Cognitive Engine: A Gesture-Driven Multimodal Learning Interface for Embodied Cognitive Training and Reflective Visualization

Tianyi Chen

University of Hong Kong, Pok Fu Lam Road, Hong Kong
m18857398643@163.com

Abstract. Most AI-supported learning systems evaluate correctness and deliver content, but provide little visibility into how a learner is thinking. This paper presents Cognitive Engine, a work-in-progress multimodal learning interface that models cognitive processes in real time through a tightly coupled pipeline of gesture-based state sensing, adaptive dialogue, and interactive visualization. Learners express cognitive states—understanding, confusion, agreement, or frustration—through lightweight hand gestures captured via a webcam-based hand-landmark recognizer. Each recognized gesture triggers a coordinated response: the AI agent adjusts its dialogue strategy, and the "Cognitive Constellation" visualization adds or transforms visual objects to externalize the current reasoning state. Four training modes target distinct cognitive processes (memory, transfer, reflection, and Socratic dialogue), and a persistent report archive supports longitudinal tracking. The system contributes a novel design space that shifts from answer-based evaluation to process-aware cognition modeling.

Keywords: HCI · multimodal interaction · gesture recognition · explainable AI · learning interfaces · cognitive modeling

1 Introduction

Intelligent tutoring systems have demonstrated consistent effectiveness for structured domains, but their interaction model remains predominantly answer-based: the learner produces a response, the system evaluates it, and feedback is delivered [1, 2]. This cycle prioritizes correctness while leaving the processes of reasoning, confusion, and metacognitive reflection largely invisible to the system.

Learning sciences research is clear that effective learning depends on those processes. Self-explanation, iterative reasoning, and metacognitive monitoring are strong predictors of conceptual understanding and knowledge transfer [4, 5]. When AI interfaces cannot sense these states, they cannot adapt to them. Learners experiencing genuine confusion may receive content calibrated for confident understanding, and vice versa.

Cognitive Engine addresses this gap through three design commitments. First, cognitive state is treated as a primary input signal, expressed through

gesture rather than typed language, which reduces the friction of explicit self-report. Second, every detected state immediately drives both the AI dialogue and a spatial visualization of the learner's reasoning trajectory. Third, sessions are recorded and archived, making cognitive development inspectable over time rather than ephemeral.

The contribution of this demo is not any individual component—gesture recognition, LLM-based tutoring, and learning visualization each have prior precedents—but their integration into a unified pipeline in which embodied input, adaptive dialogue, and persistent visualization reinforce one another.

2 System Overview

Cognitive Engine is organized around three entry points: starting a session, accessing the report archive, and viewing achievements. A session begins with material upload. Learners supply documents, images, audio, video, or web links in any combination; the system performs multimodal analysis to construct an internal knowledge representation that serves as the grounding context for all subsequent AI interaction. This avoids constraining learners to predefined question banks [3].

On entering training, a central "core concept" star is placed in the Cognitive Constellation canvas, and the AI agent generates an opening dialogue turn calibrated to the selected training mode. From this point, the session proceeds as a continuous loop: the learner responds in text, the agent replies and optionally adds visual objects to the canvas, and gesture input can alter the trajectory at any moment. At session end, the system generates a structured cognitive report and a nebula animation summarizing the interaction, both stored persistently.

3 Gesture-Based Cognitive State Sensing

3.1 Capture Mechanism

Gesture recognition is handled by a dedicated webcam component that runs continuously in a picture-in-picture panel during training. The component uses real-time hand landmark detection to classify the learner's dominant-hand posture into one of five gesture categories. Recognition is performed locally in the browser without any data leaving the device, addressing privacy concerns about continuous video capture [6].

A cursor driven by hand position allows gestural navigation of the interface for learners who prefer to avoid keyboard interaction, using a pinch gesture as a click equivalent.

3.2 Gesture-to-State Mapping

Table 1 summarizes the five recognized gestures, the cognitive state each encodes, the adaptive dialogue action it triggers, and the corresponding change to the Cognitive Constellation. The mapping was designed to minimize learning cost: all five gestures are common, culturally recognizable hand postures.

Table 1. Gesture-to-cognitive-state mapping and system responses.

Gesture	Cognitive state	Dialogue action	Visual event
Fist (1–2×)	Understood	AI advances to next concept	New star added
Fist (3+×)	Deep insight	AI deepens the current concept	Supernova triggered
Open palm	Confused	AI re-explains at reduced complexity	Nebula added
Thumbs up	Positive / clarified	AI reinforces and extends	Nebula resolves to star
Thumbs down	Frustrated / blocked	Negotiation panel opens; AI offers alternatives	Black hole replaces current object
ILoveYou (hold 2 s)	Session end	Report generation initiated	Nebula animation

3.3 Adaptive Dialogue and Visualization Coupling

Each gesture detection event produces a coordinated, simultaneous response in both modalities. When a learner signals confusion (open palm), a nebula object is added to the canvas at a randomized position and the AI agent receives a prepended context cue instructing it to simplify its next response. When understanding is signaled (fist), a star is placed in the constellation and the agent is prompted to advance. When frustration is signaled (thumbs down), a black hole appears at the centre of the canvas, the interface enters a *Negotiation* state, and the agent is instructed to offer alternative explanations or problem framings before continuing.

Accumulating three or more fist gestures within a session triggers a supernova event—a large, animated object that marks a moment of deep insight. This threshold was set to distinguish genuine sustained understanding from single confirmatory acknowledgements.

The coupling between dialogue and visualization is intentional: the canvas provides a persistent, spatial record of state transitions that the learner can inspect at any moment, reducing reliance on linear chat history as the sole source of feedback.

4 Cognitive Training Modes

Cognitive Engine provides four training modes that target distinct dimensions of the learning process, reflecting a design decision to treat cognition as multi-dimensional rather than uniform [4, 5].

Memory training. Learners reconstruct information from memory without direct reference to source materials. The AI agent uses retrieval prompts and

spaced recall challenges to support long-term retention consistent with testing-effect research [4].

Transfer training. Learners apply knowledge across novel contexts or domains. The agent generates analogical scenarios and asks learners to reframe what they know, supporting flexible abstraction [5].

Reflection training. The agent prompts learners to examine their own reasoning strategies and assumptions rather than arriving at correct answers. This mode targets metacognitive awareness and self-regulated learning [4].

Dialogue training. The agent adopts a Socratic stance, posing counter-arguments and iterative questions to develop argumentation and deepen conceptual understanding [5].

5 Visualization, Reporting, and Motivation

5.1 Cognitive Constellation

The Cognitive Constellation is the primary visualization layer. It renders an interactive canvas in which visual objects—stars (clear understanding), nebulae (confusion or uncertainty), black holes (misunderstanding or frustration), and supernovae (deep insight)—are positioned and animated to reflect the real-time cognitive state of the session. The canvas is visible throughout training, so the learner and the system share a spatial representation of the reasoning trajectory. Objects added during the session are connected by edges to the central concept star, forming a graph structure that encodes the progression and resolution of cognitive states over time.

5.2 Cognitive Report

At session end, the system generates a structured cognitive report comprising: (i) a prose summary of the session; (ii) quantitative indices for knowledge mastery, reasoning depth, and reflection frequency; (iii) identified misconception patterns; (iv) qualitative strengths and areas for growth; and (v) suggested next steps. Reports are stored in a persistent archive, enabling learners to track their cognitive development longitudinally across sessions—an important departure from AI learning interactions that are typically ephemeral [3].

5.3 Achievement System

To support sustained engagement, the system includes an achievement mechanism that rewards meaningful learning behaviors: completing a session, resolving a black hole through negotiation, generating a supernova, and using all gesture types within a single session, among others. Following design guidance for educational gamification, rewards are tied to cognitive progress rather than superficial interaction counts [7].

6 Discussion and Future Work

The primary novelty of Cognitive Engine is the tight three-way coupling of gesture-based cognitive state sensing, adaptive AI dialogue, and spatial visualization. Prior intelligent tutoring research has demonstrated the value of student modeling and adaptive scaffolding [1, 5], but these systems typically infer state from task performance metrics. Cognitive Engine instead treats embodied, real-time gesture as a low-friction channel through which learners voluntarily signal their own cognitive states—an approach that preserves learner agency while providing the system with richer moment-to-moment signal than text interaction alone.

From an HCI perspective, embedding gesture interaction directly into the learning process—rather than as a separate control layer—means that expressive interaction and content engagement occur simultaneously. The Cognitive Constellation makes this coupling visible to the learner, creating a shared artifact that both reflects and reinforces the reasoning process.

The current implementation has limitations that the demo will surface. The gesture vocabulary is limited to five postures; learners with limited hand mobility may find the interface inaccessible, and the mapping of postures to cognitive states has not yet been empirically validated against self-report or physiological measures.

Future work will proceed in three directions. First, a controlled study will examine whether gesture-based state signaling improves learning outcomes, metacognitive accuracy, and engagement relative to a text-only baseline. Second, the cognitive state model will be extended to incorporate hesitation timing, response latency, and affect signals, moving toward a richer multimodal learner model. Third, deployment in higher education and professional training contexts will test whether the transfer and reflection modes generalize beyond the controlled settings in which the current prototype was developed.

Disclosure of Interests. The author declares no competing interests relevant to this article.

References

1. Aleven, V., Koedinger, K.R.: An effective metacognitive strategy: learning by doing and explaining with a cognitive tutor. Cognitive Science **26**(2), 147–179 (2002)
2. Piech, C., Bassen, J., Huang, J., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
3. Holstein, K., McLaren, B.M., Aleven, V.: Designing for complementarity: teacher–AI collaboration in the classroom. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2019)
4. Roediger, H.L., Karpicke, J.D.: Test-enhanced learning: taking memory tests improves long-term retention. Psychological Science **17**(3), 249–255 (2006)
5. Bransford, J.D., Brown, A.L., Cocking, R.R.: How People Learn: Brain, Mind, Experience, and School. National Academy Press, Washington, DC (2000)

6. Oviatt, S.: Multimodal interfaces. In: The Human-Computer Interaction Handbook, pp. 413–432. CRC Press (2012)
7. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness. In: Proceedings of the 15th International Academic MindTrek Conference, pp. 9–15 (2011)

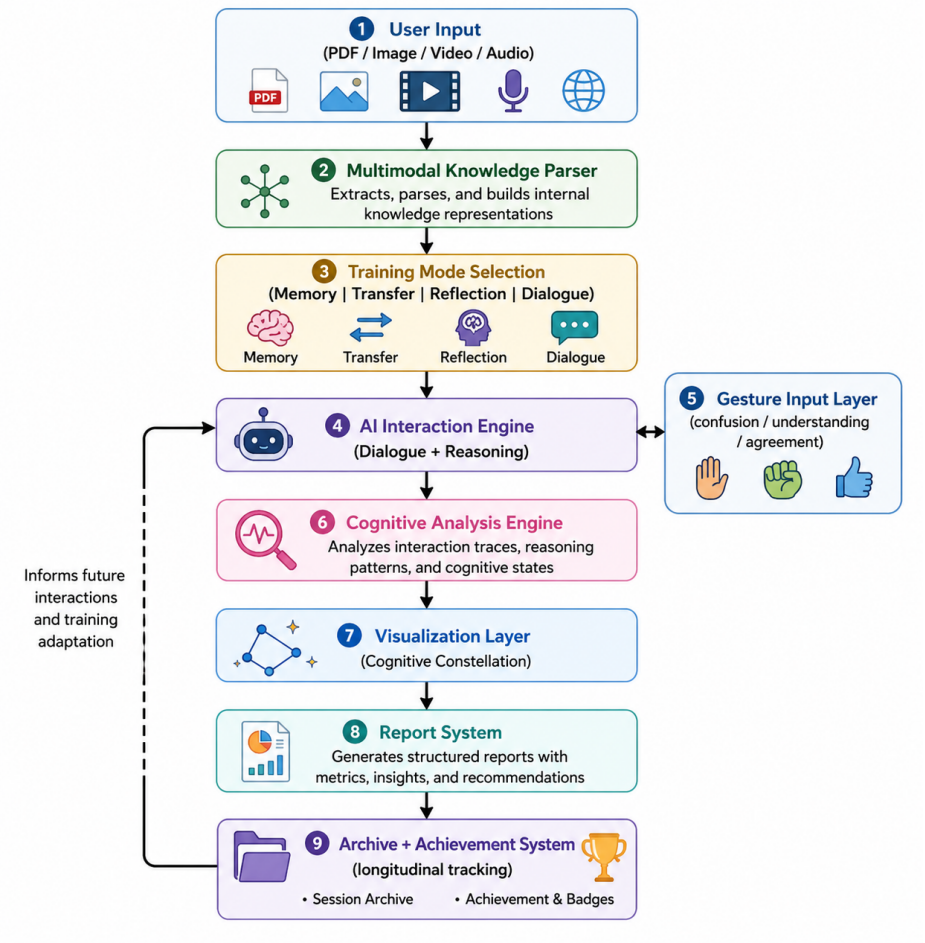

Fig. 1. System architecture of Cognitive Engine. User-provided multimodal materials are analyzed to construct a knowledge representation grounding subsequent AI interaction. During training, gesture input from the webcam-based hand recognizer triggers coordinated updates to both the AI dialogue and the Cognitive Constellation visualization. At session end, a structured cognitive report is generated and archived to support longitudinal reflection.