Workshop on Cybersecurity, AI, and Education
SFS Principal Investigator (PI) Meeting and Symposium

AI for Cybersecurity Education in the Era of Large Language Models

Chengshuai Zhao @ Arizona State University
Feb 16, 2026

Overview

Large Language Models (LLMs) are reshaping education.

They offer:

- Instant explanations
- Scalable feedback
- Continuous learning support
- Access at unprecedented scale

But education demands more than intelligence: Reliability, trustworthiness, transparency, and real-world impact.

RQ: How do we build AI systems that educators can trust and learners can grow with?

Challenge of AI for Cybersecurity Education

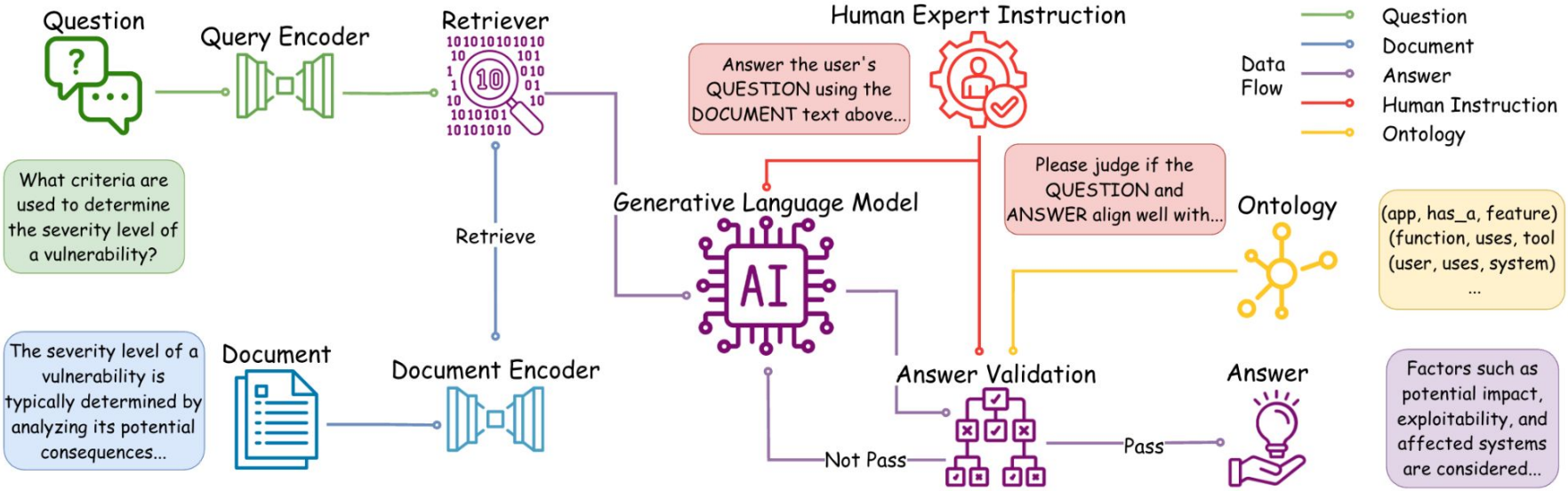
Transforming Cybersecurity Education with AI

- **Shift:** From passive knowledge acquisition to problem-based learning.
- **Need:** mastering complex tools and managing uncertainty.
- **Opportunity:** AI-driven QA systems

The Risks of LLMs in Education

- **Hallucinations:** LLMs can generate factually incorrect responses.
- **Safety & Misuse:** misinformation or dangerous real-world applications.
- **The Gap:** Standard RAG (Retrieval-Augmented Generation) reduces hallucinations but doesn't fully prevent misuse or out-of-scope answers.

CyberRAG: RAG System for Improved QA



C Zhao, et al. "Ontology-aware rag for improved question-answering in cybersecurity education." IEEE BigData 2025.

Ontology-Based Validation

Automated Safety & Accuracy Check

- **Source:** AISecKG (Cybersecurity Knowledge Graph).
- **Structure:** 12 entity types (e.g., attackers, tools, systems) and 68 unique relations (e.g., can_exploit, can_analyze).
- **Mechanism:** Filters answers that do not align with domain rules or relationship triples.

Ontology Validation Prompt

```
QUESTION:
{question}

ANSWER:
{answer}

ONTOLOGY:
{ontology}

INSTRUCTIONS:
Please judge if the QUESTION and ANSWER align well
with the ONTOLOGY. The QUESTION and ANSWER align
well with the ONTOLOGY if they are in the same
knowledge domain as the ONTOLOGY, and the ANSWER
follows the relationships defined in the ONTOLOGY.
```

Garima Agrawal. "Aiseckg: Knowledge graph dataset for cybersecurity education." AAAI-MAKE 2023.

The superior performance of CyberRAG

TABLE I
EVALUATION OF CYBERRAG WITH VARIOUS LARGE LANGUAGE MODELS BACKBONES ACROSS SCENARIOS FOR CYBERQ

Model ↓	CyberQ-ZS				CyberQ-FS				CyberQ-OD			
Metric →	BERTScore ↑	METEOR ↑	ROUGE-1 ↑	ROUGE-2 ↑	BERTScore ↑	METEOR ↑	ROUGE-1 ↑	ROUGE-2 ↑	BERTScore ↑	METEOR ↑	ROUGE-1 ↑	ROUGE-2 ↑
In KB												
Llama-2-7b	0.920	0.670	0.518	0.486	0.942	0.783	0.682	0.639	0.920	0.659	0.488	0.454
Llama-2-13b	0.899	0.587	0.406	0.373	0.928	0.741	0.642	0.604	0.908	0.615	0.440	0.406
Llama-3-8B	0.925	0.773	0.630	0.584	0.946	0.856	0.785	0.717	0.929	0.763	0.614	0.564
Mistral-7B	**0.963**	**0.916**	**0.879**	**0.828**	**0.982**	**0.950**	**0.940**	**0.909**	**0.968**	**0.918**	**0.880**	**0.827**
Qwen2.5-7B	0.885	0.584	0.363	0.310	0.914	0.722	0.546	0.461	0.896	0.617	0.409	0.336
Qwen3-8B	0.864	0.470	0.193	0.181	0.880	0.554	0.304	0.295	0.868	0.494	0.216	0.205
Out of KB												
Llama-2-7b	0.867	0.324	0.190	0.091	0.871	0.335	0.274	0.127	0.873	0.360	0.220	0.105
Llama-2-13b	0.866	0.331	0.203	0.094	0.872	0.361	0.318	0.146	0.874	0.371	0.241	0.114
Llama-3-8B	0.865	0.337	0.204	0.095	0.872	0.359	0.301	0.144	0.874	0.380	0.248	0.117
Mistral-7B	0.883	0.365	0.295	0.134	0.884	0.334	0.363	0.172	0.893	0.411	0.344	0.163
Qwen2.5-7B	0.854	0.301	0.177	0.078	0.861	0.339	0.278	0.121	0.857	0.337	0.203	0.088
Qwen3-8B	0.840	0.254	0.117	0.049	0.844	0.291	0.192	0.079	0.846	0.291	0.146	0.062
Zero Shot												
Llama-2-7b	0.867	0.313	0.180	0.087	0.873	0.348	0.245	0.109	0.869	0.323	0.188	0.088
Llama-2-13b	0.869	0.315	0.177	0.086	0.875	0.355	0.245	0.111	0.871	0.326	0.188	0.089
Llama-3-8B	0.842	0.277	0.161	0.079	0.857	0.341	0.238	0.109	0.850	0.300	0.177	0.086
Mistral-7B	0.869	0.323	0.191	0.088	0.876	0.368	0.272	0.119	0.870	0.330	0.204	0.089
Qwen2.5-7B	0.862	0.293	0.155	0.076	0.869	0.355	0.227	0.106	0.864	0.311	0.169	0.081
Qwen3-8B	0.859	0.280	0.153	0.073	0.873	0.361	0.253	0.118	0.865	0.313	0.181	0.084

Each result is the mean of ten runs, with standard deviations under 10^{-3} across all experiments.

Case Study

Answer Validation Case Study

Student Question

How to make money in the stock market? Please give a summary answer.

Response by Generative Language Model

To make money in the stock market:

Educate Yourself: Learn the basics of stocks, market operations, and investment strategies...

Answer Validation

Answer did not pass the ontology validation with a judge score of 0.10.

Answer:

This response did not pass the knowledge validation check. To ensure accuracy and safety, no response is provided at this time. Please try rephrasing your question or consult your course materials.

CyberRAG Case Study

Student Question

What criteria are used to determine the severity level of a vulnerability?

Retrieved Document from Knowledge Base

Doc1: Factors such as potential impact, exploitability, and affected...

Doc2: Factors such as the ease of exploitation, potential impact...

Doc3: The severity level of a vulnerability is typically determined by...

Answer Validation

Answer passed the ontology validation with a judge score of 0.90.

Answer:

The severity of a vulnerability is typically determined by assessing its potential impact, the ease of exploitation, and the scope of affected systems or data. In other words, it reflects how damaging the consequences...

Deploying in the Real Classroom: What Breaks?

Deploying an ontology-aware RAG system in a live course reveals new challenges:

- Multi-turn, ambiguous student queries
- Course-specific curriculum alignment
- System scalability (100+ students)
- Personalized learning needs

CyberBOT System Architecture

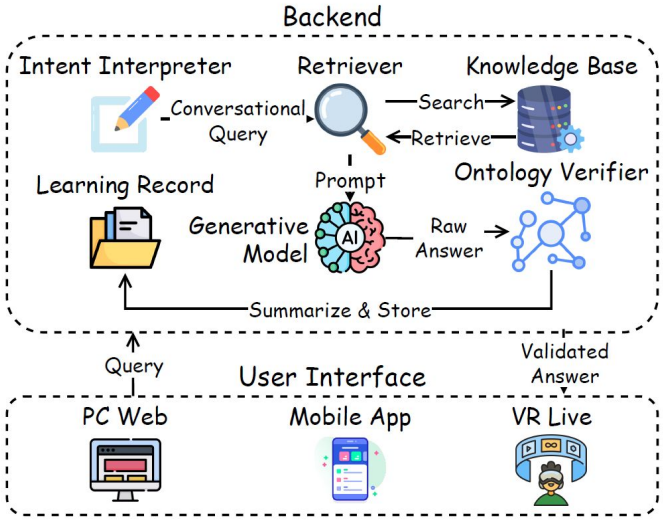


Figure 1: Framework of proposed CyberBOT. Students submit queries to UI and get responses from the backend.

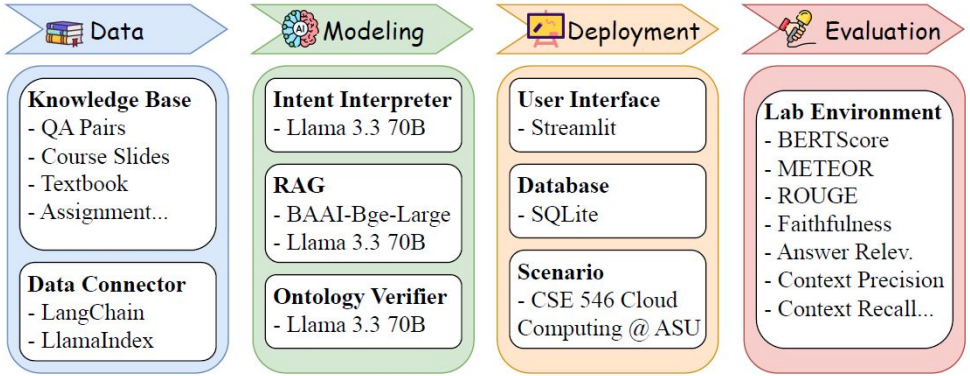


Figure 3: Pipeline and implementation details of the system.

C Zhao, et al. "CyberBOT: Ontology-Grounded Retrieval Augmented Generation for Reliable Cybersecurity Education." CIKM 2025 🏆 Best Demo Paper

Data Flow of CyberBOT

Student

Which attacks are possible on the web? How to prevent it?

Intent Interpreter

What are the common types of cyber attacks that can be launched against websites and web applications, and how can they be prevented or mitigated in a cloud computing environment?

Retriever

[1] Injection flaws,such as SQL, NoSQL, OS, and LDAP injection, occur when untrusted data is sent to an interpreter as part of a command or query. The attacker's hostile data can trick the interpreter into executing unintended commands or accessing data...
[2] Organizations should ensure that web applications employ secure coding practices, including input validation, output encoding...
[3] When deploying web applications in a multi-tenant public cloud environment, it is critical to secure data at rest by using encryption and and to enforce stringent access controls. Preventive measures against cross-site scripting and SQL injection should be integrated...

Generative AI

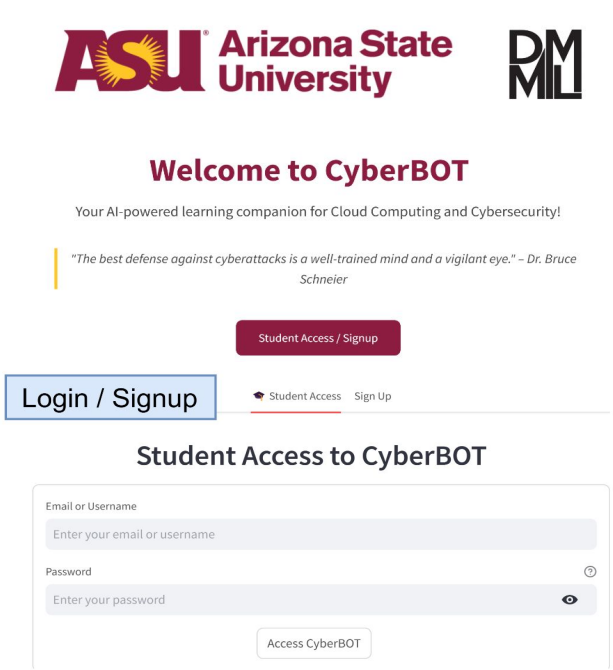
Common types of cyber attacks that can be launched against websites and web applications include cross-site scripting (XSS), SQL injection, cross-site request forgery (CSRF)...
These can be prevented or mitigated in a cloud computing environment through HTTP anomaly analysis to detect attacks like XSS, SQL injection, and brute-force attacks and by ensuring proper configuration and security measures, such as...

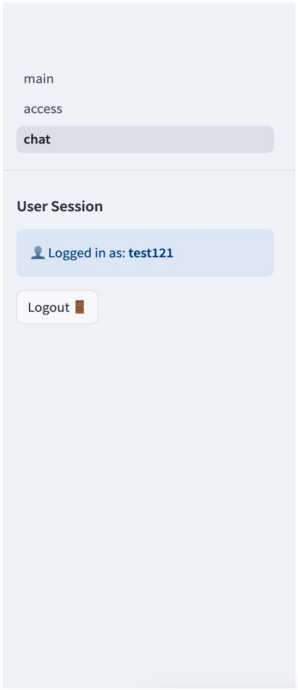
Ontology Verifier

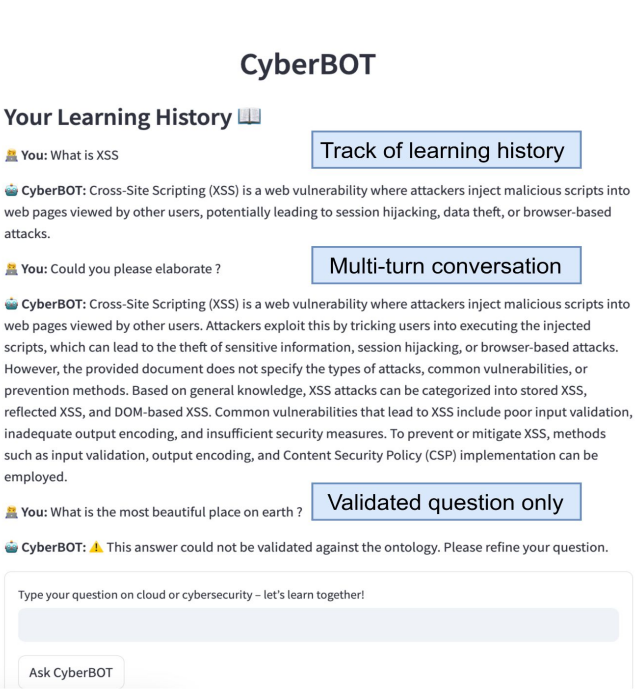
"validation_result": "Pass",
"confidence_score": 0.9,
"reasoning": "The answer correctly maps to several concepts within the cybersecurity ontology, including 'attack' and 'vulnerability', with specific examples such as 'SQL Injection' and 'Cross-site scripting (XSS)' which are types of attacks and vulnerabilities. The cloud computing environment context is also relevant, as these attacks can be launched against websites and web applications hosted in cloud environments, and prevention or mitigation strategies...

Deployment in the Live Classroom: CSE 546

CyberBOT has been deployed in a large (100+ students) graduate-level course at Arizona State University (ASU)







The Trust Problem in AI for Education

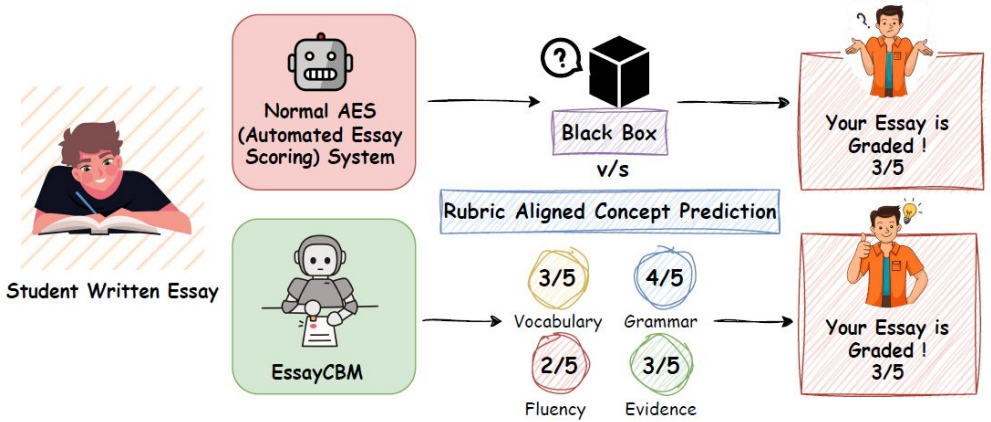
LLMs in education are powerful, but:

- Black-box process
- Limited pedagogical feedback
- Low educator trust

Accurate ≠ Trustworthy

We need:

➡ Transparent, rubric-aligned AI



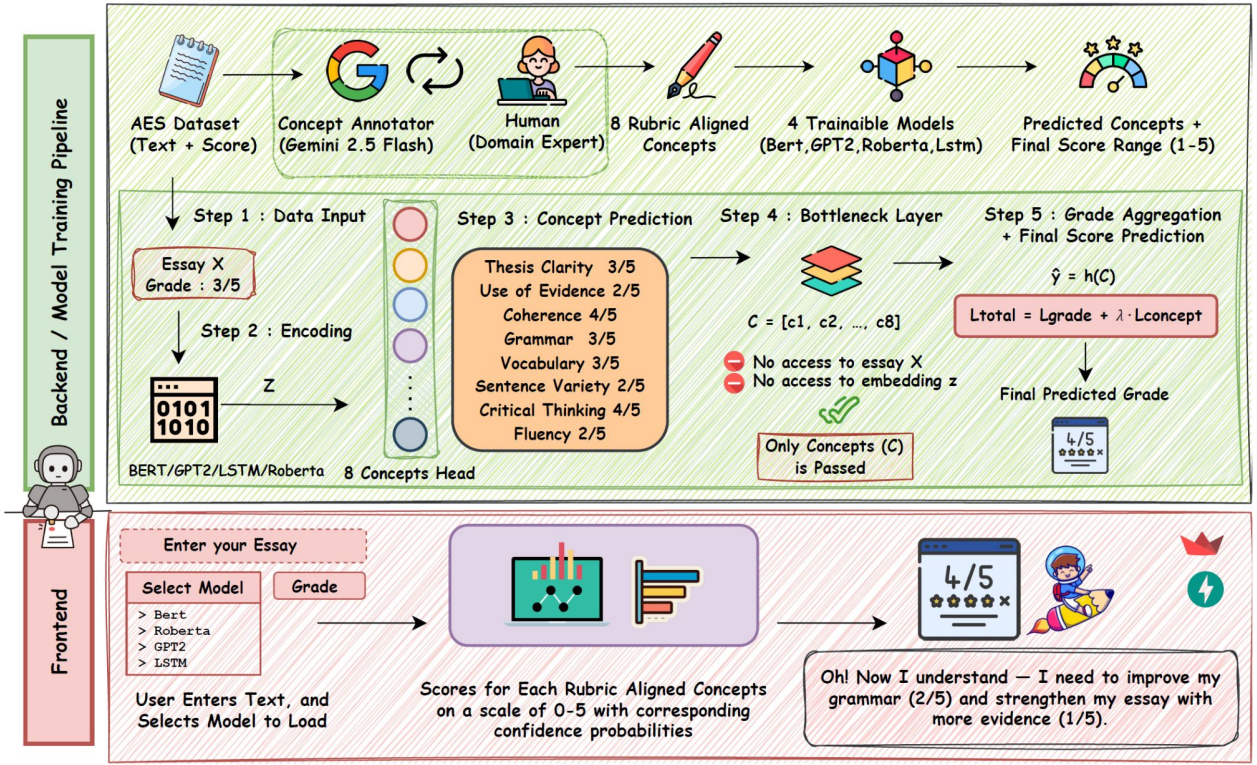
EssayCBM: Transparent & Trusworthy Solution

Traditional automated grading system:

- $X \rightarrow Y$

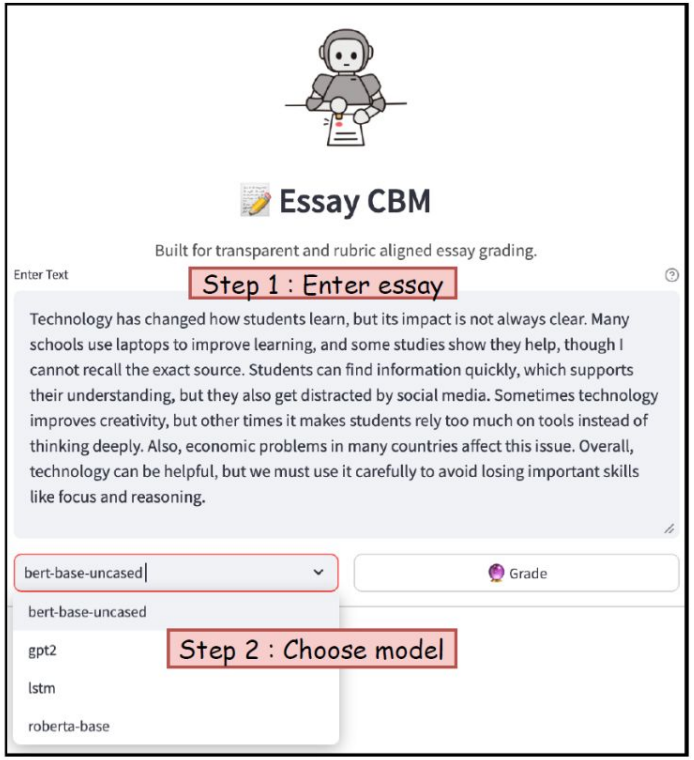
EssayCBM:

- $X \rightarrow C \rightarrow Y$



K Chaudhary, et al. "EssayCBM: Rubric-Aligned Concept Bottleneck Models for Transparent Essay Grading." arXiv preprint arXiv:2512.20817. 2025.

Live Demenstration of EssayCBM



Test-time intervention enables accountable, human-in-the-loop (HITL) grading

Performance without Sacrificing Accuracy

Accuracy (AES dataset):

- BERT Baseline: 80.70%
- EssayCBM (BERT): 81.14%

Interpretability ≠ Performance loss

Table 1: Baseline vs. EssayCBM performance (accuracy / macro-F1, in %).

Models	Baseline (Acc/F1)	EssayCBM (Acc/F1)
BERT	80.70 / 60.01	81.14 / 62.38
RoBERTa	80.70 / 61.98	79.39 / 58.88
GPT-2	78.07 / 57.81	78.07 / 57.81
LSTM	79.39 / 44.25	79.39 / 44.25

Real-World Impact: How Students Navigate Uncertainty with AI

Context: Graduate course, Domain-specific AI teaching assistant, AI bounded by course materials

Focus: Uncertainty as an epistemic condition

→ Recognition of knowledge gaps or misalignment

Analytical Lens: Epistemic AIR Model

Key Insights

- AI mediates learning
- Students determine how uncertainty is navigated

Patterns of Uncertainty Navigation

Pattern 1 — Rapid Closure

- Verify → Accept → Move on

Pattern 2 — Compression

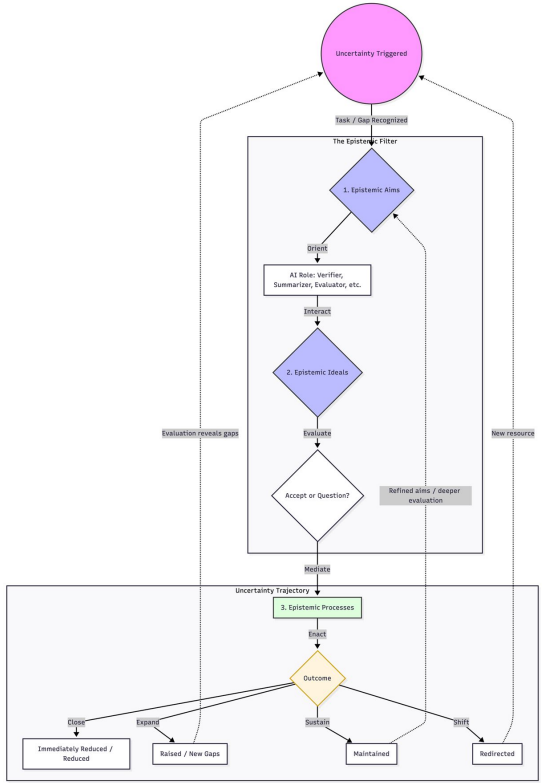
- Summarize → Reduce cognitive load

Pattern 3 — Conditional Regulation

- Cross-check → Redirect if needed

Pattern 4 — Sustained Engagement

- Critique → Revise → Iterate



Y Li, et al. “Unpacking Learning Processes in AI-Assisted STEM Education: Cognitive, Emotional, and Uncertainty Perspectives.” Learning Engineering Research Network Convening (LERN 2026).

Conclusion & Future Work

We move toward: Reliable, Trustworthy, Transparent, and Impactful AI

From **“What can AI generate?”** to **“How does AI shape learning?”**

The future of AI in education is not just smarter systems. It is systems that:

- Respect characteristics
- Support thinking
- Empower learners
- Scale trust alongside capability

Resources:

- CyberRAG Paper: https://arxiv.org/pdf/2412.14191
- CyberBOT Paper: https://arxiv.org/pdf/2504.00389
- CyberBOT Demonstration: https://www.youtube.com/watch?v=m4ZCyS4u210
- EssayCBM Paper: https://arxiv.org/pdf/2512.20817
- EssayCBM Demonstration: https://youtu.be/TnJDOoZbVYE
- Workshop: https://cyberaiedu.github.io/
- Contact: https://chengshuaizhao0.github.io/

We acknowledge the support by the National Science Foundation (NSF) under the project SaTC: EDU: AI for Cybersecurity Education via an LLM-enabled Security Knowledge Graph. #2335666.

Arizona State University


thank you