Journal of Contemporary Educational Research, 2025, Volume 9, Issue 12
http://ojs.bbwpublisher.com/index.php/JCER
ISSN Online: 2208-8474
ISSN Print: 2208-8466

Security Risks and Countermeasures for Large Language Models in Language Education: A Study Based on the DREAD Framework

Yafei Wang*

School of Humanities and Law, North China University of Technology, Beijing 100144, China

**Author to whom correspondence should be addressed.*

Abstract: This study systematically analyzes the security risks associated with the application of large language models (LLMs) in language education based on the DREAD threat assessment framework. It points out that, as "digital native speakers," LLMs are deeply integrated into the entire teaching process, introducing novel educational risks such as "language hallucinations," cultural bias, and prompt injection. These risks manifest specifically as high acquisition costs in knowledge internalization, high classroom reproducibility of risks, low exploitation thresholds, broad impact scope, and low visibility. In response, this study constructs a multi-layered dynamic governance system, proposing to reduce acquisition costs through a combination of technical filtering and manual verification, manage reproducibility and exploitation thresholds by implementing tiered access controls and full-process monitoring, and strengthen the digital literacy of both teachers and students to control the risk impact scope and enhance visibility. The research indicates that only by establishing a collaborative ecosystem led by educational principles, empowered by technology, supported by institutions, and founded on literacy can LLMs truly evolve into constructive tools that promote language proficiency development and cross-cultural understanding.

Keywords: Large language models; Language education; Security risks; Governance

Online publication: December 31, 2025

1. Background

Large language models (LLMs) represented by ChatGPT, DeepSeek, and Wenxin Yiyan are transcending their traditional roles as tools and evolving into "digital native speakers" with innate language capabilities. They can not only handle structured tasks but also demonstrate human-like contextual understanding, creative expression, and cultural adaptability, fundamentally reshaping the nature of human-computer interaction. This evolution is particularly pronounced in the field of language education, where LLMs are transitioning from auxiliary aids to intelligent partners capable of simulating authentic language acquisition processes. While driving revolutionary changes in teaching paradigms, they are also raising new ethical and security challenges. In the field of education, the application

of LLMs has gradually become a focus of academic attention. Research indicates that ChatGPT holds practical significance for language education. The English as a Second Language (ESL) academic community has primarily discussed ChatGPT's role and application in English teaching [1,2]. In foreign language classrooms, they have deeply integrated into the "input-interaction-output" chain: generating multimodal texts with one click, providing real-time error correction, simulating oral examiners, and writing academic model essays. However, behind the technological dividends, new risks such as "language hallucinations," "cultural bias," and "prompt injection" are being quietly amplified through classroom discourse. For instance, research shows that a teaching LLM developed by Stanford University failed to provide objective and impartial knowledge output when answering geopolitical questions about the Israeli-Palestinian conflict or the Russia-Ukraine war, even transmitting values prioritizing the United States [3]. In academic abstract writing tasks, it misidentified 32% of AI-generated abstracts as original ones [4]. As the application of LLMs continues to deepen, the foreign language discipline urgently needs to incorporate security risks into its conceptual system and propose countermeasures for risk prevention.

2. The DREAD framework for LLM risks

DREAD is an information security threat assessment model proposed by Microsoft Corporation, primarily used for the systematic analysis and quantitative evaluation of potential security risks. The name of this framework is derived from its five core evaluation dimensions: Damage, Reproducibility, Exploitability, Affected Users, and Discoverability. Each dimension measures the severity and impact scope of a threat from a different perspective, assisting security teams in prioritizing and making risk management decisions.

Among them, the Damage dimension assesses the extent of actual harm that may result if the threat is successfully executed, including impacts on systems, data, business operations, and organizational reputation. The Reproducibility dimension refers to the ease with which an attacker can repeatedly trigger the threat. The Exploitability dimension measures the technical difficulty, resource investment, and time cost required for an attacker to exploit the threat. The higher the exploitability, the greater the probability of the threat being successfully attacked. The Affected Users dimension quantifies the number of users or the scope of systems potentially impacted by the threat. This dimension focuses on the breadth of the security incident's impact, often measured by the proportion of affected users or the coverage of critical business functions. The Discoverability dimension evaluates how easily a vulnerability or threat can be discovered by an attacker. Due to its clear structure and strong practicality, the DREAD framework continues to be widely referenced and applied in the field of information security risk assessment. Aligning with the characteristics of language education, this study translates it into an "educational context," which is shown in **Table 1**.

Table 1. "Educational context" of DREAD

DREAD original dimension	Pedagogical dimension	Guiding questions for language education context
Damage	Acquisition cost	Does incorrect language or cultural knowledge lead to fossilization or identity misrecognition?
Reproducibility	Classroom reproducibility	Will the same prompt lead to repeated errors across different classes?
Exploitability	Exploitation threshold	Do teachers and students need advanced technical skills to trigger the risk?
Affected users	Affected learner scope	Does the risky discourse affect the whole class or specific language learner groups?
Discoverability	Risk visibility	Can teachers easily detect the risk based on their disciplinary expertise?

3. Analysis of language education risks under the DREAD framework

In language education, the five elements of the DREAD framework can be developed into acquisition cost, classroom reproducibility, exploitation threshold, affected scope, and risk visibility, respectively.

3.1. Acquisition cost (damage)

Acquisition cost refers to the persistent and classroom-amplified consequences of learners internalizing erroneous language rules, cultural cognitive biases, or identity conflicts due to incorrect, biased, or low-quality language output from LLMs.

When faced with complex tasks for which it cannot provide a solution, ChatGPT, in most cases, does not choose to refuse to answer but tends to generate text that deviates from prompt constraints, contradicts context, or does not align with world knowledge—a phenomenon known as "hallucination" [5]. In Chinese grammar teaching scenarios, most errors generated by ChatGPT are caused by "hallucination." The model may violate the prompt's requirements for a specific grammar point, generating irrelevant content that is formally similar or identical but fails to address the question. This phenomenon manifests across various language levels [6].

3.2. Classroom reproducibility (reproducibility)

Classroom reproducibility refers to the ease with which the same "problematic prompt" or "attack script" can be replicated with zero barriers across different classes and time periods, consistently triggering the same risk. Higher reproducibility leads to faster risk diffusion.

LLM-assisted content generation can provide a low-cost method for large-scale disinformation creation. For example, LLMs can lower the cost of disinformation campaigns by generating hundreds of text samples for students to choose from. Disinformation may be used to mislead the public, shape opinions on specific topics, create false "majority opinions," or flood websites with synthetic text [7].

3.3. Exploitation threshold (exploitability)

The exploitation threshold refers to the skills, tools, and time costs required for teachers, students, or external users to trigger an LLM risk event. Lower costs equate to a higher score and greater classroom security risk.

The threshold for the diffusion of LLM application risks is extremely low. Most risk events can be replicated with just a "browser + natural language," and the propagation chain is very short, allowing for clipboard-style diffusion between students, between classes, etc. Furthermore, the barrier to training large models is continuously decreasing. Attackers can frequently send a large number of queries and train a functionally similar "counterfeit" model based on the outputs [8].

3.4. Affected scope (affected users)

Affected scope refers to the number of students, teachers, classes, or schools ultimately impacted by a single-point risk event. The broader the scope, the greater the difficulty of classroom governance.

Researchers at ETH Zurich discovered that the large language models powering chatbots like ChatGPT can accurately infer a surprising amount of personal information about users—including their race, location, occupation, etc.—from seemingly harmless conversations. Anyone using such a model can be affected [9].

3.5. Risk visibility (discoverability)

Risk visibility refers to the ease with which teachers or administrators can detect errors, biases, or backdoors in LLM outputs based on their disciplinary expertise or routine audits. Lower visibility makes it easier for risks to

remain latent long-term.

This depends on the teacher's disciplinary literacy and critical thinking. AI's "judgments" rely on training data. If the data itself contains biases—such as over-representing reading cases from urban students—the generated content may favor a particular group. For example, a teacher using AI to analyze "excellent case studies of picture-based writing for lower grades" found that AI-recommended cases were mostly scenarios like "city park" or "birthday party," while materials from rural students, like "catching butterflies in the field" or "helping grandma feed chickens," were not included. Further investigation revealed that in the AI's training data, picture-based writing from urban students accounted for 75%, while rural students only accounted for 25%. If teachers directly use AI-recommended cases for instructional design, it may leave rural students with "nothing to say" [10].

4. Security risk governance countermeasures for LLMs based on the DREAD framework

The application of LLMs in language education harbors multiple risks. Their governance must adhere to the principles of systematization and dynamism, deeply integrating technical defenses with educational practice. Based on the risk analysis using the DREAD framework, this section constructs a multi-layered, coordinated, and interactive security defense system from the five dimensions of "Acquisition Cost, Classroom Reproducibility, Exploitation Threshold, Affected Scope, and Risk Visibility." This system encompasses four levels: technological optimization, classroom practice, capacity building for teachers and students, and institutional safeguards. It aims to shift security risk management from "passive response" to "active defense."

4.1. Defense strategies for high acquisition cost: Building a dual "technical filtering + manual verification" error correction mechanism

High acquisition cost stems from learners internalizing language errors or cultural biases from LLM outputs. The core of defense lies in "early detection, timely correction," blocking the path to acquiring erroneous knowledge.

4.1.1. Technical level

Firstly, to develop domain-specific "Educational Guardian Models": encourage EdTech companies and universities to collaborate on fine-tuning general-purpose LLMs or training specialized models based on high-quality, multi-source, debiased language education corpora (e.g., approved textbooks, academic papers, authentic cross-cultural communication materials). These models should embed "fact-checking" and "cultural sensitivity" modules capable of automatically annotating uncertain information when generating responses and providing multi-perspective prompts for statements involving values or historical events.

Secondly, to deploy real-time output monitoring and alert plugins: develop browser extensions or API interface plugins for teaching LLM platforms, embedding error detection algorithms based on rules and statistics. For instance, in Chinese teaching scenarios, integrate grammar error recognition models similar to those studied by Bin *et al.*. When sentences suspected of being hallucinated—violating target grammatical structures or semantic logic—are detected, alerts can pop up on the teacher's interface, or students can receive prompts like "It is recommended to discuss this content with the teacher."

4.1.2. Educational practice level

Firstly, to mandate the "AI Output Manual Verification" teaching process: make "critically examining AI-generated content" an essential component of language teaching. Teachers should design "AI Text Evaluation

Forms," guiding students to conduct group discussions and annotations on LLM outputs from four dimensions: grammatical accuracy, cultural appropriateness, logical consistency, and evidential reliability. This fosters a thinking habit of "accept-question-verify."

Secondly, to establish a dynamically updated "Common AI Error Case Database": subject groups should regularly collect and categorize typical LLM error outputs encountered in teaching (e.g., hallucination cases for specific grammar points, biased expressions on specific cultural topics), forming a school-based or regionally shared case database. This database serves not only for teacher training but also as a reference for students during "human-AI collaboration" tasks, enhancing risk identification capabilities.

4.2. Management strategies to reduce classroom reproducibility and exploitation threshold: Implementing tiered access and full teaching process monitoring

High reproducibility and low-threshold exploitation accelerate risk diffusion. The focus of countermeasures is to increase the cost of triggering risks and limit their spread through technical constraints and procedural controls.

Firstly, to implement a "Tiered LLM Access Permission" system based on teaching scenarios: schools or regional education platforms should provide teachers and students with vetted and configured educational LLM interfaces, rather than fully open commercial general models. Set different permission levels according to teaching tasks: e.g., basic writing exercises only enable grammar checking and basic rewriting functions; advanced academic writing may enable literature review assistance but automatically block the generation capability for sensitive political or historical topics. All teaching prompts and generated results should be logged in the backend for auditing.

Secondly, to construct a full teaching process risk monitoring and traceability system: utilize Learning Management Systems (LMS) to integrate LLM usage logs, recording students' submitted prompts, models used, and key snippets of generated content. When the system detects highly similar risk prompts (e.g., "attack scripts" inducing the generation of specific biased content) from multiple users, it can automatically issue warnings and trace the source, enabling early risk detection and intervention. Simultaneously, employ technologies like digital watermarking to label AI-generated lesson plans and student assignments, reducing the risk of indiscriminate dissemination of "counterfeit models" or synthetic disinformation inside and outside the classroom.

4.3. Empowerment strategies to control affected scope and enhance risk visibility: Strengthening teacher agency and digital literacy education

The expansion of the affected scope often accompanies low-risk visibility. The fundamental countermeasure lies in enhancing the agency of "people" in educational activities, particularly teachers' professional judgment and students' critical thinking.

4.3.1. Teacher capacity building level

Firstly, to conduct specialized training on "AI Literacy" and "Disciplinary Risk Identification": teacher training should not stop at the level of tool usage but must deepen to risk cognition. Training content should include: the application of risk analysis frameworks like DREAD, common AI hallucination and bias patterns in the specific discipline (e.g., grammatical hallucinations, cultural stereotypes in foreign language teaching), and rapid verification methods based on disciplinary knowledge (e.g., using authoritative dictionaries or corpora to verify language points generated by AI). Encourage teachers to become "prompt engineers" guiding LLMs to produce more reliable content through refined, structured prompts.

Secondly, to establish interdisciplinary "AI-in-Education Application Ethics Review Panels": at the school

or departmental level, form review panels comprising language teachers, educational technology experts, and ethicists. Conduct pre-implementation reviews of new AI tools intended for classroom use and widely used AI-generated teaching materials (e.g., reading texts, test questions), assessing their potential disciplinary knowledge risks, cultural biases, and data privacy issues to control the risk impact scope from the source.

4.3.2. Student literacy cultivation level

Firstly, to incorporate "Human-AI Collaborative Critical Thinking" into curriculum objectives: explicitly add the competency indicator of "using AI tools responsibly" to language skill courses. Through project-based learning, let students compare responses from different AI models to the same question in practice, analyzing their differences and potential flaws; practice writing prompts that elicit more precise and fair responses; understand how limitations in AI training data lead to output biases, thereby becoming prudent AI users rather than passive recipients.

Secondly, to promote "Data Transparency and Privacy Protection" education: educate students to recognize that every interaction with an LLM may leak personal or group information. Guide students in setting privacy options (e.g., disabling chat history for training), advise against disclosing sensitive personal information to AI, and inform them about relevant school policies on protecting educational data, thereby narrowing the affected scope of privacy leakage risks from the user end.

4.4. Building long-term mechanisms: Promoting standard development and collaborative governance

The security governance of LLMs in language education is not the responsibility of a single school or a teacher. It requires the joint efforts of the industry, academia, and regulatory bodies to establish long-term mechanisms.

Firstly, to develop "Safety Guidelines for the Application of Large Language Models in Language Education": formulate industry guidelines, led by education authorities, in collaboration with experts in linguistics, educational technology, and information security; clearly define red lines and best practices for AI use in teaching different language skills across various educational stages, providing a basis for school management.

Secondly, to establish a "Risk Case Sharing and Early Warning Platform": promote the establishment of regional or national educational AI security information-sharing mechanisms. Schools can anonymously upload newly discovered risk cases (e.g., new prompt injection attack methods, variants of specific cultural bias expressions) to the platform, enabling rapid circulation of risk intelligence and collective defense.

Thirdly, to strengthen Industry-Academia-Research Collaboration to Tackle Education-Specific Security Technologies: set up special research projects to support collaboration between universities and enterprises in developing "Explainable AI" technologies more suited to educational contexts, making model decision-making processes more transparent to teachers; develop efficient and low-cost "adversarial testing" tools to help schools proactively discover model vulnerabilities before deployment.

5. Conclusion

The integration of LLMs into language education has become an irreversible trend. The associated security risks represent a concentrated manifestation of the "double-edged sword" nature of this technology. The systematic analysis based on the DREAD framework in this study shows that these risks permeate the entire process from knowledge generation to student internalization, characterized by concealment, diffusivity, and severe educational consequences. The response must not stop at simple technological blocking or moral appeals.

Instead, it is imperative to construct a multidimensional governance ecosystem centered on "Educational Leadership, Technological Empowerment, Institutional Safeguarding, and Literacy Foundation." By pre-positioning security defenses, internalizing risk awareness, and sharing governance responsibilities, we can harness this transformative technology to truly become a beneficial tool for promoting language proficiency development and cross-cultural understanding, rather than a new source of risk. Future research could further focus on the differential manifestation of risks in different languages (e.g., less commonly taught languages) and teaching modes (e.g., online, blended), as well as the quantitative evaluation of the actual effectiveness of various countermeasures, thereby continuously improving this dynamically evolving governance system.

Funding

The research was funded by Beijing Philosophy and Social Science Project: A Study on the Mechanism of Artificial Intelligence Enhancing New-Quality Productivity of Higher Education in Beijing (24JYB013).

Disclosure statement

The author declares no conflict of interest.

References

[1] Pack AJ, Austin JL, Maloney AJ, 2023, Using Generative Artificial Intelligence for Language Education Research: Insights from Using OpenAI's ChatGPT. TESOL Quarterly, 57(4): 1571–1582.

[2] Xu JJ, Zhao C, 2024, The Role of Large Language Models in English Teaching. Foreign Language Education in China, 7(1): 3–10 + 90.

[3] Jiang GY, Yin WX, 2025, Transformation or Crisis: The Empowerment and Limits of Large Language Models in University Teaching—A Case Study Based on Stanford University. e-Education Research, 46(1): 122–128.

[4] Else H, 2023, Abstracts Written by ChatGPT Fool Scientists, viewed November 26, 2025, https://www.nature.com/articles/d41586-023-00056-7

[5] Huang L, Yu W, Ma W, et al., 2025, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 43(2): 1–55.

[6] Shuai B, Wang YM, Hu RF, 2025, Evaluation of Chinese Grammar Teaching Resource Generation Based on Large Language Models. Chinese Teaching in the World, 39(3): 405–419.

[7] Ge MY, Wu YQ, Mei YY, 2024, Security and Governance of Large Language Models in Universities. China Education Network, 2024(6): 39–42.

[8] Zhang JX, Liu S, Li K, 2025, Large Models Frequently Attacked, Security Governance Urgently Needed, Science and Technology Daily, November 20, 2025, (005).

[9] Fang X, 2023, Study: Large Models Can Highly Accurately Infer User Privacy, May Be Used by Ad Companies and Scammers, The Paper, October 29, 2023.

[10] Liu BC, Gou MH, 2023, The Impact and Countermeasures of ChatGPT and Other New-Generation AI Tools on Educational Research. Journal of Soochow University (Educational Science Edition), 11(3): 54–62.