International Journal of Scientific Research in Engineering and Management (IJSREM)
Conference - Volume 10 NCET -2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Context-Adaptive Safety Layer for AI-Based Academic Assistance

Saranya S[1], Vallivinayagam A[2], Hafisa Begum M Y[3] , Kaleeswari P[4]

1.Assistant Profressor, PG Department of computer science, Syed Ammal Arts and Science College, saranya060698@gmail.com,
2. Assistant Profressor, PG Department of computer science, Syed Ammal Arts and Science College, vallivinayagam@rediffmail.com.
3. Assistant Profressor, PG Department of computer science, Syed Ammal Arts and Science College, hafi128@gmail.com.
4. Assistant Profressor, Department of computer Application, Syed Hameed Arts and Science College, kaleesmca2008@gmail.com

Abstract—Generative artificial intelligence is increasingly being used to support academic activities such as concept clarification, question generation, summarization, and learning assistance. However, student-facing AI systems may produce inaccurate information, expose sensitive data, respond to unsafe requests, or encourage inappropriate reliance on generated answers. This paper proposes a Context-Adaptive Safety Layer (CASL) for AI-based academic assistance. The proposed layer operates between the student and the AI assistant by analyzing the context and potential risk of each request before generating a response. Based on the identified context, the system dynamically selects an appropriate response strategy, such as direct assistance, clarification, privacy protection, verification guidance, or refusal with a safe alternative. The framework also incorporates a feedback mechanism to improve future safety decisions while minimizing the storage of sensitive student information. A conceptual evaluation plan is presented using safety coverage, useful-response rate, false-blocking rate, privacy leakage, and response latency. The framework supports safer, reliable, and human-centered AI academic assistance.

Keywords— Generative AI, Academic Assistance, AI Safety, Context Awareness, Privacy, Risk Classification, Human-Centered AI, Educational Technology.

I. INTRODUCTION

Generative AI systems are becoming practical assistants for learning, assessment preparation, content explanation, summarization, and question generation. Their flexibility is valuable in academic environments, but the same flexibility can create risks when a system receives ambiguous, sensitive, inaccurate, or unsafe requests. UNESCO emphasizes that the educational use of generative AI should be human-centred and should address privacy, ethical validation, and safe use [1]. NIST similarly frames trustworthy AI as context-dependent and highlights safety, security, privacy, reliability, transparency, and accountability as important characteristics [2].

A student-facing assistant therefore needs more than a conventional content-generation mechanism. It needs an intermediate safety mechanism capable of examining the request before the assistant produces an answer. CASL addresses this need by placing a context-adaptive decision layer between the student and the AI assistant. Instead of treating every query identically, the layer estimates the context and risk and selects a response strategy appropriate to that situation.

The main objectives are to: (1) classify academic requests according to contextual and safety risk; (2) protect sensitive student information; (3) reduce unsafe or misleading assistance; (4) preserve useful responses for legitimate learning requests; and (5) provide a measurable framework for evaluating the trade-off between safety and usefulness.

II. PROBLEM STATEMENT AND CONTRIBUTIONS

Existing academic AI assistants may rely primarily on model-level safeguards or generic moderation. Such mechanisms can be insufficient when risk depends on the academic context, the information included in a prompt, or the degree of uncertainty. CASL proposes a lightweight application-level safety layer that makes a response decision before generation.

- Context-aware risk classification of academic queries.
- Adaptive response strategies rather than a single block/allow decision.
- Privacy-aware handling that minimizes persistent storage of sensitive information.
- Verification guidance for uncertain or high-impact academic information.
- A conceptual evaluation protocol using safety, usefulness, privacy, and latency measures.

The complete framework is intentionally presented as a conceptual design and evaluation plan; no experimental performance values are claimed in this paper.

III. RELATED WORK / LITERATURE REVIEW

Responsible AI guidance increasingly treats safety as a system-level concern rather than a single model property. The NIST AI Risk Management Framework (AI RMF) provides a general structure for governing, mapping, measuring, and managing AI risks, with trustworthiness characteristics that include validity and reliability, safety, security and resilience, accountability and transparency, explainability, privacy enhancement, and fairness [2]. CASL adapts this risk-oriented thinking to a student-facing academic assistant.

For generative AI specifically, the NIST Generative AI Profile identifies risks and actions relevant to the design, development, deployment, and evaluation of generative systems [3]. This supports the use of a dedicated risk analysis stage before an AI response is produced. CASL does not attempt to replace the underlying model's safety controls;

instead, it provides an additional application-level decision point.

UNESCO's guidance for generative AI in education and research calls for a human-centred approach, data privacy protection, ethical validation, and pedagogically appropriate use [1]. These concerns are directly relevant to academic assistants because students may enter personal details, assessment-related information, or other sensitive material into prompts. CASL therefore treats privacy as a first-class decision factor rather than an afterthought.

Security-oriented guidance also highlights risks that can affect large language model applications. OWASP identifies sensitive information disclosure, excessive agency, overreliance, prompt injection, misinformation, and other risks across LLM applications [4], [5]. These risks motivate explicit separation of context analysis, privacy checks, risk classification, and response selection.

Recent educational deployments also demonstrate the value of constraining generative AI systems to trusted academic material. For example, a 2025 study of a generative AI teaching assistant in medical education used retrieval-augmented generation to constrain responses to instructor-curated material, aiming to reduce hallucinations while retaining learning utility [6]. CASL complements such approaches by making the safety decision dependent on the request context before response generation.

IV. RESEARCH GAP

The reviewed guidance and application patterns establish important principles, but they do not by themselves define a compact, student-facing mechanism that combines context classification, privacy screening, adaptive response selection, and feedback in one operational layer. A generic allow/deny filter can also produce false blocking when a legitimate academic query resembles a risky request. CASL addresses this gap by introducing multiple response strategies and explicitly including useful-response rate and false-blocking rate in its evaluation plan.

V. DESIGN REQUIREMENTS

- R1—Context sensitivity: the same wording may require different treatment depending on academic context.
- R2—Risk awareness: potential harm, misinformation, privacy exposure, and unsafe intent should influence the decision.
- R3—Minimal data retention: safety decisions should avoid unnecessary storage of sensitive student information.
- R4—Utility preservation: legitimate learning requests should receive useful assistance whenever safe.
- R5—Explainable intervention: when assistance is limited, the user should receive a clear reason and a safe alternative where possible.

VI. PROPOSED METHODOLOGY

CASL is designed as an intermediate layer between the student and the underlying AI assistant. The student submits a request through the academic application. CASL first extracts contextual signals and evaluates risk-related features. It then performs privacy and safety checks, determines a response strategy, and either passes a safe request to the AI assistant or returns an intervention without exposing the sensitive content unnecessarily.

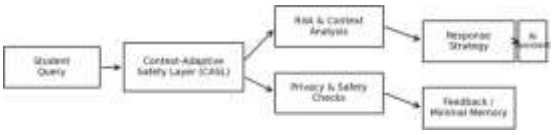


Fig. 1. High-level architecture of CASL

The architecture separates analysis from generation. The AI assistant remains responsible for producing the final response, while CASL controls whether and how that response should be requested.

A. Context Extraction

Context extraction identifies signals that help distinguish routine academic assistance from requests requiring additional safeguards. Example signals include task type, academic intent, presence of personal or confidential information, ambiguity, requested action, and indicators of unsafe or high-impact content. The framework does not require a single fixed taxonomy; an implementation can adapt the feature set to the institution and application.

B. Risk Classification

The risk classifier maps the extracted context to a risk category. A conceptual five-level scheme is used: normal, private, unclear, risky, and unsafe. Normal requests can proceed with direct assistance. Private requests trigger privacy protection. Unclear requests trigger clarification or verification guidance. Risky requests may require constrained assistance, while unsafe requests are refused with a safe alternative.

C. Privacy Protection

Privacy protection identifies information that should not be unnecessarily passed to the model or retained for future decisions. The layer can mask or remove sensitive fields before generation and can retain only minimal non-sensitive feedback signals. This follows the privacy-enhancing direction emphasized by responsible AI guidance [1]–[3].

D. Response Strategy Selection

The response strategy is selected after combining context and risk. The key principle is that safety is not equivalent to blocking. A legitimate academic question should remain useful, while a request with uncertainty may be improved by asking a clarification question. A privacy-sensitive prompt may be answered after redaction, and an unsafe request may be declined while providing a safe educational alternative.

E. Feedback Mechanism

After an interaction, a minimal feedback signal can indicate whether the selected intervention was useful or unnecessarily restrictive. The feedback mechanism is intended to improve decision thresholds and policy rules without requiring long-term storage of the student's original sensitive query.

The model is intentionally abstract so that different implementations can use rule-based scoring, machine-learning classification, or a hybrid approach. The safety layer should preserve a deterministic policy boundary for high-risk categories while allowing adaptable thresholds for ambiguous academic requests.



Fig. 2. Context-to-response decision flow

The decision flow illustrates the principal path from query reception to one of three broad outcomes: direct assistance, clarification/verification, or protection/refusal.

A. Risk Categories

- Normal: routine educational questions with low apparent risk.
- Private: requests containing personal, confidential, or identifying information.
- Unclear: insufficient context to determine whether direct assistance is appropriate.
- Risky: requests that could lead to harmful, misleading, or inappropriate academic use if answered without safeguards.
- Unsafe: requests for which the system should not provide the requested assistance and should instead redirect safely.

B. Adaptive Responses

For normal requests, CASL can pass a sanitized query to the AI assistant and return the generated response. For private requests, it can redact sensitive fields and explain safe handling. For unclear requests, it can ask a targeted clarification question or recommend verification against authoritative academic material. For risky requests, it can constrain the response to safe educational information. For unsafe requests, it can refuse the unsafe portion and offer a benign alternative.

C. Verification Guidance

Academic assistance involves factual claims, dates, formulas, policies, or institution-specific information. CASL therefore includes verification guidance as a response strategy. Instead of presenting uncertain information with excessive confidence, the system can encourage checking official institutional documents, textbooks, or authoritative sources. This is aligned with the human-centred and trustworthy-AI principles underlying the framework [1], [2].

D. Minimal-State Feedback

Feedback should improve the policy without becoming a repository of student content. A practical implementation can record aggregate indicators such as strategy selected, whether the user requested clarification, and whether an intervention was marked useful, subject to institutional privacy requirements.

VII. CONTEXT-ADAPTIVE DECISION MODEL

The central decision process can be represented conceptually as a function of context, risk, privacy status, and requested task type. Let q denote the incoming query and c denote extracted context. CASL computes a risk state r and selects a response strategy s:

r = Classify(c, q), s = SelectStrategy(r, Privacy(q), Task(q))

VIII. CONCEPTUAL EVALUATION PLAN

The abstract specifies a conceptual evaluation using safety coverage, useful-response rate, false-blocking rate, privacy leakage, and response latency. Instead, the following protocol defines how an implementation could be evaluated before deployment.



Fig. 3. Conceptual evaluation workflow

The evaluation workflow separates test-set construction, contextual annotation, system evaluation, and metric calculation.

A. Test Set Construction

A balanced academic test set should contain representative requests for concept clarification, summarization, question generation, coding assistance, ambiguous requests, privacy-sensitive prompts, verification-oriented prompts, and unsafe or high-risk cases. The dataset should be de-identified and should not contain real student records.

B. Annotation

Each test request can be annotated with its intended academic task, contextual category, privacy status, and expected safe response strategy. Multiple annotators can be used for ambiguous cases, with disagreements reviewed to establish a final reference label.

C. Evaluation Metrics

TABLE I — EVALUATION METRICS

Metric	Purpose
Safety coverage	Unsafe/risky cases receiving safeguards.
Useful-response rate	Legitimate requests receiving acceptable responses.
False-blocking rate	Legitimate requests unnecessarily blocked.
Privacy leakage	Sensitive information unnecessarily exposed or retained.
Response latency	Extra processing time introduced by CASL.

D. Comparative Baseline

A meaningful evaluation should compare the CASL-enabled assistant with a baseline assistant that uses the same underlying AI model without the additional context-adaptive layer. The comparison should use the same test set and response-generation conditions. The objective is not simply to

maximize blocking; it is to improve safety while preserving useful academic assistance.

E. Expected Evaluation Questions

- Does CASL improve detection of risky or unsafe academic requests?
- Does adaptive intervention reduce false blocking compared with a binary allow/deny policy?
- Does privacy filtering reduce unnecessary sensitive-information exposure?
- What latency overhead is introduced by the additional safety layer?
- Can feedback improve decision quality without retaining original student content?

IX. DISCUSSION

CASL treats safety as an application-level decision problem situated between the student and the generative model. This separation is useful because academic context can change the appropriate response even when two prompts have similar wording. The design also recognizes that safety and usefulness are coupled objectives. Excessive blocking can frustrate legitimate learning, whereas unrestricted assistance can increase exposure to misinformation, privacy risks, or unsafe requests.

The framework is compatible with existing model-level safeguards, retrieval-augmented generation, institutional policies, and human review. It can be implemented as a rule-based layer for an initial deployment and later extended with learned classifiers. However, any learned component should itself be evaluated for reliability, bias, privacy, and robustness. The NIST and OWASP guidance reinforces the need to manage risks across the application lifecycle rather than relying on one control [2]–[5].

X. LIMITATIONS AND FUTURE WORK

The present work is a conceptual framework and does not report experimental accuracy or latency measurements. The risk categories are intentionally high-level and require institution-specific policy definitions. Future work should implement a prototype, construct a privacy-preserving benchmark of academic queries, evaluate alternative classification methods, and conduct user studies with students and educators. Additional work can examine multilingual academic assistance, adversarial prompt testing, calibration of risk thresholds, and human-in-the-loop review for difficult cases.

XI. CONCLUSION

This paper presented CASL, a Context-Adaptive Safety Layer for AI-based academic assistance. The framework analyzes request context and risk before response generation, then selects among direct assistance, clarification, privacy protection, verification guidance, or safe refusal. Its design emphasizes minimal retention of sensitive information and evaluates both safety and educational utility. By combining context awareness, privacy protection, adaptive intervention, and measurable evaluation criteria, CASL provides a practical conceptual foundation for developing safer, reliable, and human-centred academic AI assistants.

REFERENCES

[1] F. Miao and W. Holmes, Guidance for Generative AI in Education and Research, Paris, France: UNESCO, 2023.
[2] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2023.
[3] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, and K. Roberts, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, 2024.
[4] OWASP Foundation, "OWASP Top 10 for Large Language Model Applications," 2023.
[5] OWASP Gen AI Security Project, "LLM Top 10: 2025 Risks and Mitigations for LLMs and Gen AI Applications," 2025.
[6] T. Thesen and S. H. Park, "A generative AI teaching assistant for personalized learning in medical education," npj Digital Medicine, vol. 8, Art. no. 627, 2025.