KHUNG KIẾN TRÚC HỖ TRỢ CHẤM TỰ LUẬN HƯỚNG MÔ HÌNH NGÔN NGỮ LỚN VÀ TRUY VẾT PHÊ DUYỆT

Nguyễn Sỹ Khánh[1], Bùi Thu Hải[2], Đặng Thị Thanh Hương[3], Lê Văn Vinh[1]
[1] Trung tâm Thông tin - Thư viện, ĐH Công nghệ Kỹ thuật Vinh
[2] Khoa Công nghệ thông tin, ĐH Sư phạm Kỹ thuật Nam Định
[3] Phòng Quản lý ĐT và ĐBCL, Trường CĐ Kinh tế - Kỹ thuật số 1 Nghệ An

Tóm tắt: *Chấm thi tự luận thủ công tốn thời gian và dễ thiếu nhất quán. Mô hình ngôn ngữ lớn (LLM) có thể hỗ trợ xây dựng bảng tiêu chí chấm điểm (rubric) và đề xuất đánh giá, nhưng cần kiểm soát căn cứ cho điểm và trách nhiệm phê duyệt. Bài báo đề xuất khung kiến trúc theo nguyên tắc "LLM đề xuất – giáo viên quyết định", kết hợp mô hình hóa UML, ràng buộc theo hướng OCL, kiểm tra tất định và truy vết phê duyệt. Thông qua tình huống minh họa và kiểm thử nguyên mẫu Python, kết quả bước đầu cho thấy khả năng phát hiện các vi phạm định trước và hỗ trợ quy trình kiểm tra, sửa đổi, rà soát và phê duyệt. Nghiên cứu cung cấp cơ sở đánh giá tính khả thi kỹ thuật của khung, làm tiền đề tích hợp LLM và kiểm chứng trong thực tiễn.*

Từ khóa: *Khung kiến trúc hướng mô hình ngôn ngữ lớn; Chấm thi tự luận; Ràng buộc nghiệp vụ; Truy vết phê duyệt.*

AN ARCHITECTURAL FRAMEWORK FOR LLM-BASED OPEN-ENDED GRADING SUPPORT WITH APPROVAL TRACEABILITY

Abstract: *Manual essay grading is time-consuming and prone to inconsistency. Large language models (LLMs) can support rubric development and grading proposals, but scoring evidence and approval accountability require explicit controls. This paper proposes an architectural framework based on the principle "LLMs propose, teachers decide," integrating UML modeling, OCL-style constraints, deterministic validation, and approval traceability. Through an illustrative case and Python prototype testing, preliminary results demonstrate the detection of predefined violations and support for validation, revision, review, and approval processes. The study provides initial evidence of the framework's technical feasibility, establishing a foundation for LLM integration and future evaluation in real-world grading contexts.*

Keywords: *LLM-based architectural framework; Open-ended examination grading; Business constraints; Approval traceability.*

Nhận bài: 22/09/2026 Phản biện: 25/10/2026 Duyệt đăng: 30/10/2026

I. MỞ ĐẦU

Chấm thi tự luận đòi hỏi đánh giá tính chính xác của nội dung, mức độ đáp ứng yêu cầu và chất lượng lập luận. Bảng tiêu chí chấm điểm (rubric) giúp chuẩn hóa đánh giá, nhưng độ tin cậy phụ thuộc vào thiết kế và sự thống nhất khi áp dụng (Jönsson & Svingby, 2007). Mô hình ngôn ngữ lớn (Large Language Models – LLM) mở ra khả năng hỗ trợ xây dựng rubric và đề xuất điểm số kèm giải trình (Frederick Eneye et al., 2025; Zhang et al., 2024). Tuy nhiên, mức độ thống nhất với người chấm và khả năng xử lý đáp án tương đương còn hạn chế; việc tăng chi tiết rubric cũng không mặc nhiên cải thiện độ chính xác (Yoshida, 2025; Mathew et al., 2026).

Các nghiên cứu đã đề xuất cơ chế tương tác chuyên gia trong GradeHITL (Chu et al., 2025), đánh giá dựa trên bằng chứng trong Rulers (Hong et al., 2026) và kiểm tra, sửa có giới hạn trong GradeAgentOps (Anghel et al., 2026). Tuy nhiên, việc liên kết nguồn tiêu chí, phiên bản rubric, bản đánh giá và thẩm quyền phê duyệt thành các nghĩa vụ kiểm tra xuyên suốt vẫn cần được làm rõ. Một đầu ra hợp lệ về cấu trúc và điểm số vẫn có thể sử dụng sai phiên bản rubric hoặc dựa trên báo cáo kiểm tra không còn hiệu lực sau chỉnh sửa.

Để giải quyết vấn đề này, nghiên cứu kết hợp ngôn ngữ mô hình hóa thống nhất (UML) và ngôn ngữ ràng buộc đối tượng (OCL) (Object Management Group, 2014, 2017), đề xuất khung kiến trúc theo nguyên tắc "LLM đề xuất – giáo viên quyết định". Khung gồm hai giai đoạn: xây dựng, kiểm duyệt rubric và đánh giá bài làm theo rubric đã phê duyệt, kết hợp kiểm tra ràng buộc, sửa có giới hạn và truy vết phiên bản.

Nghiên cứu vận dụng phương pháp khoa học thiết kế (Hevner et al., 2004), kết hợp đặc tả nghiệp vụ và kiểm thử nguyên mẫu Python nhằm trả lời ba câu hỏi: (RQ1) tổ chức vòng đời rubric và quy trình chấm để bảo đảm truy vết, phê duyệt như thế nào; (RQ2) biểu diễn các ràng buộc về cấu trúc, điểm số, bằng chứng và liên kết dữ liệu ra sao; (RQ3) nguyên mẫu phát hiện sai lệch và hỗ trợ luồng xử lý chấm như thế nào?

Bài báo đóng góp: (i) khung kiến trúc hai giai đoạn có truy vết phê duyệt; (ii) mô hình miền và hệ thống quy tắc kiểm tra nghiệp vụ; (iii) nguyên

mẫu Python và các ca kiểm thử minh họa tính khả thi kỹ thuật của khung, chưa nhằm đánh giá độ chính xác chấm điểm của LLM.

II. NHỮNG NGHIÊN CỨU LIÊN QUAN

2.1. Rubric và chấm tự luận bằng mô hình ngôn ngữ lớn

Rubric chuẩn hóa tiêu chí, mức đạt và phân bổ điểm, nhưng hiệu quả phụ thuộc vào thiết kế và sự thống nhất khi áp dụng (Jönsson & Svingby, 2007). Chuyển đáp án thành rubric cần làm rõ từng mức đạt, chấp nhận diễn đạt tương đương và tránh chấm trùng.

LLM mở rộng chấm tự động sang đánh giá theo rubric và giải trình điểm đề xuất (Frederick Eneye et al., 2025). Tuy nhiên, độ thống nhất với người chấm và tính ổn định còn hạn chế (Mathew et al., 2026; Zhang et al., 2024); tăng chi tiết rubric cũng không mặc nhiên cải thiện độ chính xác (Yoshida, 2025). Vì vậy, rubric và kết quả do LLM đề xuất đều cần được kiểm duyệt.

2.2. Sự tham gia của con người và cơ chế kiểm soát đầu ra LLM

Các hướng đưa con người vào vòng lặp (Human-in-the-Loop – HITL) gồm GradeHITL, sử dụng phản hồi chuyên gia để điều chỉnh rubric (Chu et al., 2025); Rulers, đánh giá neo theo bằng chứng (Hong et al., 2026); và GradeAgentOps, kiểm tra tất định, chuẩn hóa và sửa có giới hạn (Anghel et al., 2026).

Các cơ chế trên cho thấy đầu ra LLM cần kiểm tra trước khi sử dụng. Tuy nhiên, việc liên kết nguồn tiêu chí, thẩm quyền kiểm duyệt, phiên bản rubric đã khóa và tái kiểm tra sau điều chỉnh điểm vẫn cần được đặc tả thống nhất. Đây là trọng tâm của khung đề xuất.

2.3. Mô hình hóa và đặc tả ràng buộc nghiệp vụ

UML hỗ trợ mô hình hóa cấu trúc miền và quy trình nghiệp vụ (Object Management Group, 2017), trong khi OCL cho phép đặc tả bất biến, tiền điều kiện và hậu điều kiện trên mô hình đối tượng (Object Management Group, 2014). Trong hỗ trợ chấm tự luận, UML/OCL có thể biểu diễn các ràng buộc về tổng điểm, bước điểm, phiên bản rubric và thẩm quyền phê duyệt. Nghiên cứu vận dụng các điều kiện kiểm tra logic, minh họa bằng biểu thức OCL rút gọn và hiện thực hóa một phần bằng các hàm Python để kiểm thử nguyên mẫu.

III. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Cách tiếp cận và quy trình nghiên cứu

Nghiên cứu vận dụng khoa học thiết kế (Design Science Research – DSR) (Hevner et al., 2004) để xây dựng khung theo nguyên tắc "LLM đề xuất – giáo viên quyết định". Hai giai đoạn liên kết xây dựng, kiểm duyệt rubric với đề xuất, kiểm tra và phê duyệt điểm số. LLM hỗ trợ nội dung; bộ điều phối kiểm tra ràng buộc; giáo viên thẩm định và quyết định cuối cùng.

Quy trình nghiên cứu gồm năm bước: (1) phân tích đề thi, đáp án, hướng dẫn chấm để xác định yêu cầu và quy tắc nghiệp vụ; (2) xây dựng rubric và đặc tả cấu trúc, quan hệ, ràng buộc bằng UML/OCL; (3) thiết kế kiến trúc và bộ điều phối sinh–kiểm–sửa, quản lý phiên bản và thẩm quyền phê duyệt; (4) phát triển nguyên mẫu Python để kiểm tra ràng buộc và ghi nhận lịch sử xử lý; (5) kiểm thử bằng các ca hợp lệ, ca cấy lỗi và tình huống minh họa nhằm đối chiếu hành vi xử lý với kết quả mong đợi.

3.2. Đối tượng, dữ liệu và phương tiện nghiên cứu

Nghiên cứu sử dụng đề tuyển sinh lớp 10 môn Ngữ văn tỉnh Ninh Bình năm 2025 và lời giải tham khảo trên website giáo dục, không phải hướng dẫn chấm chính thức (Ban chuyên môn Loigiaihay.com, 2025). Đề gồm bốn câu Đọc hiểu (4,0 điểm) và hai câu Viết (6,0 điểm), với 24 tiêu chí trong rubric minh họa.

Nguyên mẫu được triển khai bằng Python 3, sử dụng thang điểm 10,0 với bước điểm q=0,25, tương ứng 40 đơn vị điểm. Bộ dữ liệu gồm 30 bài làm tổng hợp (SYN-001 đến SYN-030), chia đều thành ba nhóm: 10 bài đạt yêu cầu, 10 bài đạt một phần và 10 bài sai hoặc thiếu trọng tâm. Một số bài được bổ sung chỉ dẫn ngược yêu cầu hoặc cách diễn đạt tương đương để khảo sát hành vi xử lý.

Các quy tắc được kiểm thử bằng ca hợp lệ và ca cấy lỗi có chủ đích. Việc đánh giá tập trung vào khả năng phát hiện vi phạm, kiểm soát quy trình và truy vết, chưa xác định độ chính xác chấm điểm của LLM.

IV. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

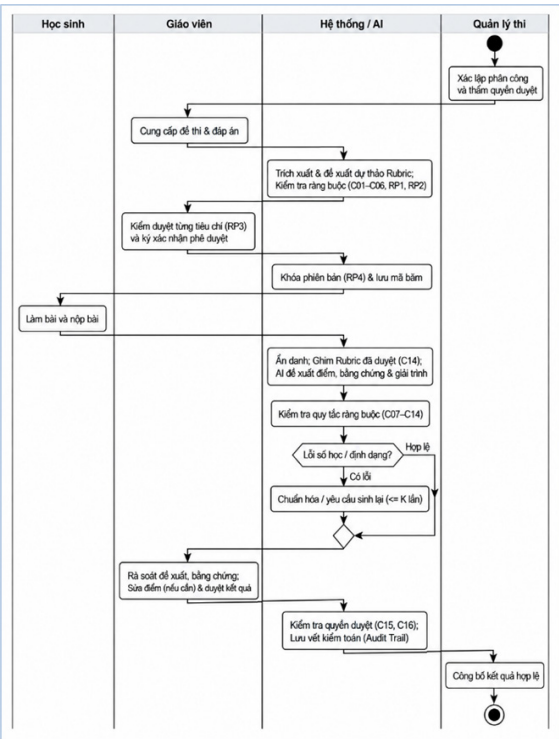
4.1. Kiến trúc tổng thể của khung

Khung gồm hai giai đoạn: xây dựng và kiểm duyệt rubric; đánh giá bài làm theo phiên bản đã duyệt. Nguyên tắc "LLM đề xuất – giáo viên quyết định" được thực thi qua kiểm tra ràng buộc và truy vết phê duyệt.

LLM đề xuất tiêu chí, mức đạt; bộ điều phối kiểm tra quy tắc; giáo viên rà soát chuyên môn và phê duyệt. Bộ phận quản lý thi phân công thẩm quyền, kiểm soát công bố. Kiểm tra tự động không thay thế quyết định chuyên môn.

Sáu thành phần gồm quản lý dữ liệu và phiên bản, dịch vụ LLM, bộ điều phối, bộ kiểm tra, giao diện kiểm duyệt và kho truy vết. Chúng phối hợp bảo toàn dữ liệu, xử lý vi phạm và lưu quyết định phê duyệt.

Giai đoạn A tạo rubric từ đề, đáp án và hướng dẫn chấm, sau đó kiểm tra và chuyển giáo viên phê duyệt. Giai đoạn B đề xuất điểm và bằng chứng theo rubric đã ghim. Định danh và phiên bản do ngữ cảnh tin cậy cấp; lỗi kỹ thuật hoặc bài thiếu dữ liệu không tự động thành điểm 0. Luồng xử lý thể hiện trong Hình 1.



Hình 1. Kiến trúc và luồng xử lý tổng thể của khung

4.2. Xây dựng và kiểm duyệt rubric

LLM đề xuất yêu cầu đánh giá, tiêu chí, mức đạt, điểm và bằng chứng từ đề, đáp án. Yêu cầu phải có trích đoạn nguồn; tiêu chí liên kết với yêu cầu. Tiêu chí do giáo viên bổ sung phải lưu danh tính, lý do và căn cứ.

Giáo viên xác nhận danh mục yêu cầu trước khi kiểm tra độ bao phủ, tránh bỏ sót nội dung mà hệ thống vẫn kết luận hợp lệ.

Mỗi tiêu chí có mã duy nhất, mô tả, điểm tối đa, các mức đạt và yêu cầu bằng chứng; phải phân biệt không đạt, đạt một phần và đầy đủ. Quan hệ phụ thuộc hoặc trần điểm cần có căn cứ chuyên môn.

Biểu diễn điểm số: Điểm số được biểu diễn bằng các đơn vị nguyên $u_i \in \mathbb{Z} \geq 0$ theo độ phân giải bước điểm q lấy từ quy tắc chấm:

$$s_i = q * u_i,\ u_i \in \mathbb{Z} \geq 0$$

Với n tiêu chí, điểm câu hỏi được tính bằng:

$$S = q \sum_{i=1}^{n} u_i$$

Ví dụ, khi q=0,25, mức 1.5 điểm tương ứng 6 đơn vị. Trong mô hình cộng điểm cơ bản, tổng điểm tối đa của các tiêu chí bằng điểm tối đa câu hỏi. Các giá trị không phù hợp với độ phân giải bước điểm q đều bị hệ thống phát hiện và báo vi phạm, tuyệt đối không làm tròn âm thầm.

Hệ thống kiểm tra cấu trúc, miền điểm, mức đạt, phụ thuộc, nguồn và liên kết tiêu chí. Kiểm tra trích đoạn chỉ xác nhận sự tồn tại, không chứng minh tính phù hợp. Giáo viên rà soát nội dung, đáp án tương đương và nguy cơ chấm trùng.

Giáo viên chấp nhận, sửa, bỏ hoặc thêm tiêu chí kèm lý do. Mọi chỉnh sửa nội dung, điểm hoặc liên kết tạo bản dự thảo mới và phải kiểm tra lại.

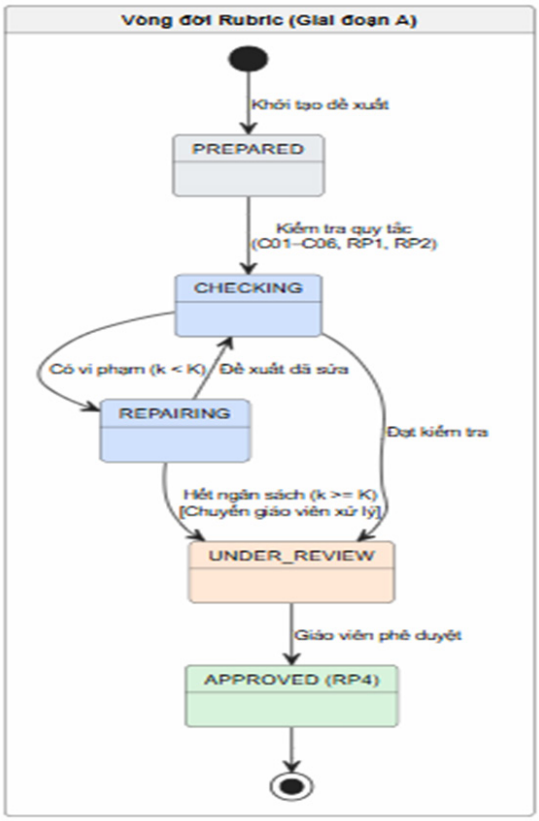
Rubric chỉ được duyệt khi đạt kiểm tra kỹ thuật, thẩm định chuyên môn và xác nhận đúng thẩm quyền. Bản duyệt được khóa kèm mã băm; thay đổi sau đó tạo phiên bản mới, không tự thay rubric đã ghim.

4.3. Mô hình miền và trạng thái xử lý

Mô hình UML biểu diễn các đối tượng, quan hệ và trách nhiệm nghiệp vụ (OMG, 2017). Bảng 2 trình bày cấu trúc các nhóm lớp chính của mô hình miền.

Mỗi nhiệm vụ chấm ghim một phiên bản rubric đã duyệt. Bản đánh giá gắn với nhiệm vụ, có đúng một mục cho mỗi tiêu chí áp dụng; mức đạt thuộc đúng tiêu chí, bằng chứng thuộc đúng câu trả lời và phiên bản văn bản. Giáo viên điều chỉnh điểm sẽ tạo bản đánh giá mới, ghi nhận người sửa và liên kết với bản trước.

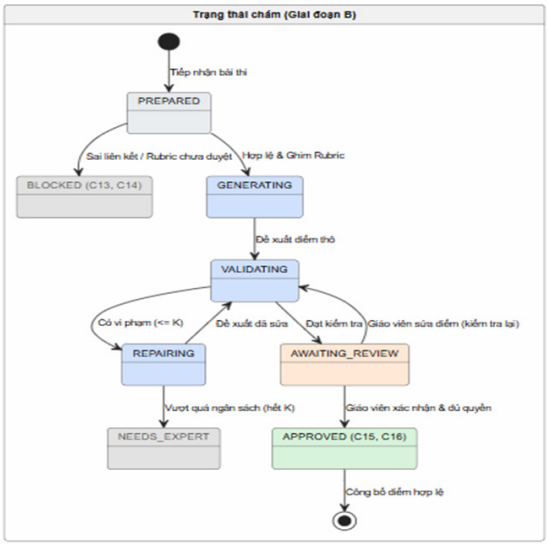
Vòng đời rubric: Gồm các trạng thái chuẩn bị (PREPARED), kiểm tra (CHECKING), sửa (REPAIRING), chờ kiểm duyệt (UNDER_REVIEW) và được phê duyệt (APPROVED). Khi hết ngân sách sinh lại (k >=K) mà vẫn còn vi phạm, đề xuất được chuyển trực tiếp cho giáo viên xử lý cùng báo cáo lỗi; rubric tuyệt đối không được phê duyệt cho đến khi mọi điều kiện bắt buộc được đáp ứng, được biểu diễn trong Hình 2.



Hình 2. Vòng đời rubric

Nhiệm vụ chấm: Gồm các trạng thái chuẩn bị (PREPARED), sinh đề xuất (GENERATING), kiểm tra (VALIDATING), sửa (REPAIRING), chờ rà soát (AWAITING_REVIEW), cần chuyên gia (NEEDS_EXPERT), bị chặn (BLOCKED) và được phê duyệt (APPROVED).

Đề xuất đạt kiểm tra được chuyển sang rà soát. Nếu giáo viên điều chỉnh, bản mới quay lại bước kiểm tra; chỉ bản cuối đạt toàn bộ điều kiện và đúng thẩm quyền mới được duyệt, được biểu diễn trong Hình 3.



Hình 3. Trạng thái xử lý chấm

4.4. Đặc tả và kiểm tra ràng buộc nghiệp vụ bằng OCL

OCL hỗ trợ biểu diễn các bất biến, tiền điều kiện và hậu điều kiện trên mô hình đối tượng, nhưng không trực tiếp điều phối xử lý hoặc sửa đổi dữ liệu (Object Management Group, 2014). Nghiên cứu xác định 20 quy tắc nghiệp vụ (Bảng 1), chia thành bốn nhóm: cấu trúc rubric (C01–C06); tính hợp lệ của bản đánh giá (C07–C12); liên kết dữ liệu, phiên bản và thẩm quyền (C13–C16); nguồn gốc đề xuất và kiểm duyệt rubric (RP1–RP4).

Bảng 1. Hai mươi quy tắc ràng buộc nghiệp vụ và kiểm thử tương ứng

Mã	Điều kiện kiểm tra	Xử lý	Ca thử (H/L)
C01	Mã tiêu chí duy nhất trong rubric	B	1/1
C02	Tổng điểm tối đa tiêu chí bằng điểm tối đa câu hỏi	B	1/1
C03	Điểm tối đa và bước điểm hợp lệ theo độ phân giải q	B	1/1
C04	Mức đạt có mã duy nhất, điểm đúng miền và đúng bước điểm	B	1/1
C05	Có mức 0 và mức đạt điểm tối đa	B	1/1
C06	Quan hệ phụ thuộc hợp lệ, không có chu trình	B	1/1
C07	Đánh giá đủ tiêu chí, không thiếu hoặc trùng lặp	S	1/1

C08	Mức đạt thuộc tiêu chí, điểm khớp mức chọn	T	1/1
C09	Tổng điểm bằng tổng số học các điểm thành phần	T	1/1
C10	Tuân thủ trần điểm của quan hệ phụ thuộc	S	1/1
C11	Bằng chứng đúng bài, phiên bản và vị trí trích đoạn	S	1/1
C12	Mức yêu cầu bằng chứng bắt buộc có bằng chứng hợp cấu trúc	S	1/1
C13	Liên kết đúng giữa câu trả lời, bài làm, câu hỏi và đề thi	B	1/1
C14	Sử dụng đúng rubric đã ghim và được phê duyệt	B	1/1
C15	Người duyệt điểm có phân công và thẩm quyền	B	1/1
C16	Phê duyệt gắn bản cuối với báo cáo kiểm tra đạt	B	1/1
RP1	Yêu cầu và tiêu chí LLM đề xuất có trích dẫn thật không rỗng từ đề/đáp án	S	1/1
RP2	Rubric phủ hết yêu cầu của đề; mọi tiêu chí gắn yêu cầu	S	1/1
RP3	Giáo viên quyết định từng tiêu chí; sửa/bỏ/thêm có lý do	B	1/1
RP4	Duyệt rubric đúng thẩm quyền, đạt kiểm tra; nội dung không đổi sau duyệt	B	1/1

Chú thích: "Ca thử (H/L)" là số ca hợp lệ (H) / số ca cấy lỗi (L) của mỗi quy tắc. Mức xử lý: T – chuẩn hóa tất định; S – yêu cầu LLM sinh lại; B – chặn và chuyển giáo viên.

Ví dụ đặc tả bằng OCL minh họa cho ba quy tắc ở các đối tượng khác nhau, trong đó RP1 thuộc giai đoạn A:

```
-- RP1: tiêu chí phải có nguồn văn bản không rỗng hoặc do giáo viên bổ sung hợp lệ
context RubricProposal
inv RP1_SourceGrounded:
  self.criteria->forAll(c |
    (c.origin = Origin::TEACHER and c.justification.size() > 0) or
    (c.source.quote.size() > 0 and
     self.referenceAnswer.text.indexOf(c.source.quote) > 0))
-- C08: mức đạt thuộc tiêu chí và điểm thành phần khớp mức chọn
context CriterionAssessment
inv C08_LevelScoreConsistent:
  self.criterion.levels->includes(self.level) and
  self.units = self.level.units
-- C10: trần điểm theo quan hệ phụ thuộc giữa các tiêu chí
context Assessment
inv C10_DependencyCap:
  self.task.pinnedRubric.dependencies->forAll(d |
    self.unitsOf(d.fromCriterion) < d.threshold implies
    self.unitsOf(d.toCriterion) <= d.capUnits)
```

Các biểu thức OCL trên minh họa một phần, chưa hoàn chỉnh và chưa được thực thi trên engine OCL: RP1 chưa kiểm tra trích dẫn của yêu cầu đề; C10 dùng hàm unitsOf chưa được định nghĩa; indexOf chưa xử lý null và nhiều vị trí khớp.

Nguồn và bằng chứng văn bản: RP1 kiểm tra căn cứ nguồn của tiêu chí rubric; C11 xác nhận trích đoạn bằng chứng thuộc đúng câu trả lời và phiên bản văn bản. Trong nguyên mẫu Python, vị trí trích đoạn được biểu diễn bằng khoảng nửa mở [start, end), tính từ 0 trên chuỗi chuẩn hóa và chuyển đổi chỉ số khi đối chiếu OCL. Các trường hợp không tìm thấy hoặc nhiều vị trí khớp phải được xử lý trước khi chấp nhận. Việc kiểm tra chỉ xác nhận sự tồn tại và liên kết của bằng chứng, không chứng minh tính phù hợp của nội dung với mức điểm.

Liên kết dữ liệu, phiên bản và thẩm quyền: C13 kiểm tra liên kết giữa bài làm, câu trả lời, câu hỏi và đề thi; C14 xác nhận bản đánh giá sử dụng đúng phiên bản rubric đã ghim và phê duyệt theo RP4. C15 và RP4 kiểm tra thẩm quyền phê duyệt; C16 bảo đảm quyết định phê duyệt gắn với bản đánh giá cuối cùng và báo cáo kiểm tra hợp lệ. Mọi chỉnh sửa đều tạo phiên bản mới và phải được kiểm tra lại để bảo đảm truy vết.

Phân loại vi phạm và truy vết: Hệ thống kiểm tra cấu trúc, kiểu dữ liệu và điều kiện nghiệp vụ, kết hợp rà soát chuyên môn của giáo viên. Vi phạm được phân loại theo dữ liệu, số học, nguồn, phiên bản, bằng chứng và phê duyệt. Báo cáo ghi nhận mã quy tắc, đối tượng, vị trí, giá trị mong đợi và quan sát, mức nghiêm trọng. Chuỗi truy vết liên kết nguồn nghiệp vụ → yêu cầu → biểu thức → thực thi → ca kiểm thử → vi phạm, làm cơ sở tự động hóa ma trận truy vết.

4.5. Mô hình ngôn ngữ lớn đề xuất đánh giá theo rubric

LLM tiếp nhận câu hỏi, đáp án, hướng dẫn chấm, phiên bản rubric đã ghim và câu trả lời để đề xuất mức đạt, điểm số, bằng chứng và giải trình cho từng tiêu chí. Bài làm được xem là dữ liệu không tin cậy; mọi chỉ dẫn thay đổi quy tắc chấm trong bài không được thực thi.

Đầu ra gồm mã tiêu chí, mức đạt, điểm đề xuất, trích đoạn bằng chứng, giải trình và cảnh báo. Định danh nhiệm vụ, phiên bản rubric và bài làm được bổ sung từ ngữ cảnh tin cậy. Khi thiếu căn cứ hoặc gặp đáp án ngoài hướng dẫn, mô hình phải đánh dấu để giáo viên xem xét. Kiểm tra tự động xác nhận tính hợp lệ của dữ liệu và liên kết bằng chứng; giáo viên vẫn chịu trách nhiệm đánh giá nội dung, lập luận và các cách trả lời tương đương.

4.6. Điều phối sinh – kiểm – sửa, phê duyệt và truy vết

Bộ điều phối thực hiện quy trình sinh–kiểm–sửa với giới hạn K lượt sửa sau lần sinh đầu tiên, gắn với phiên bản rubric đã phê duyệt . Trước khi gọi LLM, hệ thống kiểm tra liên kết dữ liệu và phiên bản rubric; các vi phạm C13–C14 bị chặn. Đầu ra được kiểm tra cấu trúc, mức đạt, bằng chứng và các ràng buộc C07–C12. Hệ thống chỉ chuẩn hóa điểm khi mức đạt hợp lệ và phải kiểm tra lại sau chuẩn hóa. Các lỗi cần sửa được xử lý trong giới hạn \(K\); khi vượt giới hạn hoặc cần thẩm định chuyên môn, đề xuất chuyển sang NEEDS_EXPERT kèm báo cáo vi phạm. Bài làm và rubric đã ghim không bị thay đổi trong quá trình sửa.

Đề xuất hợp lệ được chuyển giáo viên rà soát. Mọi điều chỉnh tạo phiên bản đánh giá mới, ghi nhận người thực hiện, thời điểm và lý do, đồng thời phải kiểm tra lại trước khi phê duyệt theo C15–C16. Cơ chế truy vết liên kết nhiệm vụ chấm, phiên bản rubric, lịch sử đề xuất, sửa đổi, báo cáo kiểm tra và quyết định phê duyệt, bảo đảm phân biệt kết quả kiểm tra kỹ thuật với quyết định chuyên môn của giáo viên.

4.7. Kết quả kiểm thử nguyên mẫu

Kết quả lưu trong tệp results.json cho thấy 20/20 ca hợp lệ và 20/20 ca cấy lỗi đạt kết quả mong đợi (40/40 ca kiểm thử). Theo nhóm quy tắc, kết quả lần lượt là 12/12 (C01–C06), 12/12 (C07–C12), 8/8 (C13–C16) và 8/8 (RP1–RP4). Trong đó, sáu ca thuộc C06, C10 và RP4 mới được kiểm tra thông qua mô phỏng logic.

Các lỗi được phát hiện gồm trùng mã tiêu chí (C01), sai tổng điểm rubric (C02), thiếu tiêu chí đánh giá (C07), sai tổng điểm (C09), bằng chứng không tồn tại (C11), sai phiên bản rubric (C14) và phê duyệt không đúng thẩm quyền (C15). Kết quả chứng minh khả năng phát hiện các sai lệch đã thiết kế và vận hành luồng xử lý trong môi trường kiểm soát, chưa phản ánh độ chính xác chấm điểm của LLM.

4.8. Minh họa biểu diễn rubric và tính điểm

Tình huống minh họa sử dụng câu Đọc hiểu 3 (1,0 điểm), yêu cầu làm rõ tác dụng của biện pháp tu từ điệp từ trong đoạn văn miêu tả nhân vật Tân. Từ lời giải tham khảo (Ban chuyên môn Loigiaihay.com, 2025), nhóm nghiên cứu xây dựng rubric gồm ba nội dung: tạo nhịp điệu (0,25 điểm); nhấn mạnh cảm giác say sưa, ngây ngất trước hương lúa đồng quê (0,50 điểm); thể hiện sự gắn bó, yêu mến quê hương (0,25 điểm).

Bảng 2. Rubric minh họa cho câu Đọc hiểu 3 (xây dựng từ lời giải tham khảo, chưa được giáo viên phê duyệt)

Tiêu chí	Căn cứ nội dung từ lời giải	Thang điểm	Đơn vị (q = 0,25)
DH3_a	Tạo nhịp điệu cho câu văn	0,25	1
DH3_b	Nhấn mạnh cảm giác say sưa, ngây ngất	0,50	2
DH3_c	Thể hiện sự gắn bó, yêu mến thiên nhiên làng quê	0,25	1
Tổng	Rubric nghiên cứu	1,00	4

Với câu trả lời dựng sẵn: "Mùi hương lúa chín làm Tân say mê; điệp từ tạo nhịp điệu và nhấn mạnh niềm vui, sự gắn bó với quê", bộ đối sánh từ khóa của nguyên mẫu gán mức tối đa 1,0 điểm. Đây là kết quả đối sánh từ khóa, không phải đánh giá ngữ nghĩa của LLM hoặc điểm chấm thực tế.

Nguyên mẫu cũng xử lý toàn bộ 30 bài làm tổng hợp và xuất bảng điểm CSV kèm trích đoạn bằng chứng. Các kết quả chưa được đối chiếu với đánh giá của giáo viên.

4.9. Thảo luận

Khung kiến trúc đề xuất cơ chế kiểm soát chấm thi tự luận thông qua liên kết nguồn tiêu chí, quản lý phiên bản rubric, kiểm tra bằng chứng và ràng buộc thẩm quyền phê duyệt. Kết quả kiểm thử phát hiện 20/20 lỗi cấy chủ đích và xử lý 30 bài làm tổng hợp, cung cấp bằng chứng bước đầu về khả năng hiện thực hóa các điều kiện nghiệp vụ và luồng xử lý trong môi trường kiểm soát.

Đóng góp của khung nằm ở việc liên kết các cơ chế kiểm tra thành chuỗi nghĩa vụ xuyên suốt, bảo đảm bản đánh giá cuối cùng được kiểm tra theo đúng phiên bản trước khi phê duyệt. Tuy nhiên, kết quả mới phản ánh tính khả thi kỹ thuật trong phạm vi nguyên mẫu đơn giản hóa. Chất lượng chấm vẫn phụ thuộc vào rubric, đề xuất của LLM và thẩm định chuyên môn của giáo viên; nghiên cứu chưa đánh giá mức độ thống nhất điểm số với người chấm hoặc hiệu quả vận hành thực tế.

Nghiên cứu tiếp theo cần hoàn thiện quy tắc nghiệp vụ, khảo sát giáo viên, mở rộng xử lý đáp án tương đương và tích hợp LLM với tài liệu chấm đã phê duyệt. Việc đánh giá cần sử dụng dữ liệu thi thực tế có thẩm định, đối chiếu kết quả với hội đồng giáo viên độc lập và đo lường chất lượng đề xuất, mức độ đồng thuận cũng như khối lượng rà soát để xác định điều kiện triển khai phù hợp.

V. KẾT LUẬN

Bài báo đề xuất khung kiến trúc hỗ trợ chấm thi tự luận hướng mô hình ngôn ngữ lớn theo nguyên tắc "LLM đề xuất – giáo viên quyết định", kết hợp quản lý bảng tiêu chí chấm điểm, kiểm tra ràng buộc nghiệp vụ và truy vết phê duyệt. Kết quả kiểm thử nguyên mẫu bước đầu cho thấy khả năng hiện thực hóa các cơ chế kiểm tra, đồng thời phân định vai trò của kiểm tra tự động và thẩm định chuyên môn của giáo viên. Nghiên cứu tiếp theo sẽ hoàn thiện bộ kiểm tra, tích hợp LLM và đánh giá trên dữ liệu thi thực tế với hội đồng giáo viên độc lập.

TÀI LIỆU THAM KHẢO

Anghel, C., et al. (2026). GradeAgentOps: A verification-first framework for evidence-anchored LLM exam grading. AI, 7(6), Article 198. https://doi.org/10.3390/ai7060198

Ban chuyên môn Loigiaihay.com. (2025). Đề thi chính thức vào 10 môn Ngữ văn – Ninh Bình năm 2025 và hướng dẫn giải chi tiết. https://loigiaihay.com/de-thi-vao-10-mon-van-ninh-binh-nam-2025-a185696.html

Chu, Y., Li, H., Yang, K., Copur-Gencturk, Y., & Tang, J. (2025). LLM-based automated grading with human-in-the-loop [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.05239

Frederick Eneye, T. A. N., Ijezue, C. F., Amjad, A. I., Amjad, M., Butt, S., & Castañeda-Garza, G. (2025). Advances in auto-grading with large language models: A cross-disciplinary survey. In Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025) (pp. 477–498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.bea-1.35

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Hong, Y., Yao, H., Shen, B., Xu, W., Wei, H., & Dong, Y. (2026). From rubrics to reliable scores: Evidence-grounded text evaluation with LLM judges (Version 3) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2601.08654

Jönsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Mathew, J. G., Taher, S., Kundu, A., & Barbosa, D. (2026). LLMs do not grade essays like humans [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.23714

Object Management Group. (2014). Object Constraint Language (Version 2.4). https://www.omg.org/spec/OCL/2.4

Object Management Group. (2017). OMG Unified Modeling Language (OMG UML) (Version 2.5.1). https://www.omg.org/spec/UML/2.5.1

Yoshida, L. (2025). Do we need a detailed rubric for automated essay scoring using large language models? [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.01035

Zhang, D.-W., Boey, M., Tan, Y. Y., & Jia, A. H. S. (2024). Evaluating large language models for criterion-based grading from agreement to consistency. npj Science of Learning, 9, Article 79. https://doi.org/10.1038/s41539-024-00291-1