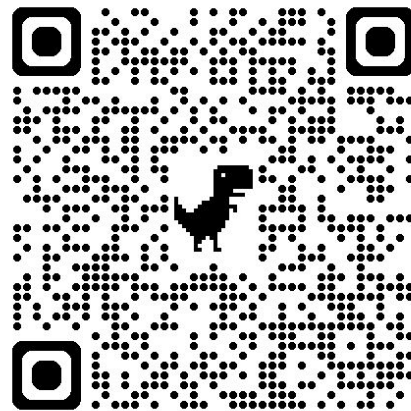


Peer and Tutor in One:

A Productive Failure-Based Architecture for Multi-Role Conversational Agents in Physics Learning



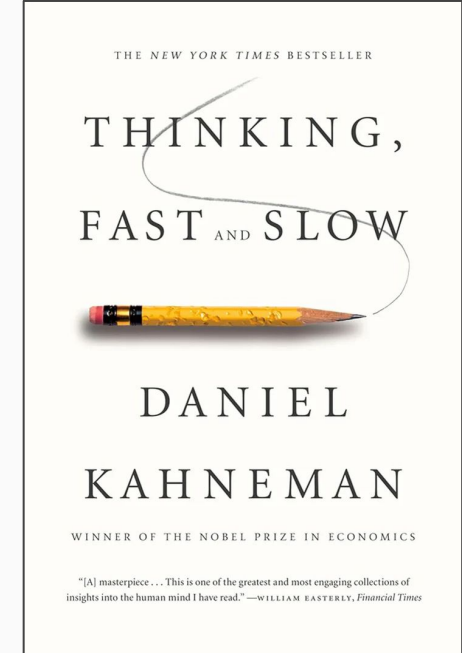
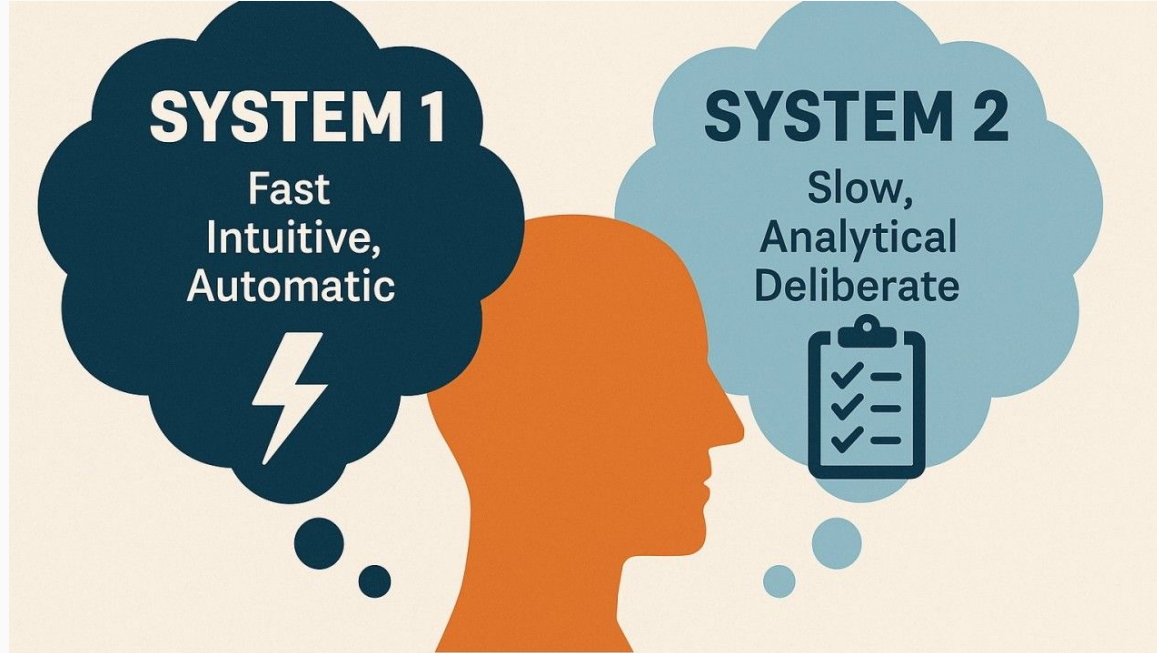
Junbo Koh¹, Minsun Cho², and Sunyoung Keum³

¹ Learning Sciences Research Institute, Seoul National University

² Department of Education, Seoul National University

³ Korea Institute for Curriculum and Evaluation





Gaps undercut today's conversational agents



Technical-feasibility focus

Systems are engineered without theoretical grounding (Stamper et al., 2024).



Perception-only evaluation

Studies report learner perceptions, leaving design fidelity (Debets et al., 2025).

1. Build dual agents to induce cognitive conflict first, and construct knowledge later applying Productive Failure (PF) learning theory.
2. Turn the pedagogy into an inspectable intent → act policy applying Dialogue Management System (DMS) architecture.

Productive Failure (Kapur, 2008)

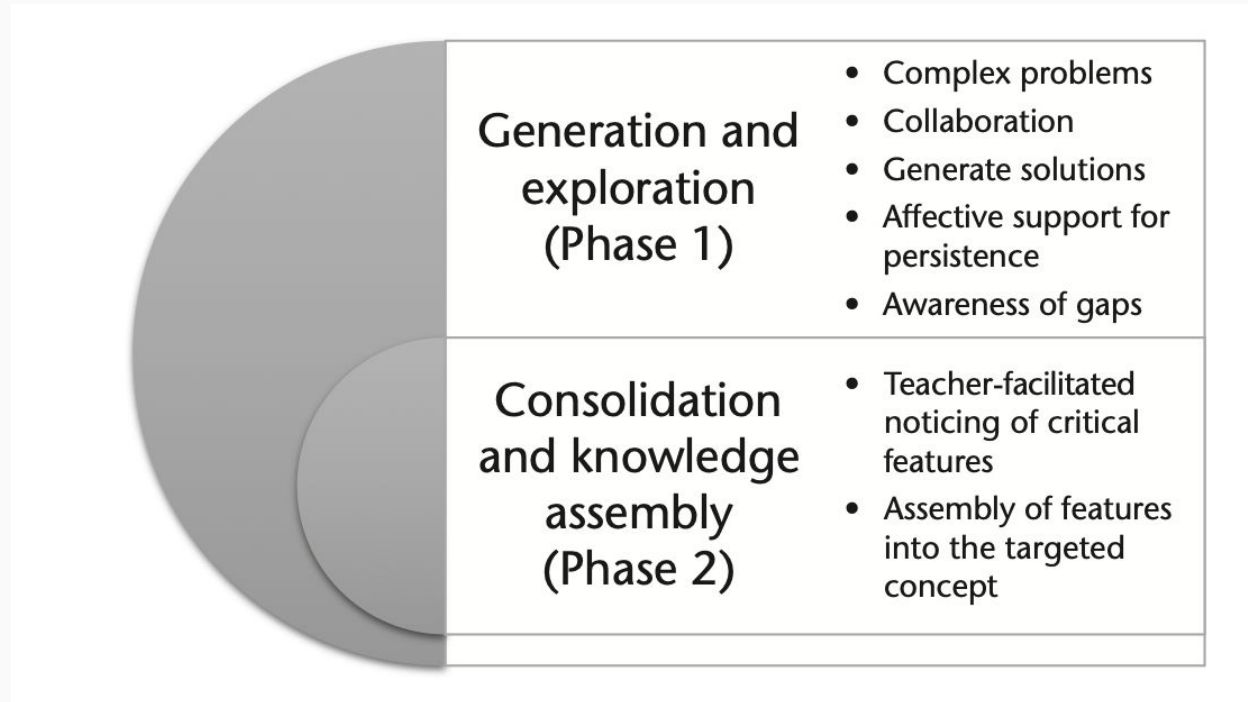


Fig. 1. Productive Failure Design (Hmelo-Silver et al., 2018).

The two phases of PF call for two contrasting agent roles

Love the honesty! Any specific scenes pop up when you think of circular motion?



Peer

Net force direction is missing!
Try explaining why it aims at the planet's center using $F=ma$.



Tutor



PHASE 1 · Exploration

Peer agent

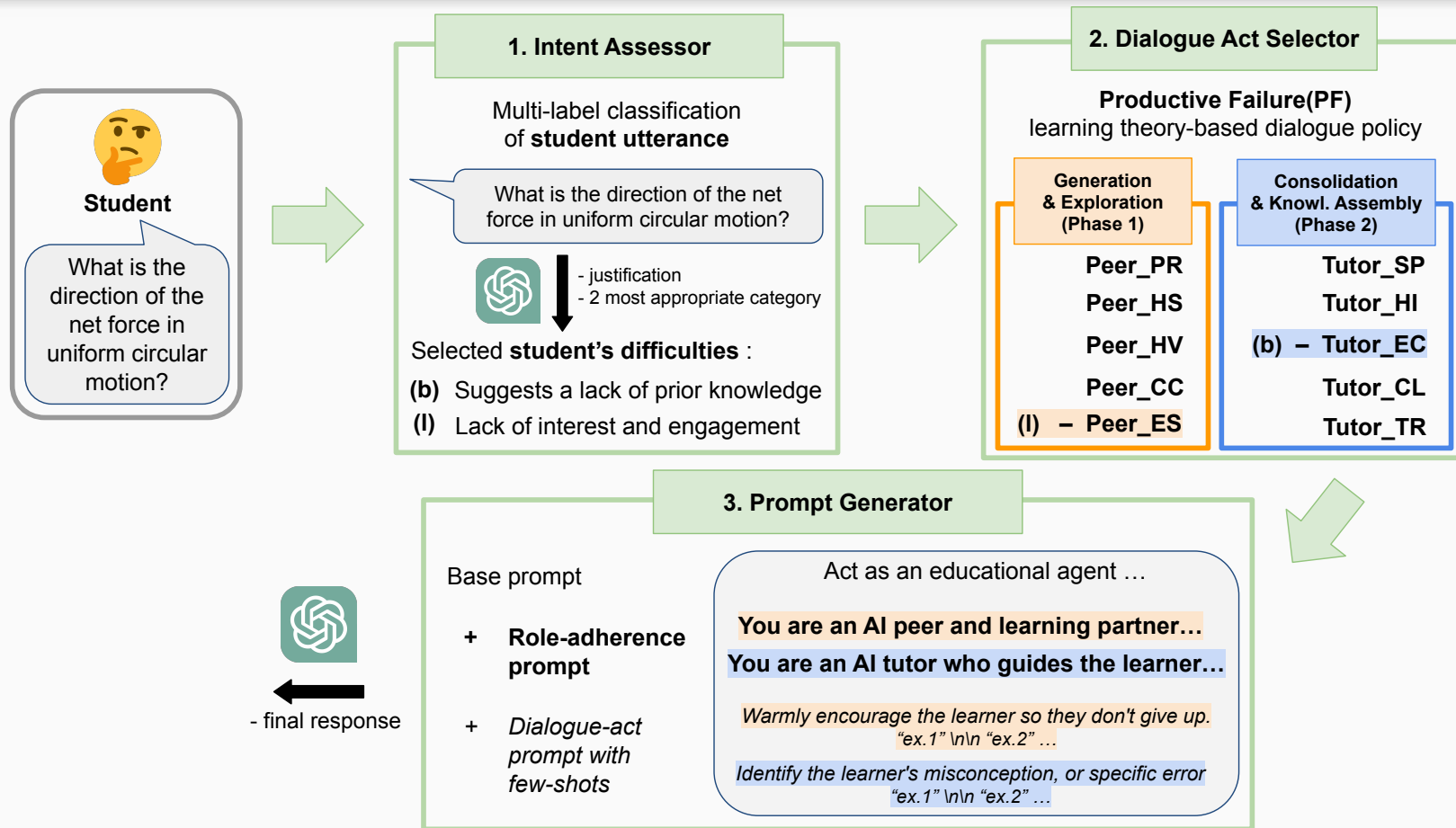
- Learners reveal prior concepts and experience intended failure.
- The peer facilitates: questions, hypotheses, and cognitive conflict — without giving answers.



PHASE 2 · Consolidation

Tutor agent

- Instruction helps learners compare their ideas with canonical solutions.
- The tutor elaborates: corrects errors, clarifies concepts, and structures knowledge.



The DMS was integrated into a simulation-based physics-inquiry platform (Koh et al., 2025) with five learning stages and multimodal input — learners draw a diagram and explain their reasoning.

The screenshot displays a web-based physics inquiry platform. On the left, a sidebar contains navigation links: 'Easy Physics', 'Main Page', 'Introduction', 'Think', 'Experiment', 'Review', and 'Apply'. Below this is a 'Chat with your AI Teacher' section showing a 'Submitted Answer' and a detailed feedback message about a misconception regarding the direction of net force in circular motion. The main area is titled 'Now try solving the problem on your own. (Draw your thinking on the canvas below)'. It features a canvas with a hand-drawn diagram of a satellite in a circular orbit around a central blue sphere, with arrows indicating the direction of motion and force. Below the canvas are 'Canvas Instructions' and a 'Submit' button. At the bottom, a green bar indicates 'Submission complete! Check your answer in the "Answer / Explanation" tab above.'



Topic

Uniform circular motion — varied prior conceptions make it well suited to a multi-role agent.



Participants

46 eleventh-grade students in South Korea; 16 joined one-on-one usability interviews.



Session

One 50-min lesson: pre-test (15 m) → AI-agent assisted learning (~20–25 m) → post-test (15 m).

Five inquiry stages

Introduction → Think → Experiment → Reflect → Apply

The agent as peer or tutor at each stage.

Two rounds of expert validation were conducted **with five specialists** (2 educational technology, 2 science education, 1 computer science).

ROUND 1

Dialogue act items: $M \geq 3.2$ (CVI 0.6–1.0; IRA .69)

Interactions with chatbot: $M \geq 3.4$

except “activation of prior knowledge” (M 2.4, CVI 0.2)
— a strict threshold caused repetitive dialogue loops.

Refinement:

supplemented the dialogue acts, clarified the distinction of roles within the chatbot system, and provided a background drawing in the answer box.

ROUND 2

Dialogue act items: $M \geq 3.4$ (CVI = 1.0 for all except PEER_HS; IRA .92)

Intent-Dialog act alignment: $M \geq 3.0$ (CVI 0.6–1.0; IRA .81)

Refinement:

clarified stage-specific intents and explicitly linked utterance analysis to act selection in the prompts.

① Role Adherence

Does one shared architecture actually produce **two distinct roles**?

→ BERT-Score on 100 sampled dialogue pairs to measure design fidelity - semantic similarity within vs. between roles

Data: Chatbot dialogue logs of 46 11th grade students in South Korea (using the system for approximately 25 minutes)

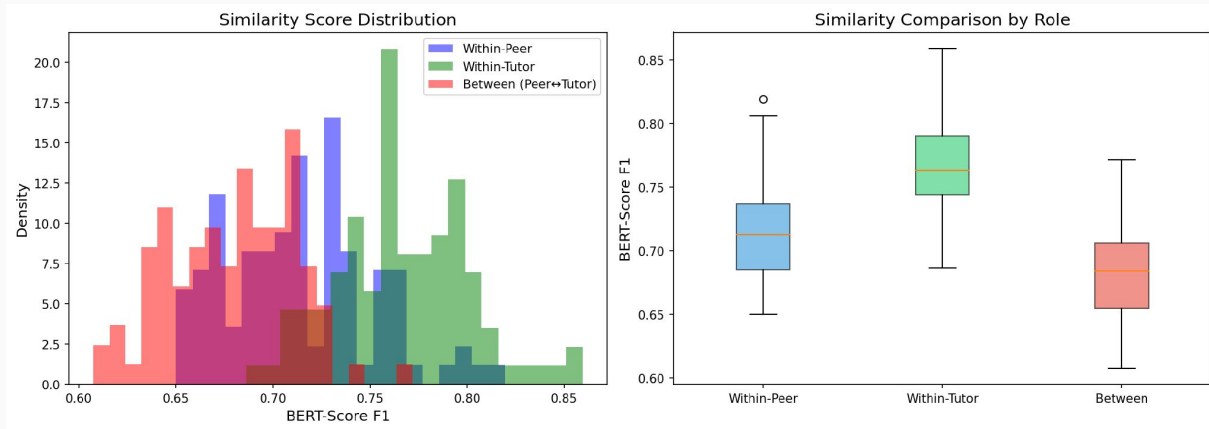


Fig. 4. BERT-Score F1 for within-role (Peer, Tutor) vs. between-role (Peer ↔ Tutor).

Results

Within-Peer M = 0.71

Within-Tutor M = 0.77

Between roles M=0.68

→ Within-Tutor > Within-Peer
> Between roles

→ Agents are semantically **distinct**.

② Rubric based evaluation

A rubric was adapted from Scarlatos et al. (2025) to evaluate generated utterances during Phases 2-4.

Peer: Guide Hypotheses(GH), Deepen Reasoning(DR), Cognitive Conflict(CC), Encouraging(Enc).

Tutor: Accuracy(Acc), Progress(Pro), Error-ID(EI), Strategic Hinting(SH).

Binary per criterion + a 5-point overall.

Table 1. Rubric scores across learning phases (peer criteria in Phase 2-3, tutor criteria in Phase 4).

Phase	GH	DR	CC	Enc	Acc	Pro	EI	SH	Overall
Phase 2	.797	.891	.516	.969	—	—	—	—	4.03
Phase 3	.873	.441	.126	.983	—	—	—	—	3.32
Phase 4	—	—	—	—	.935	.543	.935	.783	3.94

Bias-controlled LLM-as-judge: Since the system runs on GPT-4.1, evaluation was conducted by Claude-Sonnet-4.5 to avoid self-preference. An automated prompt was tuned to match human ratings; human IRA was 0.82 / 1.0 / 0.93 across Phases 2–4 before scoring the full set (335 / 165 / 46 pairs).

① Changes in science conceptual understanding

Science conceptual understanding was measured with a validated open-ended item asking students to explain the relationship between motion and net force direction in uniform circular motion, scored using a three-criteria rubric.

In this single classroom deployment, students showed a significant pre-post gain in conceptual explanation scores; this is reported as exploratory evidence, not causal effectiveness.

Table 2. Paired *t*-test results for pre–post scores in science conceptual understanding

Category	N	Pre-test		Post-test		<i>t</i>	<i>d</i>
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Science conceptual understanding	46	4.67	3.13	5.83	2.87	2.49*	.37

* $p < .05$

② Changes in attitudes toward science

Attitudes toward science were measured using a Korean-adapted 5-point Likert questionnaire covering value/enjoyment in science and reverse-scored science anxiety, with acceptable reliability for both subscales.

Students also reported more positive short-term attitudes after the activity, but these affective results should be interpreted cautiously due to self-report and novelty effects.

Table 3. Wilcoxon signed-rank test results for changes in attitudes toward science

Category	N	Pre-test		Post-test		z	r
		M	SD	M	SD		
Value and enjoyment in science	46	4.10	0.63	4.37	0.63	4.53***	.67
Science anxiety1)	46	4.06	0.76	4.35	0.93	3.76***	.56
Overall	46	4.08	0.54	4.36	0.60	4.39***	.65

*** $p < .001$

1) Science anxiety items were reverse-scored.

From 16 one-on-one usability interviews, three findings emerged:



Trust over speed

Compared to general-purpose LLM(e.g. ChatGPT), the educational AI system felt more trustworthy — fewer hallucination worries, no over-explaining.

Trade-off: conversation felt less “fast and smooth.”(Student E)



UI friction

Long text blocks were hard to scan at a glance, and revisiting responses from earlier stages was difficult.

Future improvements are required (adjust size and placement of the chatbot dialogue window, enable easier access to dialogue history).



Persona customization

Students wanted to choose the peer's interaction style — e.g., focused-and-serious vs. lighter, exploratory.(Student G)

White-Box AI Design with Educational Intent

The system transcends black-box LLM limitations by embedding explicit pedagogical intentions into a Dialogue Management System, ensuring both technical and educational validity through expert-validated design.

Role differentiation within one architecture

BERT-Score patterns were consistent with the intended role differentiation: tutor responses tended to be more convergent, while peer responses showed greater exploratory variability.

Trust–naturalness trade-off

The predefined educational scope may support learners' trust in the system, but it can also make interactions feel less natural than general-purpose LLM conversations.

A Pedagogical Policy for Multi-Role Agents

PF-aligned dialogue policy: link learner intent indicators, role-specific AI acts, and inquiry phases.

Structured coordination of peer & tutor roles across predefined inquiry phases, while **maintaining one continuous learning interaction.**

Architecture Scalability

Domain-transferable design: modify intent taxonomy & act inventory for any subject.

E.g., Economics: supply/demand misconceptions → peer agent induce conflict → tutor agent explains.

Dialogue policy framework enables multi-role agents across diverse educational contexts.



Limitations

Deterministic policy

Rule-based 1:1 intent-to-act may oversimplify complex states; an intentional trade-off for interpretability and alignment with PF.

Bounded validation context

The system was implemented in one physics topic with 46 high-school students and approximately 25 minutes of use. Thus, validation remains bounded to a short, single-topic classroom deployment.

Limited effectiveness inference

Because this study focused on chatbot development, the pre–post outcomes should not be interpreted as causal evidence of educational effectiveness.



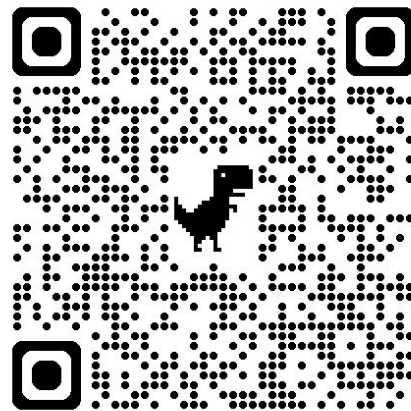
Future Work

Multi-agent orchestration. An orchestration layer monitors learner state in real time and composes roles dynamically — e.g., invoking tutor-like acts mid-exploration when a misconception persists — while preserving the overall PF sequence.

Rigorous comparison. Benchmark multi-role agents against single-role agents and general-purpose LLMs to isolate the effect of role design.

Peer and Tutor in One:

A Productive Failure-Based Architecture
for Multi-Role Conversational Agents in
Physics Learning



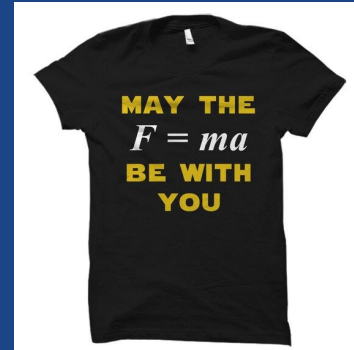
Junbo Koh (gtkobo92@snu.ac.kr)

Minsun Cho (moon05@snu.ac.kr)

Sunyoung Keum (keum0815@kice.re.kr)



May the Physics be with you:)



Junbo Koh (gtkobo92@snu.ac.kr)

Minsun Cho (moon05@snu.ac.kr)

Sunyoung Keum (keum0815@kice.re.kr)

Peer acts scaffold thinking; **Tutor** acts deliver feedback and consolidation.

PEER Phase 1 · Generation & Exploration

Peer_PR · Problem Recognition

Prompts exploration of the key elements needed to solve the problem.

Peer_HS · Hypothesis Setting

Encourages diverse exploration of possible explanations and predictions.

Peer_HV · Hypothesis Verification

Helps logically link a hypothesis to simulation / experiment results.

Peer_CC · Cognitive Conflict

Surfaces discrepancies between prior conceptions and scientific concepts.

Peer_ES · Emotional Support

Encourages the learner to persist and not give up.

TUTOR Phase 2 · Consolidation & Knowledge Assembly

Tutor_SP · Specification Request

Guides expressing intuitive understanding via learned concepts / principles.

Tutor_HI · Explicit Hint

Provides concrete guidance toward the answer when understanding is low.

Tutor_EC · Error Correction

Identifies the misconception, points out what is wrong, and redirects.

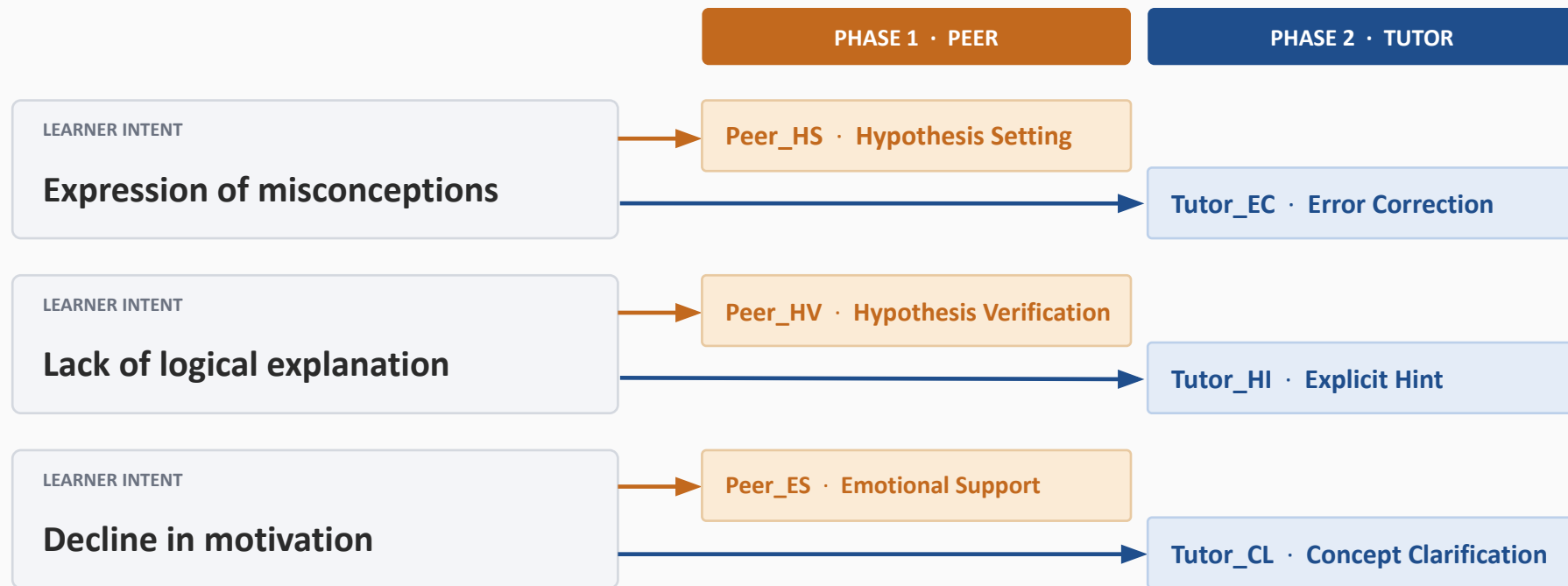
Tutor_CL · Concept Clarification

Detailed explanations that clarify and connect the target concepts.

Tutor_TR · Transfer Promotion

Guides generalizing the learning beyond the specific case.

Same intent, **different acts** — the act is decided by **phase and role**.



Rule-based 1:1 mapping · the highest-ranked indicator picks one act.