2026 IEEE 50th Annual Computers, Software, and Applications Conference (COMPSAC) | 979-8-3315-4497-3/26/$31.00 ©2026 IEEE | DOI: 10.1109/COMPSAC69091.2026.00199

Designing Psychologically Safe AI Tutors for Students: An Emotion-Aware Post-hoc Intervention for LLM-Assisted Learning

Yicheng Sun†, Jacky Keung†, Hi Kuen Yu†, Yihan Liao†, Zhenyu Mao†, Yishu Li‡, Xiaoxue Ma‡,*
†Department of Computer Science, City University of Hong Kong, Hong Kong, China
‡Department of Electronic Engineering and Computer Science, Hong Kong Metropolitan University, Hong Kong, China
{yicsun2-c, hikuenyu2-c, yihanliao4-c, zhenyumao2-c}@my.cityu.edu.hk, jacky.keung@cityu.edu.hk, sliy@hkmu.edu.hk
*Corresponding author: kxma@hkmu.edu.hk

***Abstract*—Large language models (LLMs) are increasingly integrated into students' learning processes, serving not only as cognitive tools for academic support but also as conversational partners that influence learners' emotional experience. However, existing LLM-based learning systems are not explicitly designed to account for learners' emotional vulnerability, and institutions often deploy such models as black-box services, limiting the feasibility of model-level modification. In this work, we investigate the prevalence of emotionally risky responses in real-world LLM-assisted learning and propose a non-invasive, post-hoc emotion-value gated framework that enhances emotional safety and supportive communication without altering underlying model architectures. Through a large-scale analysis of 42,172 authentic student–AI interactions and a controlled deployment involving 24 students, we show that emotionally risky responses occur with meaningful frequency in baseline usage and can be substantially reduced by our approach. Expert evaluations demonstrate significant improvements in emotional safety, supportiveness, and overall response quality, while student studies indicate increased comfort, engagement, and willingness to use the emotion-enhanced AI, with no perceived loss in response correctness. These findings highlight the importance of emotional governance in educational AI and demonstrate that system-level, deployable interventions can meaningfully improve students' learning experience in LLM-assisted environments.**

***Index Terms*—Emotion-aware AI, Large language models, Educational technology, Psychological safety, Learning support systems**

I. Introduction

Large language models (LLMs) have rapidly become embedded in students' everyday learning practices [1]–[3]. Beyond serving as tools for information retrieval, problem solving, and writing assistance, LLM-based systems are increasingly used as conversational partners that accompany students throughout the learning process [4]–[6]. In practice, students turn to AI not only for factual explanations or task completion, but also for reassurance, clarification, and emotional grounding when they experience confusion, frustration, self-doubt, or academic pressure [7], [8]. As a result, contemporary educational AI systems are no longer perceived solely as cognitive aids; they increasingly assume an affective role within learning environments, shaping how students feel about their learning as well as what they learn [9]–[12].

This shift raises important concerns regarding the emotional appropriateness of AI-generated responses [13], [14]. While modern LLMs demonstrate strong capabilities in producing fluent, coherent, and technically accurate explanations, they are not explicitly designed to account for learners' emotional vulnerability or situational sensitivity [15]. In authentic learning interactions, particularly when students repeatedly struggle or explicitly express uncertainty, even technically correct responses may unintentionally convey dismissiveness, implicit blame, or emotional coldness. Such pragmatic mismatches can undermine learners' confidence, weaken their motivation to persist, and negatively influence their overall learning experience [16]. Importantly, these risks are often subtle rather than overt: they emerge through tone, framing, or implied judgment, making them difficult to identify using conventional safety mechanisms that primarily target explicit toxicity or prohibited content [17], [18].

At the same time, institutional deployment constraints further complicate the challenge of addressing emotional risk in educational AI [19]. Many universities and learning platforms adopt LLM services through packaged, commercial, or API-based solutions due to considerations of cost, scalability, maintenance, and governance. In these settings, LLMs are typically treated as black-box systems [20]: institutions have limited or no access to model internals and cannot modify architectures, retrain parameters, or fine-tune behavior at scale. Consequently, even when educators and administrators recognize the potential emotional risks associated with AI-generated feedback, approaches that rely on model-level retraining or alignment are often infeasible in practice. There is therefore a growing need for solutions that can enhance emotional safety and supportive communication at the system level without altering the underlying LLM.

Taken together, these two limitations—firstly, the expanding emotional role of LLMs in student learning without corresponding emotional safeguards, and secondly, the black-box nature of LLM deployment in institutional contexts—highlight a critical gap in current learning technologies. Existing research has largely concentrated on improving model performance, explainability, or alignment during training, while

2836-3795/26/$31.00 ©2026 IEEE
DOI 10.1109/COMPSAC69091.2026.00199

comparatively little attention has been paid to post-hoc, deployable mechanisms that regulate the emotional quality of AI responses at the point of interaction [21]–[23]. As a result, educational platforms currently lack practical tools to systematically assess emotionally risky outputs and intervene in a timely manner, while still preserving the pedagogical usefulness and correctness of AI-generated responses.

To address these challenges, we propose a non-invasive, post-hoc emotion-value gated framework for AI-assisted learning. The framework operates entirely at the output level, evaluating each LLM-generated response for potential emotional risk and supportive value prior to delivery to learners. By combining rule-based detection with a lightweight classifier, the system identifies responses that may be psychologically unsafe or insufficiently supportive and applies targeted regeneration or rewriting when needed. Crucially, the proposed approach does not modify the underlying LLM architecture, parameters, or training process, making it readily compatible with black-box deployment scenarios that are common in institutional educational settings.

Through a large-scale analysis of authentic student–AI interactions and a controlled user study conducted in university contexts, we examine both the prevalence of emotionally risky responses in existing LLM-based learning support and the effectiveness of the proposed intervention. The results provide empirical evidence that emotion-aware post-processing can substantially improve the emotional and pedagogical quality of AI-generated responses while preserving factual correctness and practical deployability. In summary, our primary contributions can be summarized as follows:

- We provide empirical evidence that emotionally risky responses are a recurring phenomenon in real-world LLM-assisted learning interactions, even when models are used for routine academic support.
- We introduce a model-agnostic, post-hoc emotion-value gated framework that enhances psychological safety and supportive communication without modifying LLM architectures or retraining models.
- We develop a dual-objective assessment mechanism that jointly considers emotional risk and pedagogical support, enabling fine-grained, response-level intervention through regeneration and rewriting.
- We conduct expert-based evaluations and a student-centered user study to demonstrate that emotion-enhanced AI responses improve perceived support and emotional experience without compromising learning usefulness.

II. Methodology

A. Overview.

We propose a non-invasive, post-hoc intervention layer for general learning-oriented responses, designed to reduce negative affect and enhance supportive, motivating communication without modifying the architecture, parameters, or training process of the underlying large language model (LLM). Given a learner query x, a base LLM first produces a candidate response y_0 through standard inference. Our method operates entirely at the output level: y_0 is evaluated by a lightweight assessor that combines (i) rule-based pattern checks and (ii) a compact neural classifier. This posterior evaluation allows emotional risks and support deficiencies to be identified after generation, making the approach fully model-agnostic and compatible with any off-the-shelf or proprietary LLM.

If the candidate response is identified as emotionally harmful or insufficiently supportive, the intervention layer triggers a controlled correction process—either constrained regeneration or targeted rewriting—to produce the final response $y^\star$. Importantly, this process does not alter the base model's internal representations or decoding mechanism; instead, it acts as a *plug-and-play safeguard and enhancement module* that can be inserted between model inference and user delivery. This design enables fine-grained control over emotional quality while preserving the original model's knowledge, reasoning capabilities, and instructional content.

The method is built around three guiding goals. Firstly, *Safety-first*: the system enforces a strict boundary against dismissive, humiliating, blaming, threatening, or otherwise psychologically unsafe language in learning contexts, ensuring that harmful outputs are intercepted before reaching learners. Secondly, *Pedagogical integrity*: emotional intervention is explicitly constrained to avoid degrading factual correctness, explanatory depth, or task relevance, thereby preventing the system from producing responses that are merely polite but educationally unhelpful. Finally, *Deployability*: by relying on post-processing assessment and threshold-based decision rules rather than model retraining or fine-tuning, the system remains lightweight, modular, and easily deployable across diverse educational platforms, with transparent policy control that can be adjusted to different learner populations and institutional norms. To provide an end-to-end view of the proposed method before detailing each component, Algorithm 1 presents the overall emotion-value gated response generation pipeline, from initial LLM inference to final response delivery.

B. Emotion-risk and support-value dimensions.

We operationalize the emotional quality of learning-oriented responses using two complementary constructs: *harm risk* and *support value*. *Harm risk* captures negative affect triggers that may undermine learners' confidence and persistence, including ridicule, harsh judgment, subtle blame, cold dismissal, and discouraging suggestions to stop learning. *Support value* captures psychologically safe communication that preserves learner dignity while still being instructionally effective, such as normalizing difficulty, offering structured guidance, and promoting persistence through actionable encouragement rather than empty flattery. In our system, harm risk is treated as a hard constraint (must stay below a threshold), whereas support value is treated as a soft objective (should exceed a threshold) to prevent the system from returning responses that are technically correct yet emotionally neutral or pedagogically disengaging.

Algorithm 1 Emotion-Value Gated Response Generation

Require: Learner query x, base LLM $\mathcal{M}$, rule set $\mathcal{R}$, classifier $\mathcal{C}$, thresholds τ_R, τ_V, fusion weight α, max attempts $K{=}2$)
Ensure: Final response $y^\star$
$y_0 \leftarrow \mathcal{M}(x)$
$y_t \leftarrow y_0$
for $t = 0$ **to** $K-1$ **do**
  $(r_{\text{rule}}, \mathcal{Z}) \leftarrow \textsc{RuleScore}(\mathcal{R}, x, y_t)$ {$\mathcal{Z}$: triggered evidence/reasons}
  $r_{\text{clf}} \leftarrow \textsc{ClfScore}(\mathcal{C}, [x; y_t])$
  $R(x, y_t) \leftarrow \alpha \cdot r_{\text{rule}} + (1-\alpha) \cdot r_{\text{clf}}$
  $V(x, y_t) \leftarrow \textsc{SupportScore}(x, y_t)$
  if $R(x, y_t) \le \tau_R$ **and** $V(x, y_t) \ge \tau_V$ **then**
    return y_t
  else if $R(x, y_t) > \tau_R$ **then**
    $y_{t+1} \leftarrow \textsc{Regenerate}(\mathcal{M}, x, \mathcal{Z})$ {High risk: re-generate with explicit constraints from $\mathcal{Z}$}
  else
    $y_{t+1} \leftarrow \textsc{Rewrite}(\mathcal{M}, y_t)$ {Low support: add normalization + next step, preserve technical content}
  end if
  $y_t \leftarrow y_{t+1}$
end for
return $\textsc{SafeFallback}(x)$ {Bounded latency; conservative safe response}

Given a learner query x and an initial model response y_0, our assessor computes (i) a rule-based harm score $r_{\text{rule}}(x, y_0)$ and a set of triggered reasons $\mathcal{Z}$, and (ii) a classifier-based harm probability $r_{\text{clf}}(x, y_0)$. These signals are later fused into an overall risk score and used to determine whether the response is directly delivered, rewritten, or regenerated. Importantly, the workflow is *posterior and model-agnostic*: the base LLM is treated as a black box that produces y_0, and our layer only inspects and intervenes at the output stage before user delivery.

The first component is a rule-based scorer that detects emotionally harmful cues with high precision. It uses a curated set of lexical triggers and short syntactic templates tailored to learning interactions, including: (1) humiliating labels (e.g., "stupid", "useless"), (2) blame and moral condemnation (e.g., "it's your fault"), (3) dismissive refusal (e.g., "I won't explain", "figure it out"), (4) absolute negative judgments (e.g., "you will never understand"), and (5) threat-like language or coercion. Rules are implemented as weighted matches over unigrams, bigrams, and lightweight dependency-like patterns (e.g., "you are + [negative adjective]", "stop wasting time"). The output is a normalized risk score $r_{\text{rule}}(x, y) \in [0, 1]$ and an evidence set of triggered reasons $\mathcal{Z}$, which supports interpretability and downstream constrained regeneration.

To clarify what we consider "emotionally risky" in authentic learning contexts, we provide two mild examples that can occur in real LLM usage. Consider a learner query: *"I don't understand why the derivative of* $\sin(x)$ *is* $\cos(x)$*."* An LLM might respond with a technically correct explanation but add phrasing such as: *"This is basic calculus; you should already know it."* While not an explicit insult, the statement can still convey judgment and discourage learners, and our rule-based scorer flags it via patterns indicating belittlement or dismissiveness. As a second example, for a query like *"I keep getting the wrong answer in this equation, what should I do?"*, a response such as *"Just redo the steps carefully; it's probably your mistake."* may be perceived as blaming and non-supportive, especially when the learner is already frustrated. In such cases, the rule-based scorer produces a moderate r_{rule} and records reasons (e.g., blame/dismissive tone) in $\mathcal{Z}$, which later guides the rewriting/regeneration step.

Because rule matching can miss subtle dismissiveness or misfire in benign contexts (e.g., technical uses of words that also appear in negative lexicons), we add a compact classifier for context-sensitive assessment. The classifier treats the learner query and the model response as a single paired input, encoding the concatenated sequence $[x; y]$ using a pretrained RoBERTa [24]encoder followed by a lightweight linear classification head. This design leverages contextual representations to distinguish between (i) language that is harsh *toward the learner* and (ii) neutral technical wording that may superficially resemble negative expressions. The classifier outputs (a) a scalar harm probability $r_{\text{clf}}(x, y) \in [0, 1]$ and (b) optional multi-label harm types (humiliation, blame, dismissive tone, threat, coldness) to support analysis and targeted interventions.

In implementation, the encoder produces a pooled representation $\mathbf{h} = \textsc{RoBERTa}([x; y])$, using the `[CLS]` token, which is then passed to a linear layer with a sigmoid activation for binary risk prediction. When multi-label harm types are enabled, we use a parallel multi-label head trained with binary cross-entropy to estimate type-specific probabilities. Training minimizes:

$$\mathcal{L} = \mathcal{L}_{\text{risk}} + \lambda_{\text{type}} \mathcal{L}_{\text{type}}, \tag{1}$$

where $\mathcal{L}_{\text{risk}}$ is the binary cross-entropy for harm/non-harm classification and $\mathcal{L}_{\text{type}}$ is the multi-label loss (optional). This objective maintains lightweight inference while improving sensitivity to nuanced tone and learner-directed judgment.

We train and validate the classifier using an interaction dataset collected from a large university during a three-month period in 2024. The dataset contains $32{,}470$ learning Q&A interactions produced in authentic study scenarios (e.g., concept clarification, problem solving, and writing support). All samples were pre-processed to remove personally identifiable information and were then manually reviewed and annotated by trained annotators following a structured rubric. Each interaction was labeled with (i) a binary harm indicator, (ii) harm intensity on an ordinal scale (used for analysis and threshold calibration), and (iii) optional harm type tags (humiliation, blame, dismissiveness, threat, coldness). Disagreements were resolved through adjudication to ensure consistent labels, resulting in a high-quality dataset suitable for supervised training of the compact classifier.

Notably, in our study, a key design choice is to evaluate the response *conditioned on the learner query*. In educational settings, the same phrase can be interpreted differently depending on what the learner asked (e.g., whether

the learner expressed confusion or frustration). By encoding $[x; y]$ jointly, the classifier captures signals such as learner difficulty markers ("I'm stuck", "I'm confused") and assesses whether the response escalates negative affect (e.g., blame) or provides psychologically safe scaffolding. This is particularly important for general learning assistance, where emotional appropriateness is tightly coupled with the learner's expressed state.

Finally, the assessor returns a tuple $(r_{\text{rule}}, \mathcal{Z}, r_{\text{clf}}, \hat{\mathbf{t}})$, where $\hat{\mathbf{t}}$ denotes optional harm-type probabilities. This tuple enables both threshold-based gating and *actionable* interventions: $\mathcal{Z}$ and $\hat{\mathbf{t}}$ are used to generate precise constraints (for regeneration) or edit instructions (for rewriting), ensuring that the intervention corrects the specific emotional failure mode while preserving instructional content.

C. Score fusion and decision gate.

Building on the dual assessment signals, we next integrate rule-based and classifier-based judgments into a unified decision mechanism that determines whether an LLM response can be safely delivered, requires tone enhancement, or must be regenerated. This fusion step translates heterogeneous emotional assessments into explicit, threshold-controlled policies that govern downstream intervention. We combine rule and classifier signals into a single harm score:

$$R(x, y) = \alpha \cdot r_{\text{rule}}(x, y) + (1 - \alpha) \cdot r_{\text{clf}}(x, y), \qquad (2)$$

where $r_{\text{rule}}(x, y)$ captures high-precision lexical and pattern-based risk cues, $r_{\text{clf}}(x, y)$ captures context-sensitive harm probability conditioned on the learner query, and $\alpha \in [0, 1]$ controls the precision–recall trade-off between the two. In practice, a higher α prioritizes conservative rule-based screening, while a lower α increases sensitivity to subtle, context-dependent dismissiveness identified by the classifier.

In parallel, we compute a *support value* score $V(x, y) \in [0, 1]$, which is intentionally kept separate from the harm score. While $R(x, y)$ focuses on preventing emotional damage, $V(x, y)$ measures whether a response actively supports learning rather than merely avoiding harm. We accept a candidate response if and only if it satisfies:

$$\textsc{Accept}(x, y) \Leftrightarrow \big(R(x, y) \le \tau_R\big) \ \wedge \ \big(V(x, y) \ge \tau_V\big), \quad (3)$$

where τ_R is a strict risk threshold enforcing a safety boundary, and τ_V is a minimum support threshold ensuring pedagogical usefulness. In deployment, τ_R is set conservatively to minimize emotional harm, whereas τ_V can be adjusted to balance supportive depth against verbosity or cognitive load.

Consistent with the support-value dimensions defined earlier, $V(x, y)$ is computed as a weighted sum of three educationally grounded cues:

$$V(x, y) = w_1 S_{\text{safe}}(x, y) + w_2 S_{\text{act}}(x, y) + w_3 S_{\text{agency}}(x, y), \quad (4)$$

where $w_1 + w_2 + w_3 = 1$. The *psychological safety cue* S_{safe} captures non-judgmental language and normalization of difficulty (e.g., acknowledging that confusion is common). The *actionability cue* S_{act} captures the presence of concrete steps, examples, or checks that the learner can immediately perform. The *agency cue* S_{agency} captures language that enhances perceived control and self-efficacy (e.g., "start with this step", "try checking whether..."). Each cue is implemented via short pattern lists and lightweight structural checks (such as detection of enumerated steps or explicit next actions). This design avoids over-optimizing for generic positivity and ensures that support value reflects constructive educational scaffolding rather than superficial praise.

D. Intervention policies

When a candidate response fails the decision gate in Eq. 3, we apply an intervention policy determined by the type of failure. If the harm score $R(x, y)$ exceeds the safety threshold τ_R, the response is considered potentially harmful, and we trigger **regeneration** with explicit constraints derived from the rule evidence set $\mathcal{Z}$ and, when available, predicted harm types. If $R(x, y) \le \tau_R$ but $V(x, y) < \tau_V$, the response is deemed emotionally safe but insufficiently supportive; in this case, we trigger a lower-cost **rewrite** operation that preserves factual content while improving tone, structure, and learner guidance. This two-path policy prioritizes safety while controlling computational cost by reserving full regeneration for genuinely risky outputs.

Both interventions are implemented by prompting the same base LLM, preserving the architecture-agnostic nature of the system. For regeneration, we provide the learner query x together with a constraint list that explicitly prohibits detected harmful traits and requires non-judgmental, stepwise guidance. A typical regeneration prompt is as follows:

> ***System prompt (regeneration).*** *You are an educational assistant. The previous answer was flagged for emotional risk due to the following issues: {$\mathcal{Z}$}. Re-generate a response that is psychologically safe and non-judgmental, acknowledges that learning difficulties are normal, and provides clear, step-by-step guidance with a concrete next action. Avoid blaming, dismissive language, ridicule, threats, or discouraging the learner. Keep the explanation accurate and focused on helping the learner progress.*
>
> ***Learner question:*** *{x}*

For rewriting, we pass the original response y and instruct the model to adjust tone and structure without altering technical correctness:

> ***System prompt (rewrite).*** *Rewrite the following answer to improve psychological safety and supportive tone without changing its technical content. Remove any harsh, dismissive, or judgmental phrasing. Add brief normalization of difficulty and a clear next step the learner can take. Avoid excessive praise or vague encouragement.*
>
> ***Original answer:*** *{y}*

To prevent overly "sweet" or verbose outputs, both prompts include constraints that require actionable help and discourage generic praise, ensuring that emotional support remains tightly coupled with learning utility.

We iterate the assess–intervene loop for up to two attempts. If the system still fails to produce an accepted response, it returns a conservative fallback template: a neutral and polite answer that (i) acknowledges the learner's difficulty, (ii) provides a minimal but concrete next step, and (iii) suggests seeking human support (e.g., a teacher or peer) when appropriate. This design guarantees bounded latency and prevents the delivery of emotionally risky content even under adversarial, ambiguous, or repeatedly failing prompts.

Finally, the full pipeline remains modular and auditable. The assessor can be updated independently of the base LLM, thresholds τ_R and τ_V provide explicit policy control, and the rule evidence set $\mathcal{Z}$ offers interpretable justification for each intervention. This enables educators and institutions to align the system with local norms and learner populations (e.g., stricter harm thresholds for novice or younger learners). Table I summarizes the assessed dimensions, typical triggers, and their roles in the decision process.

III. Research Questions

This study aims to systematically investigate the prevalence, impact, and educational implications of emotionally risky responses generated by LLMs in authentic learning contexts, as well as the effectiveness of a post-hoc emotion-value enhancement framework. We formulate three research questions that collectively examine the necessity, performance, and experiential consequences of emotion-aware intervention in AI-assisted learning.

RQ1: How prevalent are emotionally risky responses in existing LLM-based learning support?

In this section, we examine the baseline risk landscape of LLM-generated responses in real-world student usage. Specifically, we ask: *What proportion of LLM responses in authentic learning interactions exhibit potential emotional risk, such as dismissiveness, subtle blame, or psychologically unsafe tone?* This question establishes the empirical necessity of emotional risk mitigation and motivates the proposed intervention.

To address RQ1, we collaborated with two comprehensive universities and collected all LLM-generated responses to student queries over a continuous 10-day period. During this period, students interacted with a commonly deployed, non-modified LLM-based assistant for routine learning purposes, including concept clarification, problem solving, writing assistance, and general academic inquiries. In total, we collected 42,172 LLM responses paired with the corresponding student questions.

All collected responses were independently evaluated by a panel of four domain experts, consisting of two PhD-level linguists with expertise in discourse and pragmatics, and two senior university instructors with extensive frontline teaching experience. Prior to annotation, the experts jointly developed and refined an evaluation rubric through iterative discussion, explicitly defining what constitutes emotionally risky language in learning contexts (e.g., implicit belittlement, discouraging tone, learner-directed blame, or emotional coldness). Each response was labeled for the presence or absence of emotional risk, along with a qualitative categorization of risk type when applicable. Disagreements were resolved through adjudication, ensuring a consistent and conservative estimation of risk prevalence.

RQ2: How does the proposed emotion-enhanced AI perform in improving response quality?

In this section, we evaluate the effectiveness of the proposed emotion-value gated framework in improving the emotional and pedagogical quality of AI-generated responses. To answer RQ2, we deployed the proposed framework in the same two universities, integrating it as a post-processing layer on top of the existing LLM without modifying the underlying model. Twenty-four volunteer students participated in a 12-day usage study, during which they interacted freely with the AI assistant as part of their daily academic routines. Students were instructed to use the system naturally, including asking both learning-related questions (e.g., coursework, exam preparation, writing) and broader questions related to stress, confusion, or everyday student life.

Across the study period, we collected a total of 27,000 AI responses after filtering out extremely short or non-informative interactions. These responses were again evaluated by the same four experts using a consistent rubric. Each response was assessed along at least four dimensions: (1) emotional safety (absence of harmful or discouraging tone), (2) supportive value (degree of encouragement and normalization), (3) pedagogical helpfulness (clarity, correctness, and actionability), and (4) overall response quality. This design allows a direct comparison between baseline and emotion-enhanced responses while controlling for evaluator bias and contextual variation.

RQ3: Does emotion-enhanced AI influence students' learning engagement and emotional experience?

In this section, we focus on learner-centered outcomes beyond response-level quality. The same 24 students who participated in RQ2 were invited to maintain lightweight usage logs throughout the 12-day study period, recording notable interaction experiences, emotional reactions, and perceived usefulness. Following the deployment, participants completed a post-study questionnaire measuring changes in learning enthusiasm, emotional comfort, perceived psychological safety, and willingness to continue using AI for learning. Importantly, all participants had prior experience using standard LLM-based tools without emotional intervention, enabling within-subject reflection on perceived differences before and after integration of the proposed framework.

By combining longitudinal usage logs with self-reported survey responses, RQ3 captures the experiential and affective impact of emotion-enhanced AI at the learner level. Together, the three research questions provide a comprehensive evaluation of (i) the scope of emotional risk in current LLM usage, (ii) the technical effectiveness of the proposed intervention, and (iii) its educational significance in shaping students' learning experience.

TABLE I: Assessment dimensions and their roles in the proposed emotion-value gating layer.

Dimension	Operational cues (examples)	Role in gate
Harm risk (Rule)	Insults/labels ("stupid"), blame ("your fault"), dismissive refusal ("figure it out"), absolute negation ("never understand"), threat-like phrasing	Contributes to r_{rule} in Eq. 2
Harm risk (Classifier)	Context-dependent harshness, subtle contempt, coldness given learner struggle; uses $[x; y]$ to avoid false positives	Contributes to r_{clf} in Eq. 2
Support value	Normalization ("this is common"), stepwise guidance (enumerated steps), next action ("try X"), agency cues ("start with")	Must satisfy $V(x, y) \geq \tau_V$ in Eq. 3
Decision policy	If $R > \tau_R$: regenerate; else if $V < \tau_V$: rewrite; else accept	Determines intervention type

IV. Results and Analysis

A. Prevalence of Emotional Risk in LLM Responses (RQ1)

In this section, we focus on the prevalence and characteristics of emotionally risky responses produced by a widely used LLM in authentic student learning interactions. Across the 42,172 LLM-generated responses collected during the 10-day period, expert evaluation identified a non-trivial proportion of responses containing potential emotional risk. Specifically, 6,384 responses (15.1%) were labeled as emotionally risky by at least three of the four experts. While the majority of responses were technically correct and neutral in tone, this result indicates that emotionally problematic outputs are not isolated edge cases but occur with meaningful frequency in routine educational use.

Table II summarizes the distribution of different emotional risk categories identified by the expert panel. The most common risk type was *dismissive or emotionally cold tone*, followed by *learner-directed blame* and *implicit belittlement*. Explicitly hostile or threatening language was rare, suggesting that most risks manifest in subtle, everyday phrasing rather than extreme expressions typically targeted by existing safety filters.

Importantly, most emotionally risky responses were not overtly offensive or abusive. Instead, they often involved subtle pragmatic cues that could be interpreted as judgmental, dismissive, or emotionally unsupportive in a learning context—especially when students explicitly expressed confusion or repeated difficulty. Experts noted that such responses would likely pass conventional content moderation systems but could still negatively affect learners' confidence and motivation.

One linguistics expert commented that "many of these responses are linguistically polite but pragmatically cold, which can be discouraging when learners are already uncertain." A teaching professor similarly observed that "from a pedagogical perspective, telling students that something is 'basic' or 'obvious' may unintentionally signal that their struggle is illegitimate." These expert reflections highlight that emotional risk in learning-oriented AI is often context-dependent and tightly coupled with instructional discourse norms.

Although a 15.1% risk rate may appear moderate, its impact should be interpreted in light of usage frequency. Given that students may interact with AI systems multiple times per day, repeated exposure to emotionally unsupportive responses could accumulate and influence learners' willingness to seek help, persist through difficulty, or trust AI-assisted learning tools. This is particularly concerning for novice learners or students experiencing academic stress.

The findings also suggest that existing safety mechanisms, which typically focus on preventing explicit toxicity or prohibited content, are insufficient for educational contexts. The majority of risky responses identified in Table II would not be flagged by standard filters, as they do not violate explicit safety policies. This gap underscores the need for domain-specific emotional assessment that accounts for pedagogical and affective norms.

Answer to RQ1: The results demonstrate that emotionally risky responses are a recurring phenomenon in real-world LLM-assisted learning, even when models are used for routine academic support. These findings empirically justify the need for a dedicated, post-hoc mechanism that can detect and mitigate subtle emotional risks while preserving instructional value. This motivation directly informs the design of our emotion-value-gated framework and guides its evaluation in subsequent studies.

B. Improvement in Emotional and Pedagogical Quality (RQ2)

In this section, we examine whether the proposed emotion-value gated framework improves the quality of AI-generated responses compared to the baseline LLM. We focus on both the reduction of emotionally risky outputs and improvements across multiple expert-evaluated quality dimensions. We first compare the distribution of emotionally risky response types before and after integrating the proposed framework. Using the same expert rubric as in RQ1, we evaluate the 27,000 responses generated with emotion-value gating and contrast them with the baseline distribution reported in Table II. As shown in Table III, the proportion of emotionally risky responses decreases substantially across all categories.

Overall, the total proportion of emotionally risky responses drops from 15.1% to 4.5%, representing a relative reduction of more than 70%. Notably, the largest absolute reductions occur in dismissive tone and learner-directed blame, which were also the most prevalent risk types in baseline usage.

Beyond risk reduction, we examine whether emotion-value gating improves the overall quality of responses. Each response was rated by experts on a 0–5 Likert scale across four dimensions: emotional safety, supportive value, pedagogical helpfulness, and overall quality. Table IV reports the average scores for baseline and emotion-enhanced responses.

Across all four dimensions, emotion-enhanced responses achieve consistently higher scores. Improvements are most pronounced in emotional safety and supportive value, directly reflecting the design objectives of the proposed framework.

TABLE II: Distribution of emotionally risky response types in baseline LLM-generated learning support (N=42,172).

Risk type	Count	Proportion	Illustrative examples (excerpt)
Dismissive / emotionally cold tone	2,547	6.0%	"This is straightforward. Just follow the steps." "The explanation is already clear from the formula." "There is no need to go into more detail here." "You can verify this on your own." "This does not require further discussion."
Learner-directed blame	1,812	4.3%	"You made a mistake in the previous step." "This happens because you did not apply the rule correctly." "The issue comes from how you set up the equation." "Your confusion is due to an earlier error." "You may have overlooked a basic assumption."
Implicit belittlement	1,204	2.9%	"This is basic knowledge you should already know." "Most students learn this early on." "This concept is usually not difficult." "At this level, this should be familiar." "This is a standard result."
Discouraging persistence	602	1.4%	"If this still confuses you, it may not be worth focusing on." "You might want to move on to something simpler." "It may be better to skip this for now." "This topic might not be essential for you." "You do not need to spend too much time on this."
Other subtle negative cues	219	0.5%	"There is nothing special to explain here." "This follows directly from the definition." "The result should be obvious from the context." "The answer is already implied." "No additional clarification is necessary."
Total emotionally risky	**6,384**	**15.1%**	–

TABLE III: Comparison of emotionally risky response proportions before and after emotion-value gating.

Risk type	Baseline (RQ1)	Emotion-enhanced
Dismissive / emotionally cold tone	6.0%	1.8%
Learner-directed blame	4.3%	1.2%
Implicit belittlement	2.9%	0.7%
Discouraging persistence	1.4%	0.6%
Other subtle negative cues	0.5%	0.2%
Total emotionally risky	**15.1%**	**4.5%**

TABLE IV: Expert evaluation of response quality across four dimensions (0–5 scale).

Dimension	Baseline LLM	Emotion-enhanced AI
Emotional safety	3.42	4.24
Supportive value	3.67	4.39
Pedagogical helpfulness	3.91	4.46
Overall response quality	3.38	4.05

Importantly, pedagogical helpfulness also improves, indicating that emotional intervention does not dilute instructional quality.

Expert feedback further clarifies the nature of these improvements. One teaching professor noted that "the enhanced responses not only avoid discouraging language, but actively guide students toward manageable next steps, which aligns well with effective instructional practice." A linguistics expert observed that "the revised responses demonstrate better pragmatic alignment with learner vulnerability, especially when students express confusion or stress."

We observe a systematic increase in response length after emotion-value gating. On average, emotion-enhanced responses are 18.42% longer than baseline outputs. Importantly, expert analysis indicates that this increase does not stem from unnecessary verbosity or repetitive phrasing. Instead, the additional content primarily consists of explanatory scaffolding, normalization of learning difficulty, and the inclusion of explicit next actions that learners can take. In many cases, responses expand by clarifying intermediate reasoning steps or by explicitly acknowledging common points of confusion, thereby making the instructional guidance more accessible. This pattern suggests that the proposed framework promotes richer and more structured instructional discourse, rather than generic elaboration or inflated explanations.

A further qualitative finding concerns the way emotion-enhanced responses handle situations of sustained confusion, emotional distress, or non-academic concerns. Compared to baseline outputs, the post-processed AI more frequently encourages students to seek appropriate external support—such as consulting a course instructor, tutor, or counselor—when continued self-guided interaction may be insufficient. Notably, these suggestions are framed in a supportive and non-dismissive manner, emphasizing that seeking help is a normal and responsible step rather than a sign of failure. Experts characterized this behavior as *educationally responsible*, highlighting that it aligns with good teaching practice by recognizing the limits of automated assistance.

Despite the overall increase in supportive tone, experts did not observe a corresponding rise in formulaic or superficial encouragement, such as repetitive motivational slogans or empty reassurance. This distinction is important: the enhanced

TABLE V: Post-study survey results on students' perceptions of emotion-enhanced AI (N=24).

Survey item	Mean score
The AI responses felt emotionally supportive	4.63
The tone of the AI responses improved compared to previous use	4.59
I felt more comfortable asking questions to the AI	4.24
The AI encouraged me to continue learning when I was confused	4.06
The correctness of AI answers changed significantly	3.28
I am willing to continue using this AI for learning	4.41

responses provide emotional support that is tightly coupled with concrete guidance and situational awareness, rather than relying on generic praise. This observation is consistent with the framework's explicit constraints against empty flattery and supports the claim that emotion-aware post-hoc intervention can foster substantive, context-sensitive support without compromising instructional rigor.

Answer to RQ2: The results demonstrate that the proposed emotion-value gated framework substantially improves the emotional safety and pedagogical quality of LLM-generated responses. The framework achieves significant reductions in emotionally risky outputs while simultaneously enhancing supportiveness and instructional value, validating the effectiveness of post-hoc emotional intervention.

C. Student Engagement and Emotional Experience (RQ3)

In this section, we examine how sustained interaction with the emotion-enhanced AI influences students' learning engagement and emotional experience. We combine quantitative survey data with qualitative analysis of usage logs and open-ended responses to capture both perceived changes and lived interaction experiences. At the end of the 12-day study period, all 24 participants completed a post-study questionnaire using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The survey measured perceived changes in learning enthusiasm, emotional comfort, psychological safety, response correctness, and willingness to continue using the AI system. Table V summarizes the aggregated results.

Notably, students reported a clear improvement in emotional tone and comfort, while perceiving little change in response correctness. This distinction is critical: participants generally viewed the emotion-enhanced AI as *emotionally different rather than technically different*, suggesting that the framework improves affective experience without altering perceived cognitive accuracy.

The relatively neutral score for perceived correctness (mean = 3.28) indicates that students did not attribute major gains or losses in technical accuracy to the intervention. Instead, improvements were concentrated in emotional and experiential dimensions. This aligns with the design goal of preserving pedagogical integrity while enhancing emotional support, and helps mitigate concerns that emotional intervention may dilute instructional rigor.

To further understand how students experienced the emotion-enhanced AI in practice, we conducted a qualitative analysis of usage logs and open-ended questionnaire responses using NVivo. All textual data were coded independently by two researchers, with discrepancies resolved through discussion. The analysis yielded several recurring themes, summarized in Table VI.

Students frequently described increased willingness to engage with the AI, particularly when experiencing confusion or repeated difficulty. Several logs indicated that students who previously avoided asking "basic" questions felt more comfortable doing so. Importantly, the emotion-enhanced AI did not position itself as a replacement for human support; instead, it occasionally encouraged students to consult instructors or counselors when appropriate, which participants perceived as responsible and supportive.

A recurring theme across logs and open-ended responses was a clear preference for the emotion-enhanced AI over prior experiences with standard LLM-based assistants. Students described the post-processed responses as "more considerate," "less mechanical," and "more like a supportive tutor." This preference suggests that emotional tone meaningfully shapes user acceptance, even when technical performance remains largely unchanged. While short-term studies cannot fully capture long-term motivational change, several participants reported increased willingness to persist through difficult tasks during the study period. Logs often linked this persistence to the AI's non-judgmental tone and explicit normalization of struggle, which reduced feelings of inadequacy and frustration.

Answer to RQ3: The results indicate that emotion-enhanced AI positively influences students' emotional experience and engagement with AI-assisted learning. Students perceived clear improvements in tone, comfort, and psychological safety, while viewing response correctness as largely unchanged. The combination of increased acceptance, reduced anxiety, and responsible guidance toward human support underscores the educational value of post-hoc emotional intervention.

V. Discussion

This study set out to examine emotional risk in LLM-assisted learning and to evaluate whether a post-hoc, emotion-aware intervention can improve students' learning experience without altering underlying model architectures. Across three complementary research questions, our findings consistently demonstrate that emotionally risky responses are a recurring phenomenon in authentic educational use, and that targeted emotional intervention at the output level can meaningfully improve both response quality and learner experience.

First, the results of RQ1 challenge the common assumption that emotional risk in educational AI is rare or confined to extreme cases. Although the baseline LLM rarely produced overtly offensive language, a substantial proportion of responses exhibited subtle but pedagogically consequential risks, such as dismissiveness, implicit belittlement, or emotional coldness. These risks are particularly salient in learning contexts, where students' questions often reflect uncertainty or vulnerability. This finding aligns with emerging concerns in educational technology that affective misalignment [25],

TABLE VI: Themes identified from NVivo analysis of usage logs and open-ended responses.

Theme	Frequency	Representative excerpt
Increased willingness to ask questions	High	"I felt less hesitant to ask even simple questions."
Emotional reassurance during difficulty	High	"The AI made me feel that being confused is normal."
Preference for emotion-enhanced AI	Medium–High	"I prefer this version because it feels more considerate."
Encouragement to seek human support	Medium	"It suggested talking to my teacher when I was really stuck."
Reduced anxiety during learning	Medium	"I was less stressed when the answer was not judgmental."

[26]—rather than explicit toxicity—poses a greater threat to sustained learning engagement.

Second, the RQ2 results demonstrate that emotional quality can be systematically improved without sacrificing pedagogical integrity. The proposed emotion-value gated framework substantially reduced the prevalence of emotionally risky responses while simultaneously increasing expert-rated emotional safety, supportiveness, and overall quality. Importantly, pedagogical helpfulness also improved rather than declined, countering the concern that emotionally supportive language may dilute instructional rigor. These findings suggest that emotional scaffolding and cognitive scaffolding are not competing objectives, but can be jointly optimized through careful system-level design.

From a learner-centered perspective, RQ3 highlights the experiential significance of emotional intervention. Students consistently reported improved comfort, psychological safety, and willingness to engage with the AI, while perceiving little change in response correctness. This distinction is critical: the value of emotion-enhanced AI lies not in making the system appear "smarter," but in making it feel safer and more supportive to interact with. Qualitative evidence further indicates that students preferred the post-processed AI and were more open to help-seeking behaviors, including consulting teachers or counselors when appropriate. This suggests that emotionally responsible AI does not replace human support, but can function as a constructive bridge within the learning ecosystem.

When compared to existing paradigms [27], [28], our work differs in both focus and methodology. Prior approaches to improving AI behavior in education have largely emphasized model-centric solutions, such as fine-tuning, reinforcement learning, or alignment during training [29], [30]. While effective in controlled settings, these approaches are often impractical for institutional deployment, where LLMs are accessed as black-box services. In contrast, our framework adopts a system-level, post-hoc perspective, treating emotional governance as an operational layer rather than a modeling problem. This shift enables compatibility with off-the-shelf LLMs and aligns more closely with real-world constraints faced by educational institutions.

More broadly, our findings contribute to an emerging reframing of AI's role in learning. Rather than viewing LLMs solely as cognitive tools that deliver information, this study underscores their affective function as conversational partners that shape learners' emotional states and learning dispositions. Emotional safety, supportiveness, and responsible guidance emerge as essential dimensions of educational AI quality, alongside accuracy and clarity. Designing for these dimensions requires moving beyond generic safety filters toward domain-sensitive, pedagogically grounded emotional assessment.

Finally, the results suggest practical implications for educational platforms and policymakers. By demonstrating that emotion-aware intervention can be achieved without modifying core models, this work lowers the barrier for institutions to adopt emotionally responsible AI systems. Threshold-based gating and interpretable rules provide transparent control over emotional standards, allowing alignment with institutional values and learner populations. Taken together, our findings advocate for a shift from model-only optimization toward holistic, deployable frameworks that integrate emotional governance into the design of AI-assisted learning environments.

VI. Threats to Validity

a) External validity: The empirical evaluation was conducted in collaboration with two comprehensive universities and involved a relatively small group of student participants in the intervention studies. Although the datasets include a large number of authentic AI-generated responses, the student population and institutional context may not fully represent learners from other educational systems, disciplines, or cultural backgrounds. Consequently, the prevalence of emotional risk and the observed benefits of emotion-enhanced AI may vary in different settings.

b) Construct validity: Emotional risk and supportiveness were assessed using expert-defined rubrics and subjective judgments, which, despite careful design and adjudication, inevitably involve interpretive elements. While we employed multiple experts from complementary backgrounds and used consistent criteria across studies, alternative operationalizations of emotional safety or learner support could lead to different assessments. In addition, self-reported survey measures in RQ3 may be influenced by participants' perceptions or expectations rather than solely by actual interaction quality.

c) Ecological and temporal validity: The intervention studies spanned a limited time window (12 days), which constrains our ability to draw conclusions about long-term effects on learning motivation, emotional resilience, or academic outcomes. Students' positive responses may partially reflect novelty effects associated with interacting with a newly enhanced system. Longer-term deployments are needed to examine whether the observed improvements in emotional experience and engagement persist over time.

VII. Conclusion

This study demonstrates that emotional risk is a non-trivial and recurring issue in LLM-assisted learning, even when models are used for routine academic support, and that such risks can be effectively mitigated through post-hoc,

system-level intervention. By introducing an emotion-value gated framework that operates entirely at the output level, we show that emotional safety and supportive communication can be significantly improved without modifying underlying model architectures or compromising pedagogical usefulness. Empirical results from expert evaluations and student-centered studies indicate that emotion-enhanced AI responses reduce subtle emotional risks, improve perceived psychological safety, and increase learners' willingness to engage with AI-assisted learning tools. Future work will examine longer-term use in diverse educational settings, adaptive thresholds for different learner groups, and the integration of emotion-aware intervention with learning analytics to support personalized and responsible educational AI systems.

ACKNOWLEDGMENT

The work described in this paper is partially supported by the Hong Kong Metropolitan University Research Grant (Project Reference No. RD/2025/1.24 and RD/2025/1.21), and partially supported by the grant from the Research Grant Council (RGC) of the Hong Kong Special Administrative Region, China (Project Reference No. UGC/FDS16/E27/25).

REFERENCES

[1] S. García-Méndez, F. de Arriba-Pérez, and M. d. C. Somoza-López, "A review on the use of large language models as virtual tutors," *Science & Education*, vol. 34, no. 2, pp. 877–892, 2025.

[2] Y. Sun, H. K. Yu, J. Keung, Y. Cao, and Y. Liao, "Stulac: An adaptive llm-driven framework for scalable student feedback analysis in software-driven educational systems," in *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2025, pp. 121–130.

[3] W. Xing, N. Nixon, S. Crossley, P. Denny, A. Lan, J. Stamper, and Z. Yu, "The use of large language models in education," *International Journal of Artificial Intelligence in Education*, vol. 35, no. 2, pp. 439–443, 2025.

[4] D. Teng, X. Wang, Y. Xia, Y. Zhang, L. Tang, Q. Chen, R. Zhang, S. Xie, and W. Yu, "Investigating the utilization and impact of large language model-based intelligent teaching assistants in flipped classrooms," *Education and Information Technologies*, vol. 30, no. 8, pp. 10 777–10 810, 2025.

[5] C. Kooli and N. Yusuf, "Transforming educational assessment: Insights into the use of chatgpt and large language models in grading," *International Journal of Human–Computer Interaction*, vol. 41, no. 5, pp. 3388–3399, 2025.

[6] X. Sun, Q. Liu, K. Zhang, S. Shen, L. Yang, and H. Li, "Harnessing code domain insights: Enhancing programming knowledge tracing with large language models," *Knowledge-Based Systems*, vol. 317, p. 113396, 2025.

[7] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, "The role of large language models in personalized learning: a systematic review of educational impact," *Discover Sustainability*, vol. 6, no. 1, pp. 1–24, 2025.

[8] Y. Sun, Y. Liao, and X. Ma, "Trusting ai to detect ai? a systematic evaluation of the reliability and robustness of current aigc detection tools for student academic work," *Computers & Education*, p. 105616, 2026.

[9] T. Shahzad, T. Mazhar, M. U. Tariq, W. Ahmad, K. Ouahada, and H. Hamam, "A comprehensive review of large language models: issues and solutions in learning environments," *Discover Sustainability*, vol. 6, no. 1, p. 27, 2025.

[10] Y. Zhu, C. Zhu, T. Wu, S. Wang, Y. Zhou, J. Chen, F. Wu, and Y. Li, "Impact of assignment completion assisted by large language model-based chatbot on middle school students' learning," *Education and Information Technologies*, vol. 30, no. 2, pp. 2429–2461, 2025.

[11] P. Wang, K. Yin, M. Zhang, Y. Zheng, T. Zhang, Y. Kang, and X. Feng, "The effect of incorporating large language models into the teaching on critical thinking disposition: An "ai+ constructivism learning theory" attempt," *Education and Information Technologies*, pp. 1–23, 2025.

[12] N. Raihan, M. L. Siddiq, J. C. Santos, and M. Zampieri, "Large language models in computer science education: A systematic literature review," in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, 2025, pp. 938–944.

[13] A. Dorigoni and P. L. Giardino, "The illusion of empathy: evaluating ai-generated outputs in moments that matter," *Frontiers in Psychology*, vol. 16, p. 1568911, 2025.

[14] V. S. Diwanji, M. Geana, J. Pei, N. Nguyen, N. Izhar, and R. H. Chaif, "Consumers' emotional responses to ai-generated versus human-generated content: the role of perceived agency, affect and gaze in health marketing," *International Journal of Human–Computer Interaction*, pp. 1–21, 2025.

[15] E.-A. Diordiev, V. A. Stenberdt, and G. Makransky, "Designing emotionally intelligent generative ai: enhancing self-regulated learning and socio-emotional outcomes in education," *PsychArchives. https://doi.org/10.23668/psycharchives*, vol. 16423, 2025.

[16] Y. Cui and H. Zhang, "Can student accurately identify artificial intelligence generated content? an exploration of aigc credibility from user perspective in education," *Education and Information Technologies*, pp. 1–26, 2025.

[17] C. Wang, Y. Du, and B. Zou, "Learners' acceptance and use of multimodal artificial intelligence (ai)-generated content in ai-mediated informal digital learning of english," *International Journal of Applied Linguistics*, 2025.

[18] O. Shugurova, ""it's thrilling and liberating, but i worry": Explorations of ai-induced emotions from a dialogical perspective," *Culture & Psychology*, p. 1354067X251340190, 2025.

[19] S. Gupta, "Ai agents collaboration under resource constraints: Practical implementations," *INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT*, vol. 3, no. 1, pp. 51–63, 2025.

[20] S. N. M. Pothukuchi, "Llmops: A comprehensive guide to deploying large language models in production," *IJSAT-International Journal on Science and Technology*, vol. 16, no. 1, 2025.

[21] H. Wang, Y. Li, S. Wang, G. Chen, and Y. Chen, "Milora: Harnessing minor singular components for parameter-efficient llm finetuning," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 4823–4836.

[22] H. Huang, X. Bu, H. Zhou, Y. Qu, J. Liu, M. Yang, B. Xu, and T. Zhao, "An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 5880–5895.

[23] E. V. Naderi, "Intelligent tutoring systems in the age of llm-based agentic frameworks-adapting small on-device language models for fact-checking and student compliance detection," 2025.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[25] D. A. Bull, "Impact of curriculum misalignment and assessment practices on student learning outcomes in higher education: A prisma-guided qualitative content synthesis," *International Journal of Interdisciplinary Research and Innovations*, vol. 13, no. 3, pp. 65–87, 2025.

[26] O. Gatete, "Revisiting tpack: A critical review and contextual extension for the digital age," *Journal of Educational Technology Systems*, p. 00472395251382942, 2025.

[27] A. Dehbi, A. Bakhouyi, A. Khaddar, and M. Talea, "Education and smart technologies: Towards a new pedagogical paradigm," *International Journal of Evaluation and Research in Education*, vol. 14, no. 1, pp. 297–309, 2025.

[28] R. Zhong and Y. Zhao, "Education paradigm shifts in the age of ai: A spatiotemporal analysis of learning," *ECNU Review of Education*, p. 20965311251315204, 2025.

[29] K. Misiejuk, S. López Pernas, E. A. Oliveira, J. Delannoy, C. Dujardin, H. Ahmed, and M. Saqr, "Facets of ai personalization: A systematic review of fine-tuned large language models for teaching and learning," *Available at SSRN 5287369*, 2025.

[30] T. Wang, Y. Zhan, J. Lian, Z. Hu, N. J. Yuan, Q. Zhang, X. Xie, and H. Xiong, "Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system," in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 510–519.