AAAI 2026
January 20 – 27, 2026
Singapore

Learning to Be Taught: A Structured SOEI Framework for Modeling and Evaluating Personality-Aligned Virtual Student Agents

Yiping Ma*, Shiyu Hu*, Xuchen Li, Yipei Wang, Yuqing Chen, Shiqing Liu+, Kang Hao Cheong+

Yiping Ma

- PhD in East China Normal Univeristy (ECNU)
- Visiting Student in Nanyang Technological University (NTU)
- mayiping98@163.com; 52275901020@stu.ecnu.edu.cn

Scan to download this slide

WeChat for the first author

华东师范大学 EAST CHINA NORMAL UNIVERSITY

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

东南大学 SOUTHEAST UNIVERSITY

Motivation： From Tutor to Student

Notable Platforms Leveraging LLM as virtual teachers for Student Support

Duolingo Max	Khan Academy	Socratic	TAL’s MathGPT	Youdao’s ZiYue	Squirrel AI
Duolingo integrates GPT-4 to offer personalized language practice and adaptive exercises, providing instant corrections and guidance to learners.	Khan Academy uses LLM to serve as a tutor, offering interactive feedback and real-time explanations to help students deepen their understanding across subjects.	Socratic offers intelligent, step-by-step explanations to students, particularly in science and mathematics, enhancing problem-solving skills.	TAL Education Team developed MathGPT to assist students with complex math problems, offering step-by-step breakdowns and personalized feedback.	Youdao offers a variety of personalized educational services, from homework assistance to interactive learning, targeting student needs in real time.	Squirrel AI uses LLMs to analyze student performance and deliver a highly personalized curriculum that adapts dynamically to student progress.

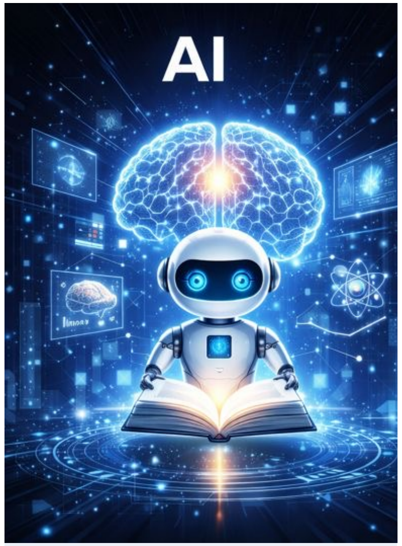
Motivation： From Tutor to Student

What if we shift our focus to the students...

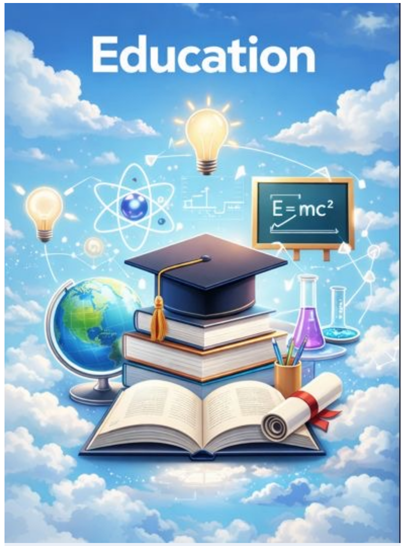
Role-playing(Digital Puppets) ➤ Programmatically Predefined ➤ LLMs-based Agent

- Insufficient authentic personality modeling
- Limited dynamic development mechanisms
- Absence of a systematic evaluation framework

Motivation： From Tutor to Student

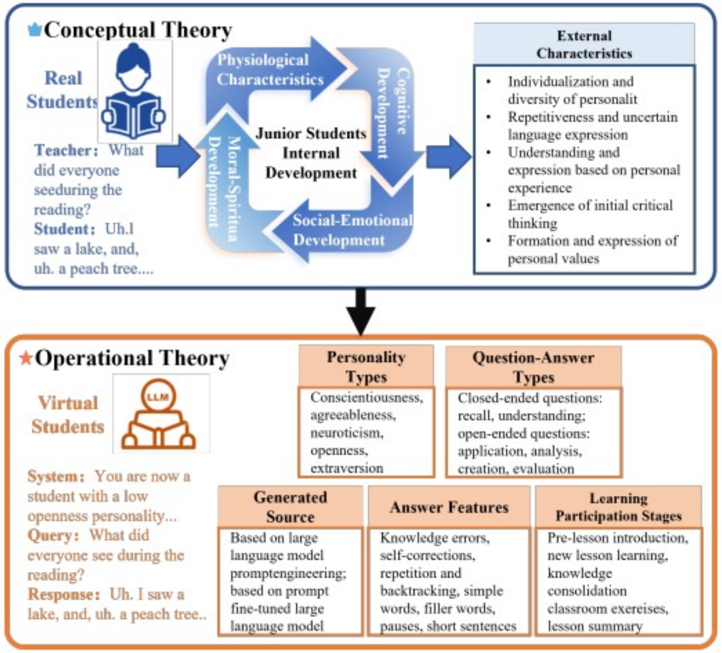


Design of Efficient, Transferable, Optimizable Model Architectures



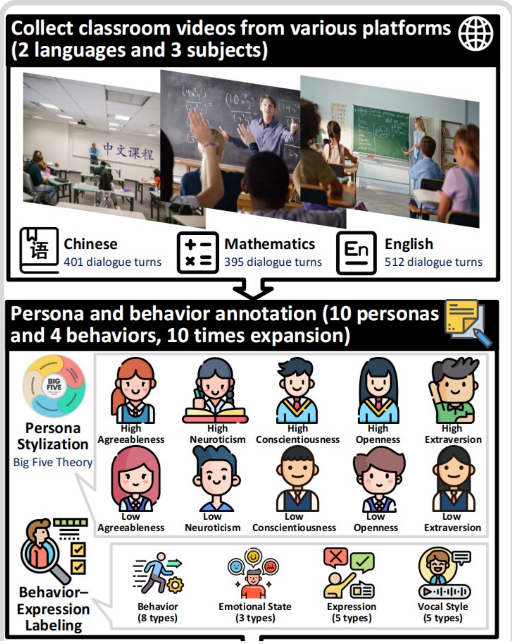
Interpretability of Student Development, the Fidelity of Individual Differences, the Intervenability of Instructional Interactions

Modeling virtual students is a genuinely **interdisciplinary challenge**, whose complexity lies not in the use of tools, but in the **integration of paradigms.**

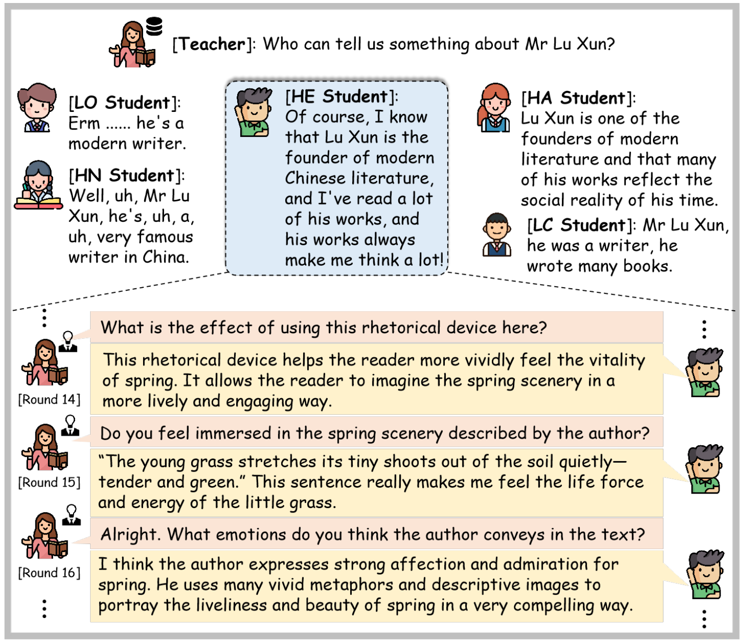


Our Series of Work

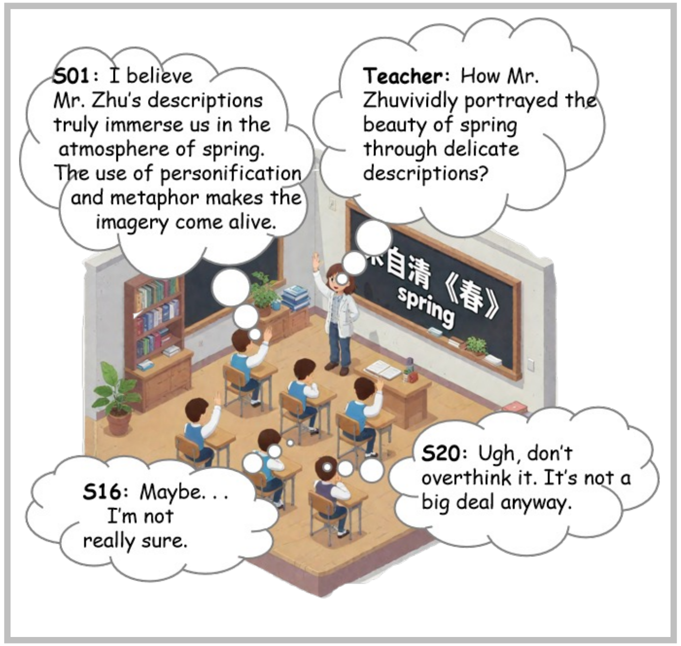
EduPersona



SOEI Framework



EduVerse



Data Level

What do real students look like?

Single-Agent Level

How do we model & evaluate one student?

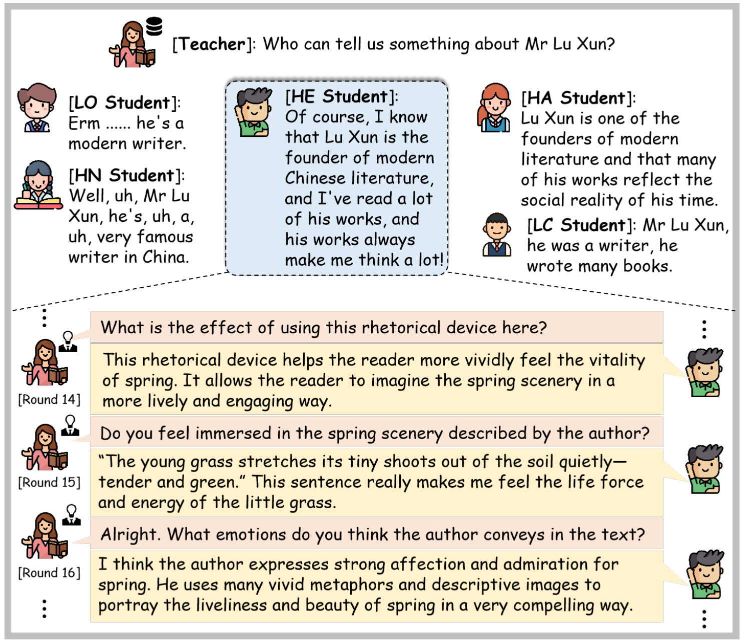
Multi-Agent Level

What emerges in classroom interaction?

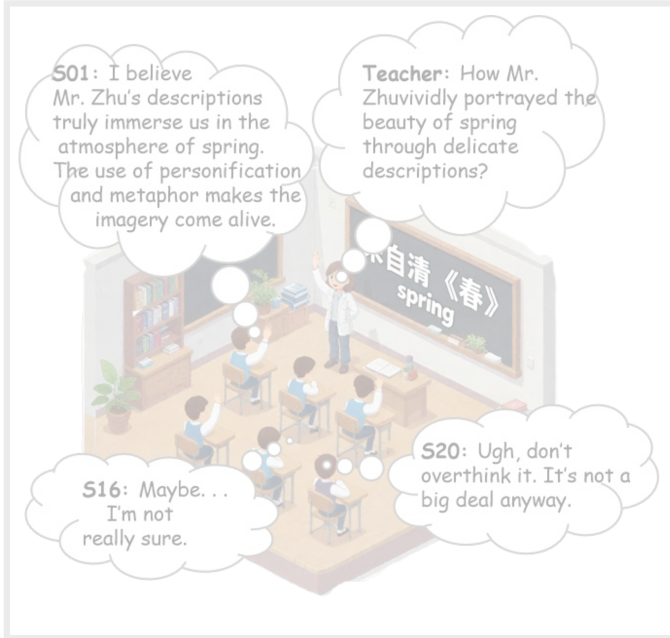
Our Series of Work

EduPersona

SOEI Framework



EduEval

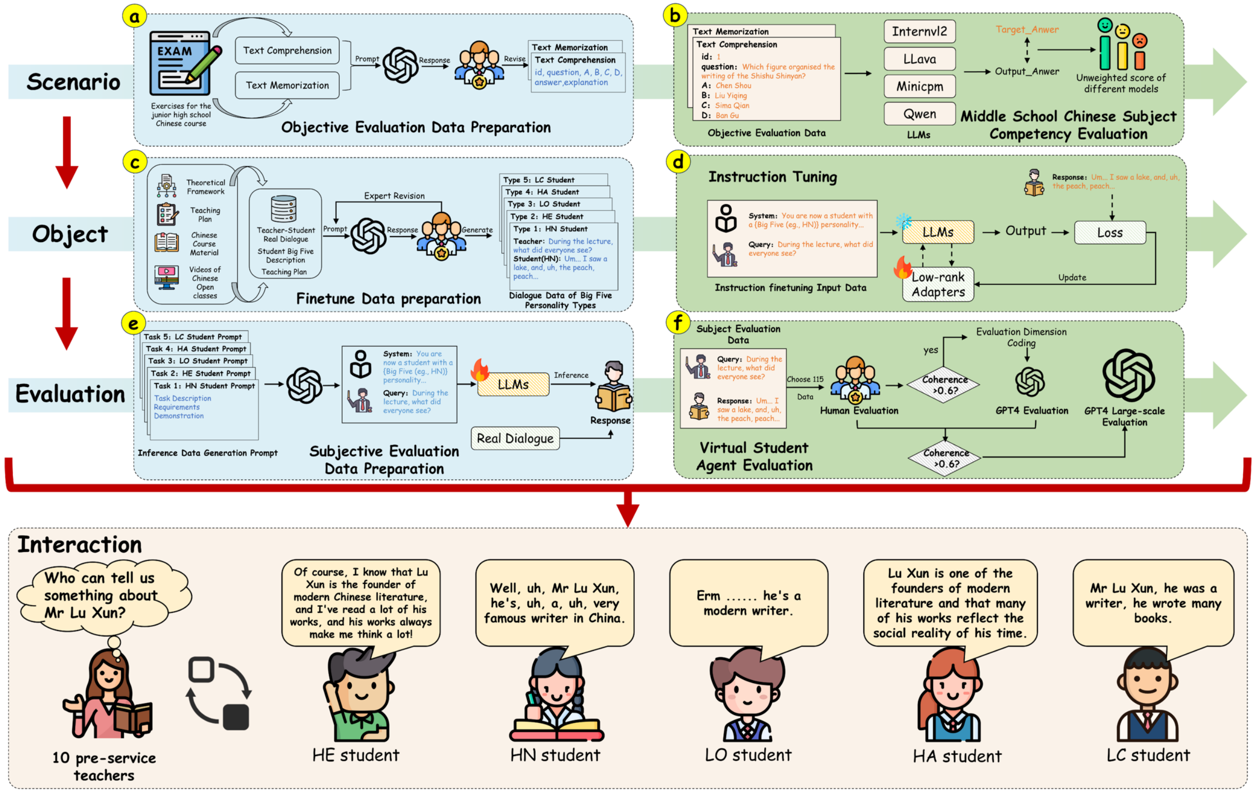


Data Level
What do real students look like?

Single-Agent Level
How do we model & evaluate one student?

Multi-Agent Level
What emerges in classroom interaction?

Research Outline



RQ1：In what scenarios do we model?

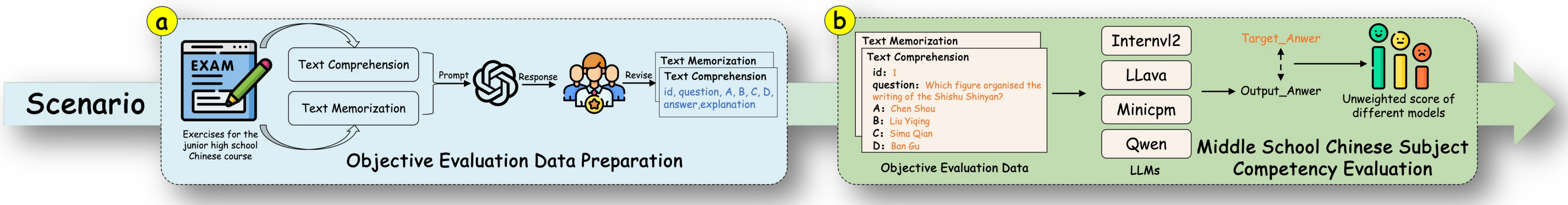
RQ2：What types of virtual students do we model?

RQ3：How do we scientifically evaluate the performance of these virtual students?

RQ4：How capable are virtual students in multi-turn interactions?

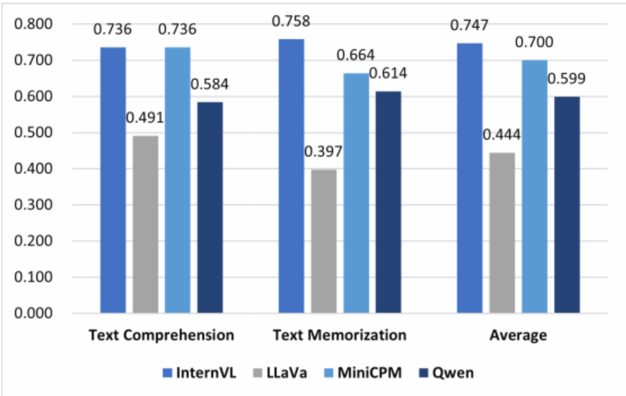
🎯 **Core Objective**: Building personalized virtual students with personality consistency, stylistic expression, and behavioral controllability

RQ1： In what scenarios do we model?



Scene Modeling: Structured Educational Task Design

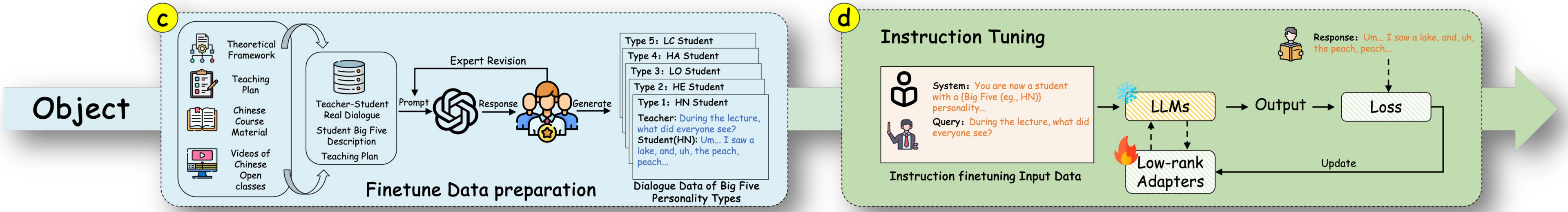
- **Five-Element Structure**: Course Content → Teaching Phase → Question Type → Language Style → Personality Traits
- **Real Classroom Grounding**: Based on authentic middle school Chinese language instruction

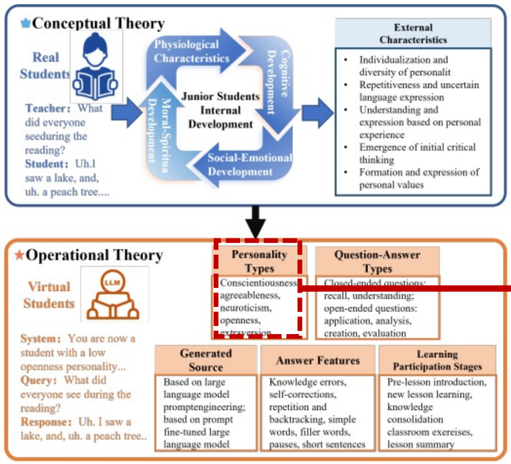


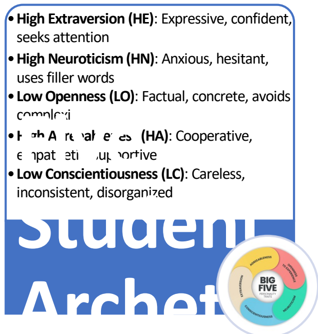
✅ **Model Performance on Chinese Language Tasks:**

- InternVL: 74.7% accuracy (comprehension: 73.6%, memorization: 75.8%)
- MiniCPM: 70.0% accuracy (comprehension: 73.6%, memorization: 66.4%)
- Demonstrates that foundation models can handle structured educational scenarios

RQ2: What types of virtual students do we model?



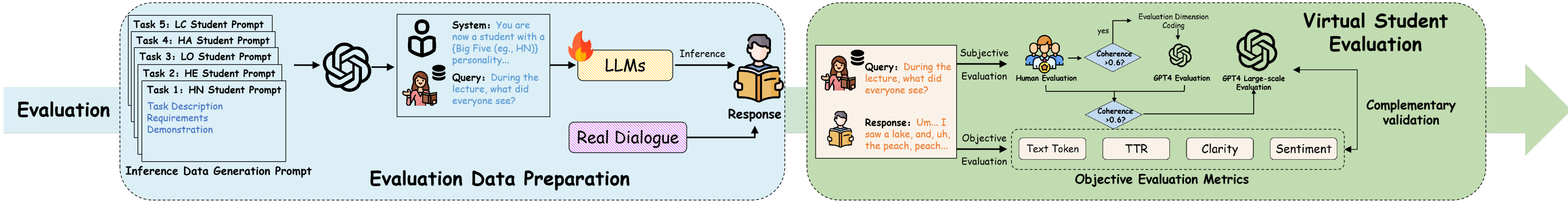


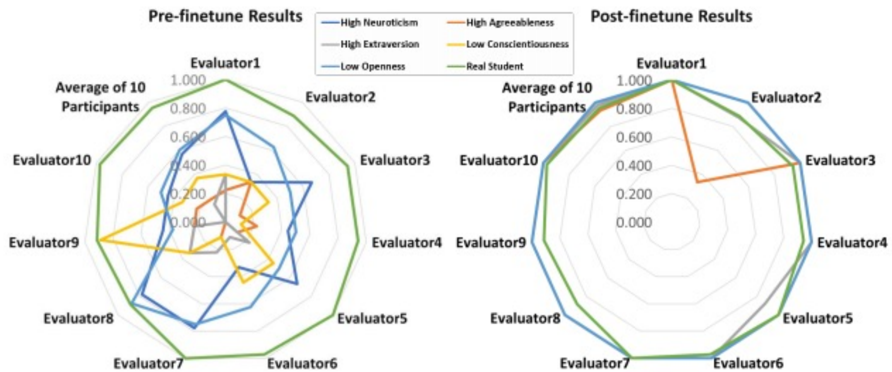


(a) HE (b) HN (c) LO (d) HA (e) LC

Based on Big Five personality theory, we construct five types of personalized virtual students, each with unique linguistic styles and cognitive characteristics.

RQ3: How do we evaluate the performance of these virtual students?





Multi-level Evaluation Mechanism:

- **Subjective Turing Test**: 10 experts evaluating virtual student vs the real student
- **GPT-4 Large-scale Scoring**: Scalable evaluation of 12,232 samples (Fleiss's Kappa =0.6806)
- **Objective Language Metrics**: Text length, TTR, sentiment polarity analysis

Using a hybrid evaluation system, results show post-fine-tuning virtual students are indistinguishable from real students.

RQ3：How do we evaluate the performance of these virtual students?

Table 2: The experiment results of different LVSA types.

	InternVL		LLaVa		MiniCPM		Qwen		Average		Student P-value
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	
HN	58.19%	94.31%	16.89%	80.45%	54.96%	94.62%	49.86%	94.62%	44.98%	91.00%	0.005**
HA	33.99%	81.19%	14.52%	66.46%	24.75%	73.93%	43.89%	80.86%	29.29%	75.61%	<0.001***
HE	44.64%	73.88%	12.69%	44.40%	19.78%	60.82%	30.34%	72.76%	26.86%	62.97%	0.002**
LC	54.96%	50.49%	18.69%	52.67%	30.16%	34.43%	30.82%	39.02%	33.66%	44.15%	0.294
LO	79.21%	91.33%	13.33%	92.33%	55.67%	88.00%	47.83%	83.67%	49.01%	88.83%	0.066**
Average	54.20%	78.24%	15.22%	67.26%	37.06%	70.36%	40.55%	74.19%	36.76%	72.51%	0.009**
Model P-value	0.058**		0.004**		0.013**		0.007**		0.006**		

Note: (1) **HE**, **HN**, **LO**, **HA**, and **LC** are abbreviations for **High Extraversion**, **High Neuroticism**, **Low Openness**, **High Agreeableness**, and **Low Conscientiousness** LVSA, respectively. (2) ** means significant; *** means highly significant.

Table 3: The experiment results of different learning stages.

	InternVL		LLaVa		MiniCPM		Qwen		Average		Learning P-value
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	
PI	64.59%	82.32%	20.06%	66.60%	48.45%	79.73%	49.48%	81.79%	45.65%	77.61%	0.012**
NL	54.71%	80.06%	16.59%	68.04%	38.04%	69.63%	41.41%	75.15%	37.69%	73.22%	0.008**
KC	49.71%	77.32%	11.86%	69.52%	37.29%	71.19%	37.76%	70.51%	34.16%	72.14%	0.011**
CE	48.63%	77.43%	16.14%	67.36%	33.89%	64.09%	37.37%	70.47%	34.01%	69.84%	0.006**
LS	55.00%	76.97%	12.25%	69.14%	32.92%	72.10%	39.50%	76.18%	34.92%	73.60%	0.012**
Average	54.53%	78.82%	15.38%	68.13%	38.12%	71.35%	41.10%	74.82%	37.28%	73.28%	0.009**
Model P-value	<0.001***		<0.001***		<0.001***		<0.001***		<0.001***		

Note: (1) **PI**, **NL**, **KC**, **CE**, and **LS** are abbreviations for **Pre-lesson Introduction**, **New Lesson Instruction**, **Knowledge Consolidation**, **Class Exercises**, and **Lesson Summary** students, respectively. (2) ** means significant; *** means highly significant.

Table 4: The experiment results of different question types.

	InternVL		LLaVa		MiniCPM		Qwen		Average		Question P-value
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	
CQ	58.61%	78.38%	17.72%	67.08%	43.19%	72.91%	47.64%	75.13%	41.79%	73.38%	0.015**
OQ	50.40%	79.23%	12.98%	69.22%	32.81%	69.67%	34.47%	74.51%	32.67%	73.16%	0.006**
Average	54.51%	78.81%	15.35%	68.15%	38.00%	71.29%	41.06%	74.82%	37.23%	73.27%	0.009**
Model P-value	0.117		0.041**		0.068		0.117		0.078		

Note: (1) **CQ**, **OQ** are abbreviations for **Closed-ended Question**, **Open-ended Question**, respectively. (2) ** means significant; *** means highly significant.

Evaluation Results with Different Personality Traits:

- Fine-tuning significantly improved average evaluation scores across five personality types (36.76% → 72.51%).
- Paired t-tests confirmed statistically significant improvements for all models ($p < 0.05$).

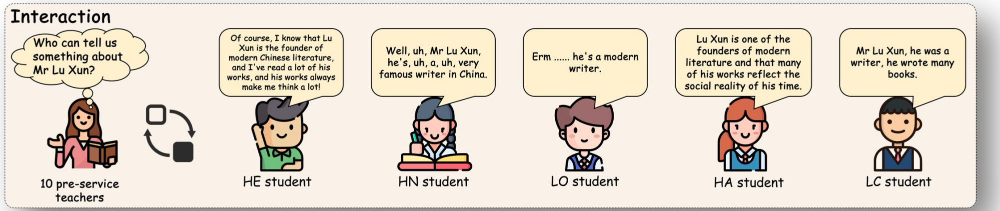
Evaluation Results with Different Learning Stage:

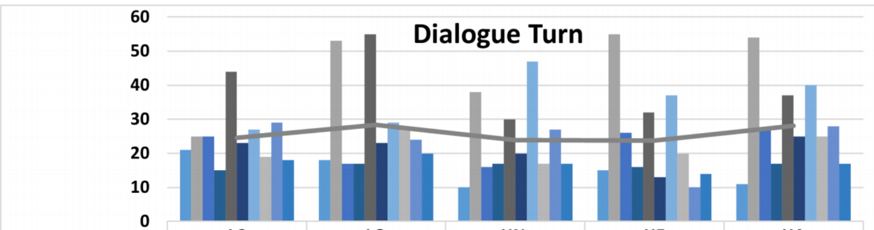
- The average performance of the four models increased by 36%, with paired t-test results showing strong statistical significance ($p < 0.001$).
- Fine-tuning based on learning stages is more effective than fine-tuning based on virtual students' personality traits.

Evaluation Results with Different Question Types:

- Paired t-tests showed statistically significant improvements across closed-ended, open-ended, and overall questions ($p < 0.05$).
- Performance differences reflect task complexity: closed-ended questions rely on factual recall, while open-ended questions require more complex reasoning and creativity.

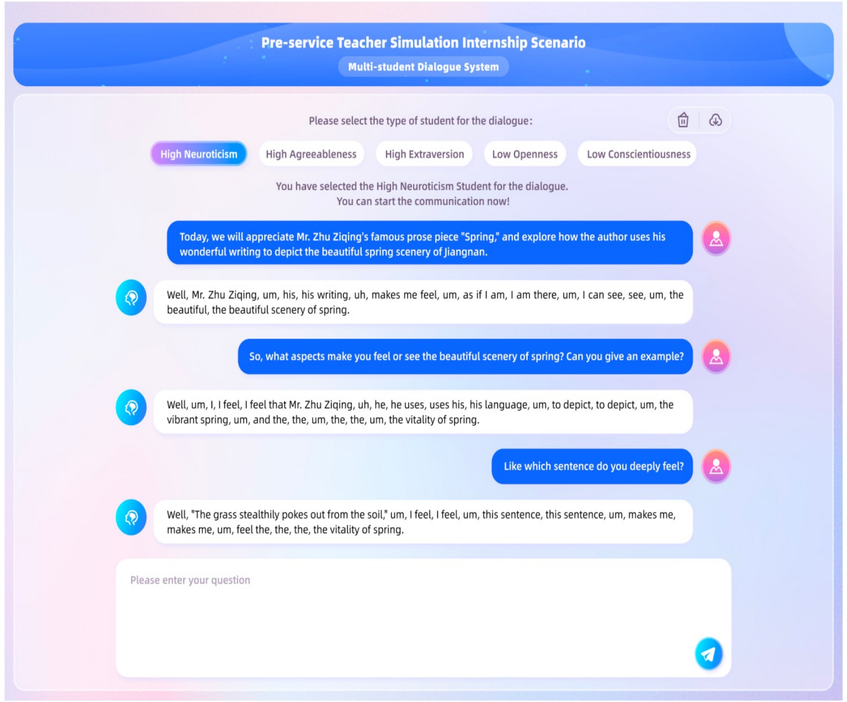
RQ4： How capable are virtual students in multi-turn interactions?





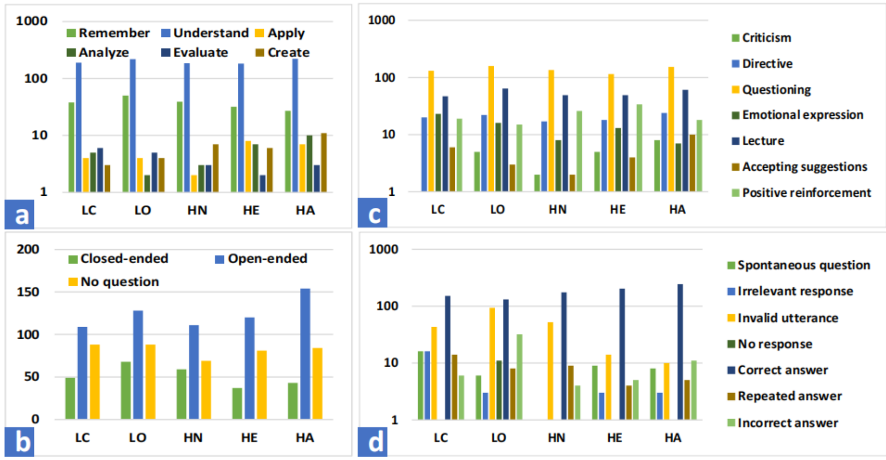
	LC	LO	HN	HE	HA
No1	21	18	10	15	11
No2	25	53	38	55	54
No3	25	17	16	26	27
No4	15	17	17	16	17
No5	44	55	30	32	37
No6	23	23	20	13	25
No7	27	29	47	37	40
No8	19	28	17	20	25
No9	29	24	27	10	28
No10	18	20	17	14	17
Average	24.6	28.4	23.9	23.8	28.1

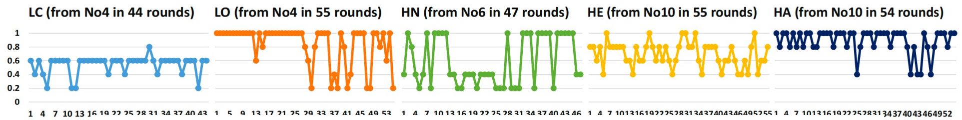
Figure A19: Dialogue turns per teacher (No1–No10) across five LVSA personality types.



Validated through 10 pre-service teachers × 1,288 real dialogue turns: virtual students maintained personality stability during long conversations

RQ4： How capable are virtual students in multi-turn interactions?





	No1	No2	No3	No4	No5	No6	No7	No8	No9	No10	Ave.
LC	0.48	0.66	0.61	0.54	0.61	0.62	0.45	0.58	0.56	0.63	**0.57**
LO	0.79	0.72	0.67	0.84	0.79	0.85	0.79	0.66	0.87	0.83	**0.78**
HN	0.66	0.91	0.81	0.86	0.62	0.60	0.87	0.84	0.72	0.84	**0.77**
HE	0.63	0.75	0.81	0.85	0.79	0.72	0.76	0.76	0.84	0.70	**0.76**
HA	0.87	0.97	0.94	0.97	0.93	0.93	0.93	0.91	0.98	0.89	**0.93**

Table 7: Average personality prediction scores per participant and personality type.

10 Pre-service Teachers × 1,288 Dialogue Turns

- High Extraversion students → More open-ended questions and higher-order cognitive guidance
- Low Openness students → More challenging, requiring more scaffolding
- **Teacher Adaptability**: 70% of teachers proactively adjusted teaching strategies to adapt to student personalities

Research Summary

Our Contributions

- We ***propose a structured framework*** for modeling and evaluating personality-aligned virtual student agents.
- We introduce an ***education-theory-driven framework*** to guide the construction of fine-tuning data.
- We incorporate ***human subjective*** evaluation criteria into GPT-4 prompt design.
- We conduct ***large-scale, multi-dimensional, and multi-level evaluations*** using GPT-4 to validate the intelligence of virtual student agents.

Future Work

- Extend the proposed paradigm to a wider range of ***academic subjects*** beyond the current scope.
- Utilize ***multi-agent simulations*** to reconstruct more realistic classroom settings, allowing systematic analysis of interaction dynamics among virtual students.

AAAI 2026
January 20 – 27, 2026
Singapore

Thanks for listening!

The authors gratefully acknowledge the support of the
MOE Tertiary Education Research Grant (Grant No. MOE2024-TRF-004).

2026.01.26 in Singapore

Yiping Ma

- PhD in East China Normal Univeristy (ECNU)
- Visiting Student in Nanyang Technological University (NTU)
- mayiping98@163.com; 52275901020@stu.ecnu.edu.cn

If you have any questions, please feel free to contact us!

Scan to see our series work

Scan to download this slide

WeChat for the first author