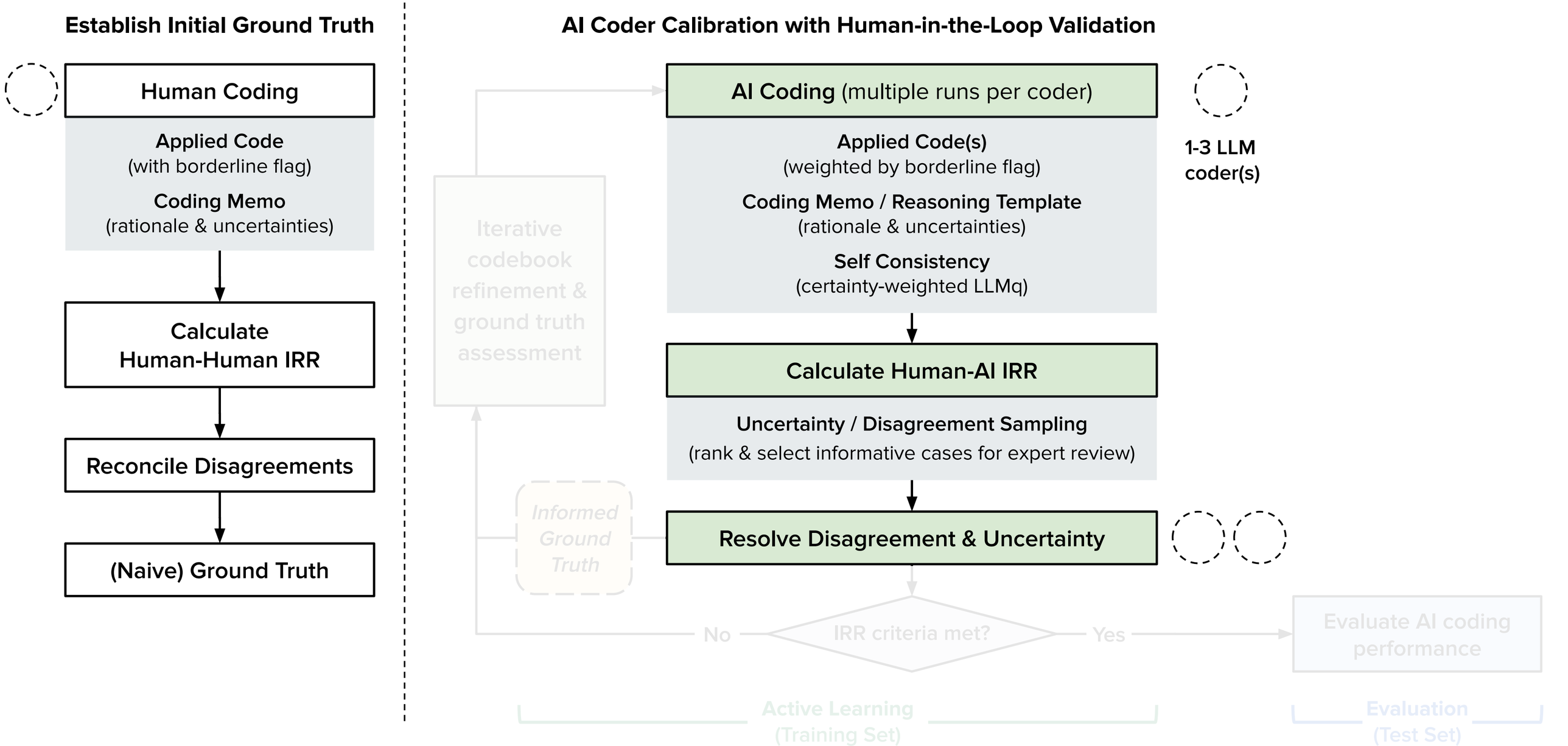
Yes

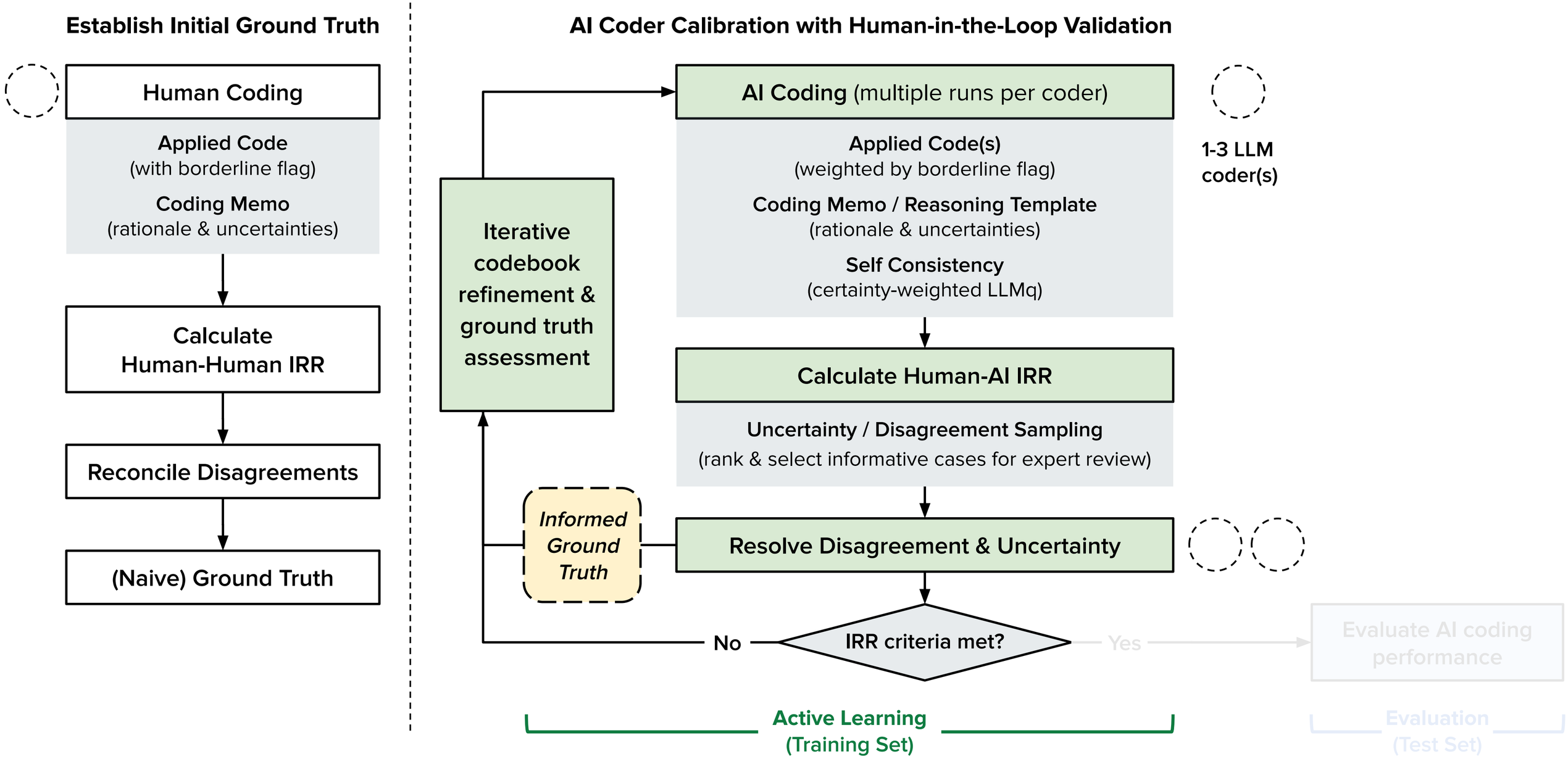
Evaluate AI coding performance

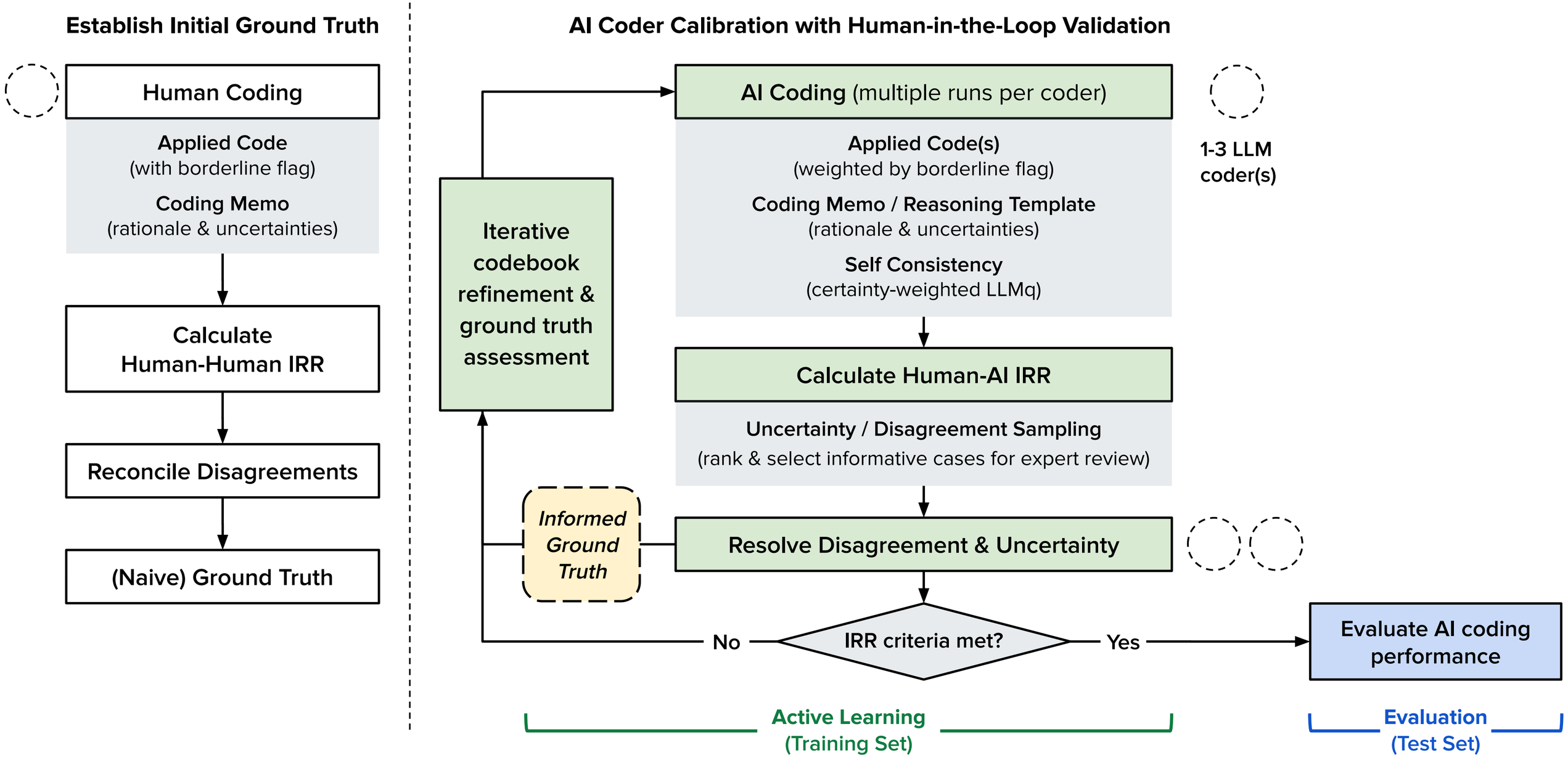
Active Learning (Training Set)

Evaluation (Test Set)


Establish Initial Ground Truth
Human Coding
Applied Code
(with borderline flag)
Coding Memo
(rationale & uncertainties)
Calculate
Human-Human IRR
Reconcile Disagreements
(Naive) Ground Truth
AI Coder Calibration with Human-in-the-Loop Validation
AI Coding (multiple runs per coder)
Applied Code(s)
(weighted by borderline flag)
Coding Memo / Reasoning Template
(rationale & uncertainties)
Self Consistency
(certainty-weighted LLMq)
1-3 LLM
coder(s)
Iterative
codebook
refinement &
assessment
Calculate Human-AI IRR
Uncertainty / Disagreement Sampling
(rank & select informative cases for expert review)
Informed
Ground
Truth
Resolve Disagreement & Uncertainty
No
IRR criteria met?
Yes
Evaluate AI coding
performance
Active Learning
(Training Set)
Evaluation
(Test Set)


Establish Initial Ground Truth
Human Coding
Applied Code
(with borderline flag)
Coding Memo
(rationale & uncertainties)
Calculate
Human-Human IRR
Reconcile Disagreements
(Naive) Ground Truth
AI Coder Calibration with Human-in-the-Loop Validation
AI Coding (multiple runs per coder)
Applied Code(s)
(weighted by borderline flag)
Coding Memo / Reasoning Template
(rationale & uncertainties)
Self Consistency
(certainty-weighted LLMq)
1-3 LLM
coder(s)
Calculate Human-AI IRR
Uncertainty / Disagreement Sampling
(rank & select informative cases for expert review)
Resolve Disagreement & Uncertainty
Iterative
codebook
refinement &
ground truth
assessment
Informed
Ground
Truth
No
IRR criteria met?
Yes
Evaluate AI coding
performance
Active Learning
(Training Set)
Evaluation
(Test Set)

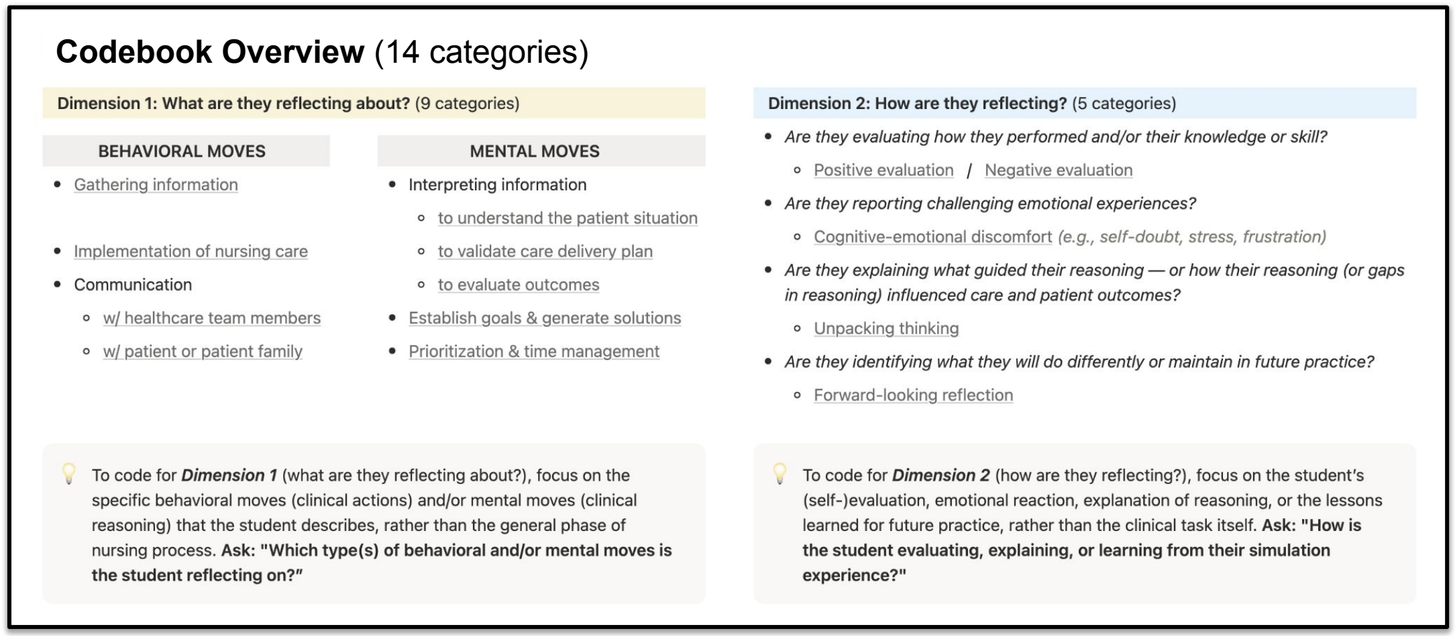

Establish Initial Ground Truth
Human Coding
Applied Code
(with borderline flag)
Coding Memo
(rationale & uncertainties)
Calculate
Human-Human IRR
Reconcile Disagreements
(Naive) Ground Truth
AI Coder Calibration with Human-in-the-Loop Validation
AI Coding (multiple runs per coder)
Applied Code(s)
(weighted by borderline flag)
Coding Memo / Reasoning Template
(rationale & uncertainties)
Self Consistency
(certainty-weighted LLMq)
1-3 LLM
coder(s)
Calculate Human-AI IRR
Uncertainty / Disagreement Sampling
(rank & select informative cases for expert review)
Resolve Disagreement & Uncertainty
Informed
Ground
Truth
Iterative
codebook
refinement &
ground truth
assessment
IRR criteria met?
No
Yes
Evaluate AI coding
performance
Active Learning
(Training Set)
Evaluation
(Test Set)


Establish Initial Ground Truth
Human Coding
Applied Code
(with borderline flag)
Coding Memo
(rationale & uncertainties)
Calculate
Human-Human IRR
Reconcile Disagreements
(Naive) Ground Truth
AI Coder Calibration with Human-in-the-Loop Validation
AI Coding (multiple runs per coder)
Applied Code(s)
(weighted by borderline flag)
Coding Memo / Reasoning Template
(rationale & uncertainties)
Self Consistency
(certainty-weighted LLMq)
1-3 LLM
coder(s)
Iterative
codebook
refinement &
ground truth
assessment
Calculate Human-AI IRR
Uncertainty / Disagreement Sampling
(rank & select informative cases for expert review)
Informed
Ground
Truth
Resolve Disagreement & Uncertainty
No
IRR criteria met?
Yes
Evaluate AI coding
performance
Active Learning
(Training Set)
Evaluation
(Test Set)

Proof-of-Concept Pilot

Context, Data and Coding Scheme

- **Nursing students' post-simulation written reflection** (213 event-level reflections, 16 students); Human coders double-coded 60 reflections (IRR: 0.63 < α < 0.91)
- **AI committee set up:**
  3 LLMs × 10 runs each
- **Focus code:**
  - **VALIDATE:** ensuring the correctness, safety, and readiness of a planned nursing intervention for implementation



Codebook Overview (14 categories)

Dimension 1: 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?"**

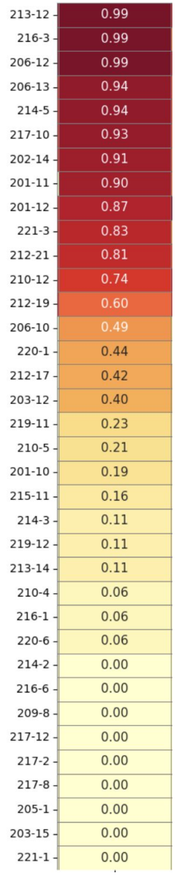
Dimension 2: How are they reflecting? (5 categories)

- *Are they evaluating how they performed and/or their knowledge or skill?*
  - Positive evaluation / Negative evaluation
- *Are they reporting challenging emotional experiences?*
  - Cognitive-emotional discomfort *(e.g., self-doubt, stress, frustration)*
- *Are they explaining what guided their reasoning — or how their reasoning (or gaps in reasoning) influenced care and patient outcomes?*
  - Unpacking thinking
- *Are they identifying what they will do differently or maintain in future practice?*
  - Forward-looking reflection

To code for ***Dimension 2*** (how are they reflecting?), focus on the student's (self-)evaluation, emotional reaction, explanation of reasoning, or the lessons learned for future practice, rather than the clinical task itself. **Ask: "How is the student evaluating, explaining, or learning from their simulation experience?"**

VALIDATE Code Trajectory

Active learning round 1:



Human-AI agreement (Krippendorff's alpha) = 0.641

Number of cases with human-AI or AI-AI disagreement = 12

Disagreement analysis:

Review top 6 high-entropy cases (out of 36 training examples)

text id	consensus entropy	human	o3	claude	deepseek
213-12	0.99	1.00	0.79	0.09	0.30
216-3	0.99	0.70	0.94	0.31	0.27
206-13	0.94	0.70	1.00	0.31	0.56
214-5	0.94	0.70	1.00	0.57	0.30
217-10	0.93	0.00	0.77	0.44	0.18
206-12	0.93	0.70	1.00	0.63	0.30

Note: When human coders initially disagreed, reconciled labels were entered as .3 / .7 (rather than 0 / 1) to indicate the uncertainty identified.

VALIDATE Code Trajectory

Highest-entropy example in active learning round 1:

Reflection 213-12 | Coding Category: VALIDATE (Interpret Info to Validate Care Delivery Plan)

Did this event segment contain a key moment that influenced the outcome of the simulation?

"I was stressed with medication calculation."

(o3)	1 flag	1 flag	1 flag	1 flag	1 flag	1 flag	1	1 flag	1	1	→ LLMq-c (o3)	= 0.79
(claude)	0	0	0 flag	0	0	0 flag	0	0	0	0 flag	→ LLMq-c (claude)	= 0.09
(deepseek)	0 flag	0 flag	0 flag	0 flag	0 flag	0 flag	0 flag	0 flag	0 flag	0 flag	→ LLMq-c (deepseek)	= 0.3

Ensemble Consensus Score **0.39**

Sample Rationale (Code = 0):

"While 'medication calculation' nominally connects to an indicator, **the lack of detail about validation behaviors creates uncertainty**. Codebook requires evidence of cognitive work to validate care plans, not just acknowledgement that calculations occurred. **However, one could argue the mere act of doing calculations implies validation attempt**."

Sample Rationale (Code = 1):

"The statement is extremely brief and does not explicitly describe performing or checking the calculation, merely being stressed. **One could argue it does not provide enough detail about the mental process**. Nonetheless, the **explicit reference to medication calculation—as a key moment—provides minimal but sufficient evidence** under a generous reading of the codebook. Borderline but leaning present."

VALIDATE Code Trajectory

Added heuristics: Apply this code when students are reflecting on interpreting information to validate care delivery plan, regardless of how they are reflecting on it. That is, **reflection on interpreting information to validate care delivery plan can take various forms, including** but not limited to describing what they did or failed to do during the simulation, evaluating their performance, **expressing feelings or challenges about the experience,** identifying lessons learned, and planning for future practice. All of these reflection types qualify as evidence for coding.

Sample Rationale BEFORE Codebook Clarification: "While 'medication calculation' nominally connects to an indicator, the lack of detail about validation behaviors creates uncertainty. Codebook requires evidence of cognitive work to validate care plans, not just acknowledgement that calculations occurred. However, one could argue the mere act of doing calculations implies validation attempt."

Sample Rationale AFTER Codebook Clarification: "Minor uncertainty exists because the student does not explicitly describe the validation process (e.g., verifying dosage accuracy) but only references stress associated with the task. **However, the codebook's heuristics emphasize that outcomes or emotional reactions related to validation work still qualify as evidence.**

VALIDATE Code Trajectory

Tracking performance across three **active learning (AL) rounds**:

AL Round 1	AL Round 2	AL Round 3
α = 0.641, # disagreement = **12**	α = 0.739, # disagreement = **9**	α = 0.845, # disagreement = **5**

Performance on **held-out test set**:

Initial Codebook	Revised Codebook (Final)
α = 0.741	α = 0.941

Concluding Thoughts

Scan for full paper

bit.ly/lak26-clarify

Objectives	AI roles	Human roles
AI coding as a supplementary verification layer	Detect drift in code application across the entire dataset (Zambrano et al., 2023)	Resolve inconsistencies observed through human-in-the-loop validation (Cohn et al., 2024; Ramanathan et al., 2025; Zambrano et al., 2023)
	Assess intra-coder consistency through recursive model queries (Ramanathan et al., 2025; Tai et al., 2024) Use convergence of codes to quantify decision confidence (Wang et al., 2023)	Evaluate trends in code stabilization (Tai et al., 2024; Ramanathan et al., 2025) Review low-confidence coding decisions (Li et al., 2026)
AI coder(s) as collaborative partners for identifying and resolving gaps in the current coding scheme	Articulate reasoning behind coding decisions, which either prompts human reflection on their rationales and assumptions or exposes unclear code definitions that cause coder confusions (Ramanathan et al., 2025; Zambrano et al., 2023; Cohn et al., 2024)	Perform disagreement analysis and resolve AI coder confusions through codebook improvements and/or few-shot examples with human-corrected reasoning (Ramanathan et al., 2025; Zambrano et al., 2023; Cohn et al., 2024)
	Provide suggestions for updating the code definition upon human request (Zambrano et al., 2023)	Decide whether to accept/reject changes suggested by LLM (Zambrano et al., 2023; Ramanathan et al., 2025)