From Objective to Subjective: A Benchmark for Virtual Student Abilities

EduPersona: Evaluating Subjective Abilities in Educational AI

Buyuan Zhu, **Shiyu Hu**, Yiping Ma, Yuanming Zhang, **Kang Hao Cheong**

AAAI AI4EDU Workshop, 2026 • Singapore

School of Physical and Mathematical Sciences, NTU Singapore

Shanghai Institute of AI for Education, ECNU

State Key Laboratory of Robotics and Systems, HIT

College of Computing and Data Science, NTU Singapore

Background & Motivation

Virtual student agents are increasingly used for classroom simulation and teacher training, offering controllable and reproducible environments for studying teaching strategies.

The Gap in Current Evaluation

✓ Well-Studied: Objective Abilities

- QA accuracy
- Knowledge correctness
- Question generation quality

⚠️ Overlooked: Subjective Abilities

- Emotional responses
- Personality traits
- Behavioral authenticity

Research Question: How can we systematically evaluate subjective abilities essential for authentic classroom interaction?

Our Contributions

1. Large-Scale Benchmark

2 languages, 3 subjects, 10 Big Five personas
1,308 rounds, 12,814 Q&A turns → 128k+ samples

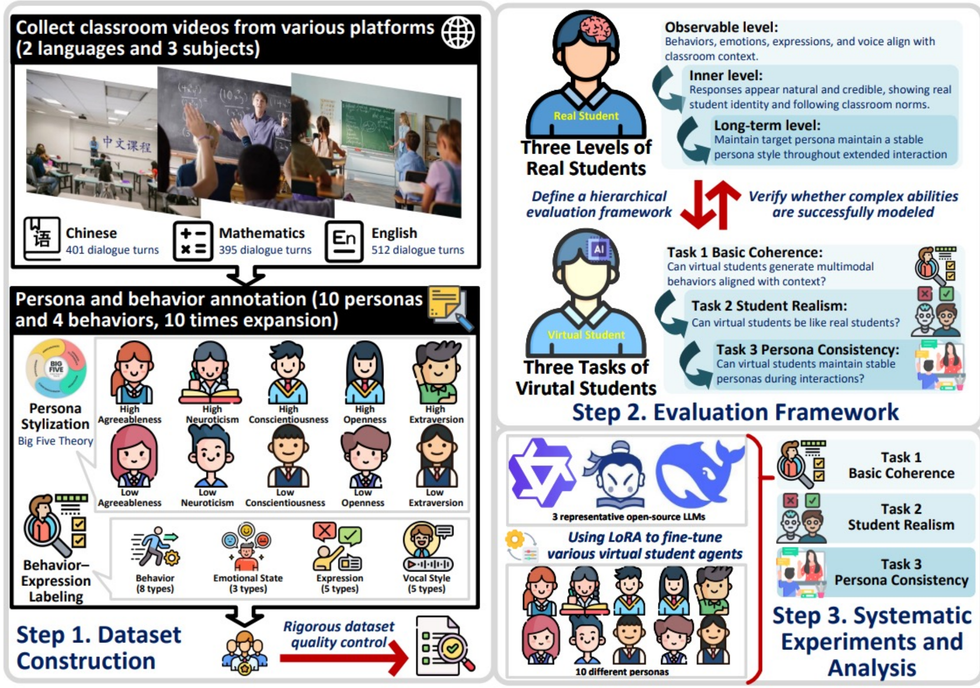
2. Three-Task Framework

Progressive evaluation:
Label Coherence → Student Realism → Persona Consistency

3. Systematic Evaluation

Experiments on 3 representative open source LLMs + 30 fine-tuned variants show:

- **Task 1 (Coherence): +33.6%**
- **Task 2 (Realism): +30.6%**
- **Task 3 (Consistency): +14.9%**



Workflow Overview of EduPersona Benchmark, consisting of three steps: dataset construction; a three-task evaluation framework and systematic experiments and analysis.

Dataset Construction

Persona Stylization

Each dialogue is rewritten (by GPT-4o) into **10 persona variants** (High/Low × 5 traits), preserving semantic meaning while adapting linguistic style, behavior patterns, and emotional expressions.

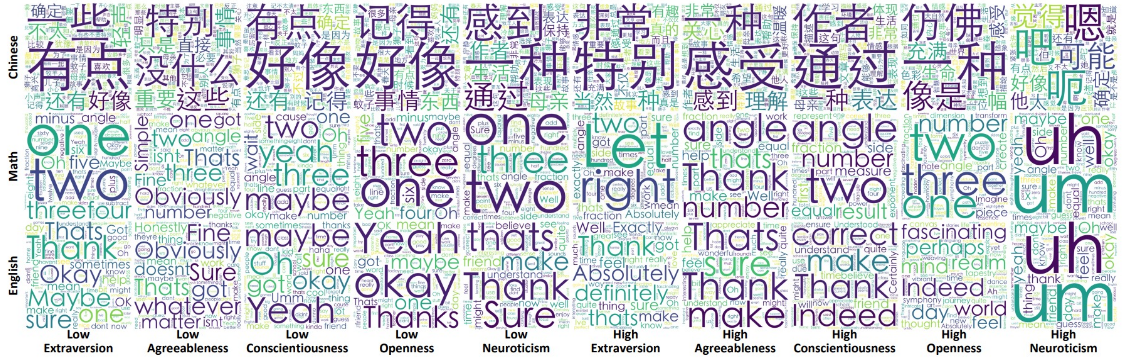
Extraversion (E)
Active participation vs. Reserved responses

Conscientiousness (C)
Organized & accurate vs. Careless responses

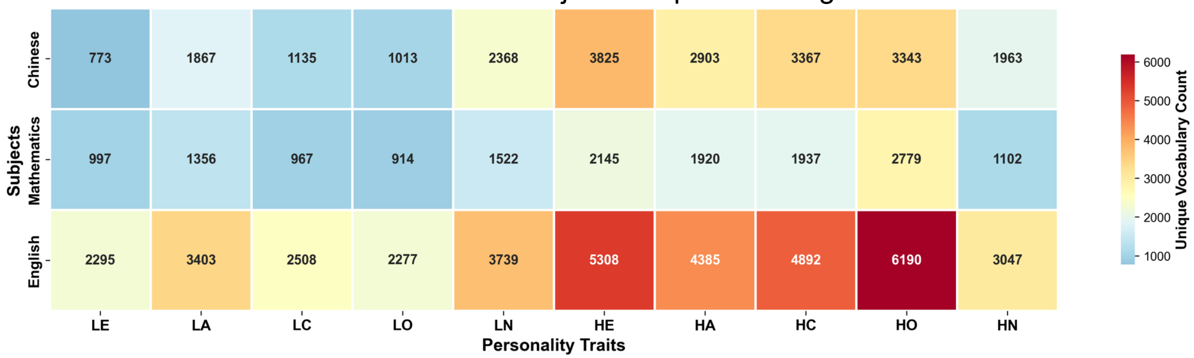
Openness (O)
Creative & curious vs. Conservative & traditional

Agreeableness (A)
Cooperative vs. Competitive behaviors

Neuroticism (N)
Anxious & hesitant vs. Calm & stable



Word Cloud of cross-subject and persona linguistic variation.



Vocabulary richness across subjects and personas.

Dataset Construction

Multi-Dimensional Coverage

Cross-Lingual

Chinese & English

Cross-Subject

- Chinese (401 rounds)
- Math (395 rounds)
- English (512 rounds)

Cross-Persona

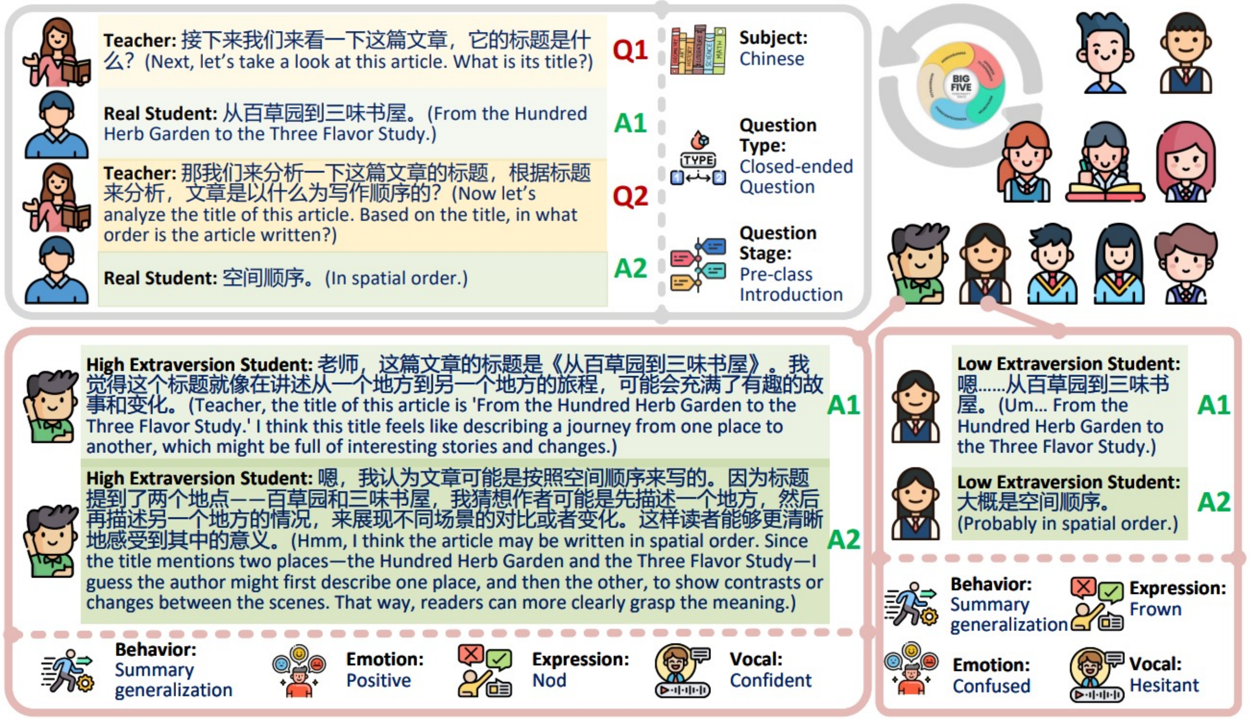
10 Big Five personas (High/Low × 5 traits)

Multimodal Labels

Behavior, Emotion, Expression, Voice (annotated by GPT-4o)

Dataset Statistics

1,308	12,814	128k+
Dialogue Rounds	Q&A Turns	After Expansion



Chinese classroom example with persona-conditioned responses. This example illustrates the full EduPersona pipeline (raw dialogue **->** persona stylization **->** behavior–expression labeling) and demonstrates how different personas yield distinct linguistic and non-verbal behaviors within the same teaching context.

Evaluation Framework: Three Progressive Tasks

Model Selection

Qwen3-8B
Strong CN-EN instruction following

InternLM3-8B
Robust in Chinese conversation

DeepSeek-R1-14B
Enhanced math & reasoning

Task 1: Basic Coherence

Question: Can virtual students generate multimodal behaviors aligned with context?

Metrics: Response Rate, Validity Rate, Label Accuracy (Behavior, Emotion, Expression, Voice)

Task 2: Student Realism

Question: Can virtual students behave like real students?

Evaluation: Using expert-derived criteria including linguistic naturalness, identity credibility, as prompts to guide API based evalutator

Task 3: Persona Consistency

Question: Can virtual students maintain stable personas during interactions?

Scope: Short-term (single-turn) and Long-term (10-turn classroom dialogues)

Evaluation Framework: Three Progressive Tasks

Model Selection

Qwen3-8B
Strong CN-EN instruction following

InternLM3-8B
Robust in Chinese conversation

DeepSeek-R1-14B
Enhanced math & reasoning

Fine-Tuning Strategy

Method: LoRA (rank=16, α=32)

Configuration: 3 base models × 10 personas = 30 fine-tuned variants

Training: AdamW optimizer, lr=3×10^{-4}, batch size 8, up to 5 epochs

Data Split: 60% training (D_t) / 40% testing (D_{test})

Evaluation Settings

Task 1: Quantitative metrics on behavior-expression alignment

Tasks 2 & 3: GPT-4o evaluator scoring realism and consistency

Task 1: Basic Coherence - Evaluation Metrics

Evaluating prediction $\hat{b}$ vs reference b across dimensions $I = \{beh, emo, exp, voi\}$

1. Availability & Validity

(1) Response Rate

→ *Ratio of non-empty outputs*

$$Resp = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}[\hat{b}_{t,i} \neq \emptyset]$$

(2) Validity Rate

→ *Ratio of valid-format outputs*

$$Valid = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{\sum_t \mathbf{1}[\hat{b}_{t,i} \in B_i]}{\max(1, \sum_t \mathbf{1}[\hat{b}_{t,i} \neq \emptyset])}$$

2. Accuracy Quality

(3) Raw Accuracy

→ *Correctness on non-empty samples*

$$Raw = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{\sum_t \mathbf{1}[\hat{b}_{t,i} = b_{t,i}]}{\max(1, \sum_t \mathbf{1}[\hat{b}_{t,i} \neq \emptyset])}$$

(4) Validated Accuracy

→ *Correctness on valid-format samples*

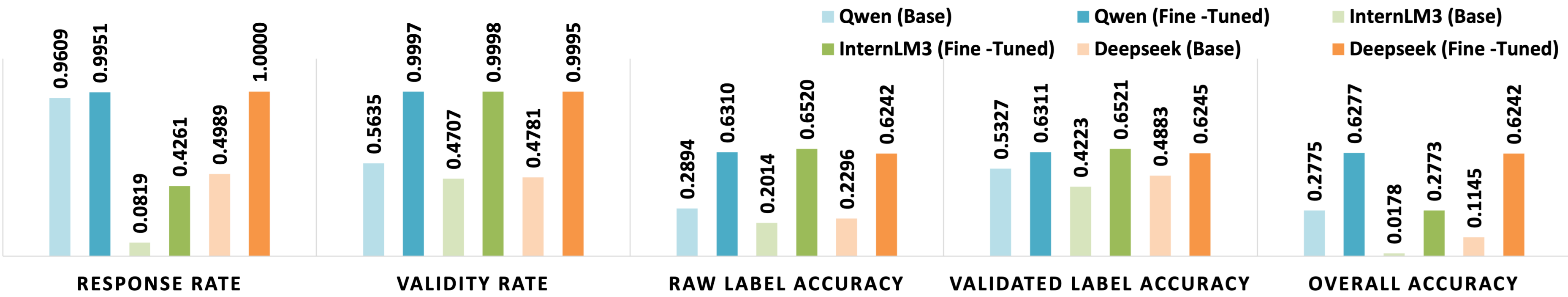
$$Val = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{\sum_t \mathbf{1}[\hat{b}_{t,i} = b_{t,i}]}{\max(1, \sum_t \mathbf{1}[\hat{b}_{t,i} \in B_i])}$$

(5) Overall Accuracy

→ *Strict end-to-end success rate*

$$All = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}[\hat{b}_{t,i} = b_{t,i}]$$

Results: Task 1 - Basic Coherence



Dimension-Level Analysis

Emotion	Expression	Voice	Behavior
Easiest	Medium	Medium	Hardest

✓ Persona fine-tuning markedly improves multimodal alignment (+33.6%)

✓ Qwen & DeepSeek achieve OverallAcc ~0.62 after fine-tuning

Results: Task 2 - Student Realism

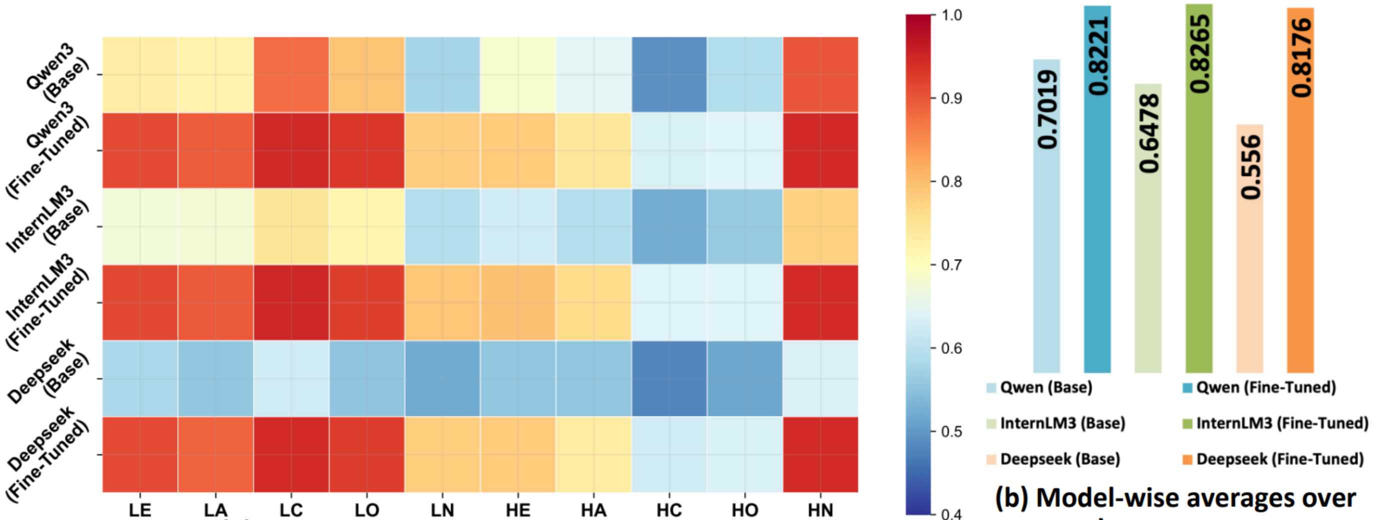
Persona-Specific Patterns

Higher Realism

High Neuroticism (0.891), Low Conscientiousness (0.887), Low Openness (0.871)

Lower Realism

High Conscientiousness (0.748), High Openness (0.764)



(a) Persona-wise consistency across mode

(b) Model-wise averages over ten personas

Overall Performance

✓ All models converge around 0.82 after fine-tuning (+30.6%)

✓ Persona conditioning harmonizes performance across model families

Results: Task 3 - Persona Consistency

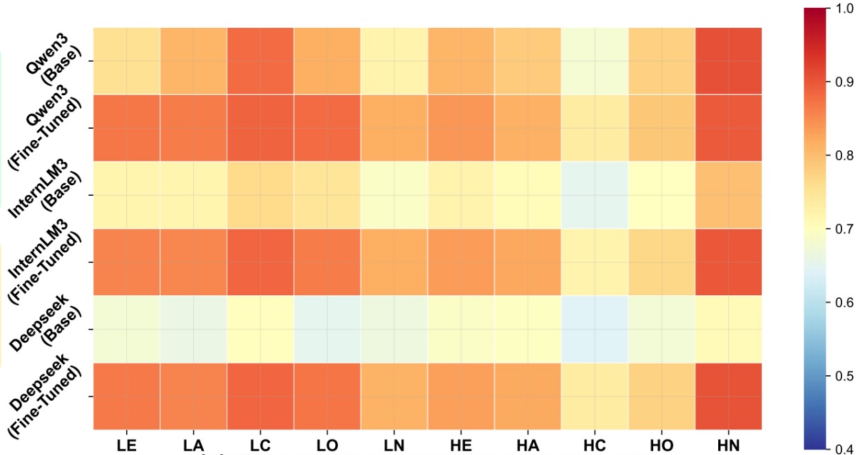
Persona-Specific Patterns

Higher Consistency

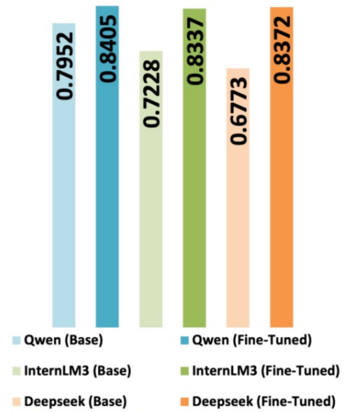
High Neuroticism (0.901), Low Conscientiousness (0.887), Low Openness (0.873)

Lower Consistency

High Conscientiousness (0.731), High Openness (0.779)



(a) Persona-wise consistency across mode



(b) Model-wise averages over ten personas

Overall Performance

✓ Fine-tuning brings all models to ~0.84 convergence (+14.9%)

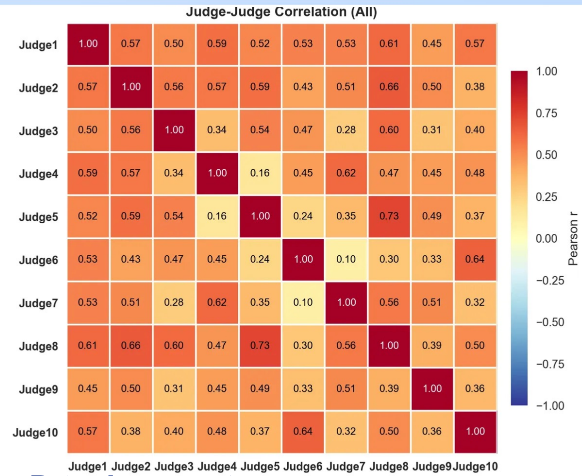
✓ Long-term stability: LoRA 0.920±0.042 vs GPT-4o 0.480±0.262

Human-AI Evaluation Alignment

GPT-4o serves as a scalable evaluator implementing expert-defined criteria for Tasks 2 & 3, validated against human judgment

Strong Positive Correlation

Larger performance improvements correlate with higher human-AI agreement, validating GPT-4o's reliability



Moderate Inter-Rater Agreement

Human judges show moderate consistency (mostly 0.4-0.6), reflecting subjective nature of realism assessment

Validation Results

✓ Scalable & Consistent

GPT-4o provides reproducible evaluation at scale

✓ Expert-Grounded

Criteria from 10 experts ensure pedagogical validity

Cross-Task Insights

Consistent Performance Progression Across Tasks

Task 1	Task 2	Task 3
~0.62	~0.82	~0.84
Basic Coherence	Student Realism	Persona Consistency

Clear progression: structural alignment → perceptual realism → long-horizon stability

Persistent Cross-Task Persona Patterns

Easier: HN/LC/LO
Hesitation and partial responses align with authentic student behaviors

Harder: HC/HO
Structured personas resemble default LLM outputs, reducing authenticity

Cross-Task Consistency Pattern

Observation: The same persona difficulty ranking (HN/LC/LO easier, HC/HO harder) persists across all three tasks, confirming that evaluation challenges stem from inherent persona characteristics rather than task-specific artifacts.

Subjective abilities depend on persona modeling, not model scale—revealing unique educational AI challenges.

Conclusion

EduPersona: First Large-Scale Benchmark for Subjective Abilities

First comprehensive benchmark evaluating virtual students across coherence, realism, and consistency—2 languages, 3 subjects, 10 personas, 1,308 rounds, 128k+ samples

Progressive Three-Task Framework

Task 1: Coherence
Multimodal alignment (→0.62)

Task 2: Realism
Authentic behaviors (→0.82)

Task 3: Consistency
Long-term stability (→0.84)

Key Experimental Findings

✓ Fine-Tuning Effectiveness
Consistent gains: +33.6%, +30.6%, +14.9% across all tasks

✓ Model Convergence
LoRA brings diverse models to similar performance bands

⚠️ Persona Hierarchy
HC/HO challenging (0.731–0.779); HN/LC/LO stable (0.873–0.901)

⚠️ Long-Term Stability
Fine-tuned 0.920±0.042 vs GPT-4o 0.480±0.262 over 10 turns

Impact: EduPersona establishes the first reproducible evaluation paradigm for human-like virtual student agents, providing systematic metrics and decoupled task framework to advance trustworthy AI in teacher training and educational research.

Future Work

1. Comprehensive Virtual Student Modeling

Current Progress:

EduPersona demonstrates improvements in subjective abilities: basic coherence (0.62), student realism (0.82), and persona consistency (0.84).

Remaining Gap:

Achieving truly holistic student simulation requires seamless integration of cognitive reasoning capabilities, emotional regulation mechanisms, and collaborative social learning behaviors.

Future Direction:

Develop unified multi-dimensional architecture that integrates knowledge state tracking, affective dynamics modeling, and authentic classroom interaction patterns.

2. Human-in-the-Loop Educational Applications

Real-World Validation:

Deploy virtual student agents in authentic teacher training programs and conduct controlled classroom experiments to evaluate real-world effectiveness and usability

Practitioner Feedback:

Systematically gather insights and feedback from practicing teachers to guide iterative model refinement, identify critical performance gaps, and ensure pedagogical validity.

Downstream Tools:

Co-design domain-specific applications with educators for teacher preparation and professional development.

Acknowledgements

This work is supported by the

MOE Tertiary Education Research Grant

Grant No. MOE2024-TRF-004

We thank all collaborators, annotators, educators, and the open-source community for their contributions to this research.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Association for the Advancement of Artificial Intelligence

Thanks for listening!

2026.01.26 in Singapore

BUYUAN001@e.ntu.edu.sg

SOEI *EduVerse*

Scan to see our series work

Scan to download the slides

WeChat of the first author