Speeding Up Thesis Feedback with a Custom GPT Reviewer: A Mixed-Methods Evaluation in Construction Engineering

Ítalo Sepúlveda
Faculty of Architecture, Construction and Environment, Universidad Autónoma de Chile, Chile
Thomas Mandel
Faculty of Architecture, Construction and Environment, Universidad Autónoma de Chile, Chile
Yaely Barrios
Faculty of Architecture, Construction and Environment, Universidad Autónoma de Chile, Chile
Gabriela Peterssen
Faculty of Architecture, Construction and Environment, Universidad Autónoma de Chile, Chile
Jesús Ortega
Faculty of Architecture, Construction and Environment, Universidad Autónoma de Chile, Chile
italo.sepulveda@uautonoma.cl

Abstract

Delayed supervisory feedback is a persistent bottleneck in capstone theses. We report a semester-long pilot of a program-specific GPT configured as a rubric-bound, first-pass reviewer for Construction Engineering theses. The system generates instant, section-level comments aligned

to program rubrics, while instructors retain final disciplinary judgment. We evaluated the pilot across three touchpoints—student drafting, advisor support, and pre-defense committee preparation—using a mixed-methods design: two 9-item, 5-point Likert surveys (students n = 7; instructors n = 7), descriptive usage observations, an artifact audit, and analysis of open-ended responses. Students reported consistently high agreement on clarity, rubric alignment, usefulness, timeliness, and intention to reuse, with modest neutrality on perceived fairness and GPT–instructor coherence. Instructors endorsed the report's diagnostic precision, perceived time savings, increased cross-instructor consistency, and usefulness for defense-readiness decisions; neutrality persisted for fairness and time savings. Open-ended comments emphasized speed and pinpointing of weaknesses, and highlighted governance needs (pedagogical tone, bounded context, human oversight). Overall, the pilot suggests that a rubric-bound GPT can shift feedback earlier, standardize the first pass, and turn waiting time into learning time—without replacing human expertise. We outline guardrails (human-in-the-loop, transparency, academic integrity) and propose scaling to larger cohorts to test reliability and learning outcomes.

Keywords: Generative Artificial Intelligence; Thesis Supervision; Construction Engineering.

1. Introduction

Delays in supervisory feedback are a persistent friction point in capstone theses within professional degrees such as Construction Engineering. Because capstone courses compress problem framing, method selection, and academic writing into a short timeframe, the turnaround of formative comments becomes a bottleneck for progress, often forcing students to wait days or weeks between iterations. This "waiting cost" undermines visibility into standards and reduces opportunities for deliberate practice. Programs also face large cohorts and uneven availability of instructors, which complicate consistency and fairness of feedback across sections.

A robust body of research shows that feedback is one of the most powerful influences on learning when it is timely, specific, and usable. Foundational syntheses and meta-analyses converge on three principles: comments that are elaborated (what/why/how to improve), aligned to clear goals or rubrics, and delivered quickly

have larger effects on performance than generic praise or correctness signals [1–4]. In computer-based environments, elaborated feedback again outperforms simple verification, particularly for higher-order outcomes [4]. Yet effectiveness also depends on students' capacity to make sense of and act upon comments—so-called feedback literacy—which is strengthened when criteria are transparent, dialogue is encouraged, and revision plans are explicit [5,6].

Against this backdrop, large language models (LLMs) have emerged as flexible tools for generating formative critique, exemplars, and rubric-aligned checklists. Reviews of AI in higher education suggest that LLMs can increase productivity and provide immediate scaffolds, but they also foreground risks—accuracy, bias, authorship, and over-reliance—and consistently recommend human-in-the-loop governance, transparency, and strong alignment with learning outcomes [7,8]. Evidence specific to thesis supervision and capstone settings remains scarce: most reports describe general writing support or course-level assistants rather than domain-specific, rubric-bound reviewers evaluated across an authentic semester.

This study addresses that gap. We designed a custom GPT ("Thesis Reviewer – Construction Engineering") that operationalizes program outcomes and the official thesis format into explicit criteria, generates section-by-section, actionable feedback, and positions instructors as final arbiters of disciplinary judgment. Over one semester we deployed the system at three touchpoints—student drafting, advisor support, and pre-defense committee preparation—and gathered perceptions from students and instructors alongside artifact audits.

Building on prior work that formalized a GenAI-based methodology for data analysis in industrialized construction—emphasizing prompt design, contextualization, and human-in-the-loop governance—we adapt those principles to the thesis-feedback setting in Construction Engineering [9].

We ask: (RQ1) How do students and instructors perceive the clarity, rubric alignment, fairness, and overall usefulness of GPT feedback? (RQ2) To what extent does the GPT accelerate feedback cycles and enable meaningful pre-instructor revision? (RQ3) Does the GPT contribute to cross-instructor consistency and coherence with

human feedback? (RQ4) What risks and governance needs emerge when embedding a rubric-bound GPT into thesis supervision?

By reframing waiting time as learning time through instant, elaborated, rubric-anchored comments, our contribution is twofold: (i) a replicable blueprint for configuring LLM-based first-pass review in capstone theses, and (ii) empirical evidence from a semester-long deployment in Construction Engineering, with implications for scalable, defensible feedback practices.

2. Methodology

We conducted a semester-long pilot with a mixed-methods design, as shown in Figure 1, combining (a) perception surveys for students and instructors, (b) descriptive usage observations of how and when the custom GPT was employed, and (c) an audit of thesis artifacts focusing on structural and formatting compliance. The study evaluates a rubric-bound GPT configured to deliver first-pass, section-level feedback while preserving the instructor's role for summative judgment.

The intervention ran across capstone courses in Construction Engineering that culminate in a written thesis (Thesis I and Thesis II). The GPT reviewer was made available to all enrolled students and to faculty involved in supervision and committee activities. For the pilot evaluation we collected two independent perception samples: students ($n = 7$) who used the tool during early drafting or final adjustments, and instructors ($n = 7$) who used the reports to support supervision or pre-defense preparation.

The custom GPT ("Thesis Reviewer – Construction Engineering") operationalizes the program's thesis format and learning outcomes into explicit criteria. The configuration instructs the model to: (i) check thesis structure and format (required sections; pagination; margins; APA author–date citations and reference list); (ii) assess coherence among problem, objectives, methods, and results; (iii) provide section-by-section observations (intro/problem/objectives/justification/background/methods/results/conclusions/references/appendices); (iv) generate actionable recommendations (what to fix, why, how); and (v) issue a balanced verdict (approve with minor fixes vs. needs revision) without grading.

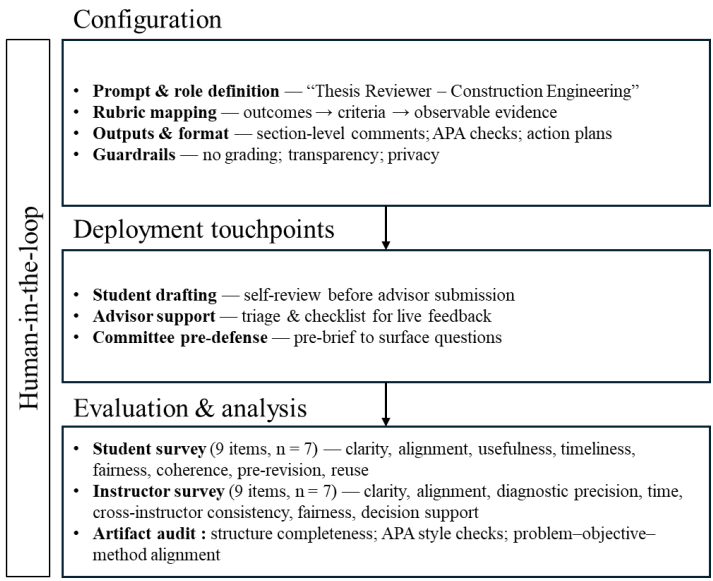
Workflow. Students used the GPT for rapid self-review before submitting drafts to instructors. Faculty could request a GPT report to triage issues and structure live feedback. Examination committees optionally consulted the report as a pre-brief to surface salient points for defense questions. At all times, human-in-the-loop governance applied: the GPT did not replace instructor evaluation.

We administered parallel 9-item, 5-point Likert questionnaires (1 = strongly disagree; 5 = strongly agree).

Student survey constructs: clarity of comments, rubric alignment, perceived fairness, overall usefulness, timeliness, relevance for improvement, ability to revise before instructor review, coherence with instructor feedback, and intention to reuse.

Instructor survey constructs: clarity, rubric alignment, fairness, overall usefulness, diagnostic precision (strengths/weaknesses), support for shaping live feedback, time optimization, cross-instructor consistency, and decision support for defense readiness. Open-ended prompts invited brief examples of effective/ineffective GPT feedback.

Fig. 1. Pilot design: configuration of the rubric-bound GPT, deployment touchpoints, and evaluation and analysis pipeline (human-in-the-loop)



Participation in the perception surveys was voluntary, with informed consent. Use of the GPT was optional and did not affect grades. Drafts reviewed by the GPT excluded personally identifiable information; reports were shared only within the course's instructional team. We adopted guardrails: disclosure of AI use, citation and integrity reminders, prohibition of automated grading, and instructor oversight of any consequential decisions (e.g., defense readiness). Data were stored in de-identified form for analysis.

To enhance credibility, we triangulated survey perceptions with artifact audits and instructor notes. Nevertheless, this remains a pilot with small samples, limited objective outcomes, and potential self-selection bias. The study aims to establish feasibility and acceptability and to inform the design of a larger, confirmatory evaluation.

3. Results

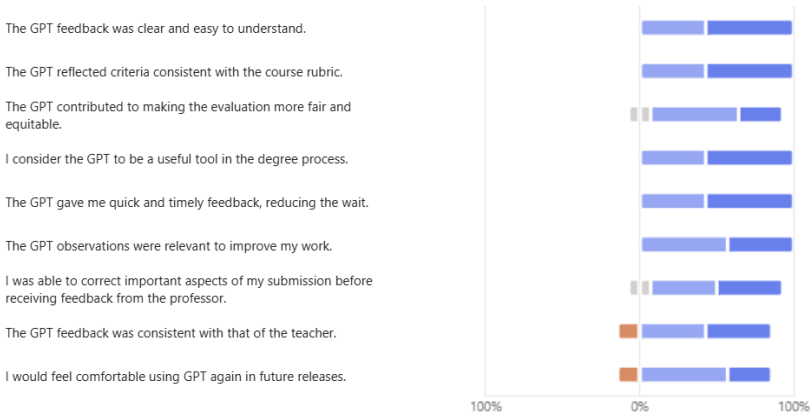
Adoption and use

Across the semester-long pilot, the custom GPT was used at three touchpoints: (i) student drafting, (ii) advisor supervision, and (iii) committee pre-defense preparation. Usage concentrated in Thesis I during early framing (problem/objectives, background, and method sections) and in Thesis II during final polishing (results, conclusions, formatting, and references). Students typically invoked the GPT prior to submitting a draft to their advisor, treating the report as a first-pass triage of structure, APA, and coherence. Instructors reported two dominant patterns: requesting a GPT report to prime their own review (so they could focus on disciplinary nuance) and sharing selected GPT comments as a checklist for live feedback sessions. Examination committees used the report selectively as a pre-brief to surface issues for questioning (e.g., scope–method alignment), while keeping human judgment central to pass/fail decisions.

Student perceptions

The 9-item Likert survey for students (n = 7) showed consistently high agreement across core dimensions, as shown in Figure 2. Students found the GPT's comments clear and easy to understand and well aligned with course rubrics; they rated the tool as useful within the capstone process and timely in reducing waiting time between iterations. Most respondents indicated that the feedback was relevant for improving drafts and that they would use the GPT again in future milestones. Areas of mixed views were limited to fairness/equity and coherence with instructor feedback: while broadly positive, a minority of neutral/disagree responses suggests that not all students perceived the same weighting of criteria or emphasis as their instructors. This aligns with the pilot's design goal—complement, not replacement—and points to continued prompt/rubric calibration.

Fig. 2. Student perceptions of the GPT reviewer (n = 7)

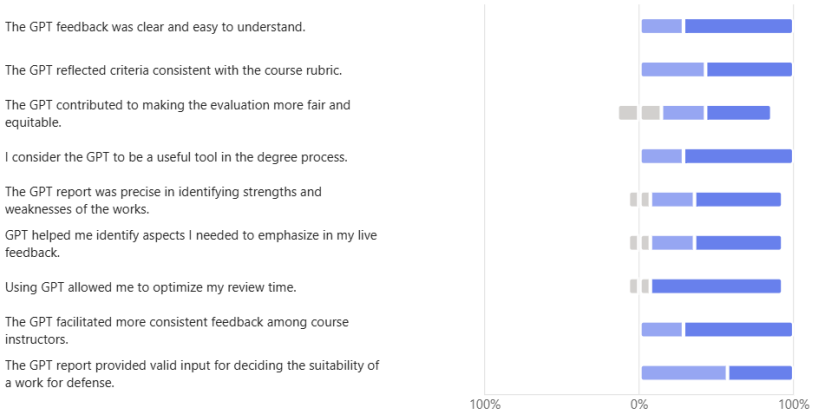


Instructor perceptions

Instructor responses (n = 7) mirrored the student pattern with strong endorsement of the GPT as a complementary, first-pass diagnostic, as shown in Figure 3. Instructors highlighted clarity, rubric alignment, and especially the report's precision in identifying strengths and weaknesses as valuable for shaping live feedback. Respondents

generally agreed that the GPT helped optimize review time and increase cross-instructor consistency, though a small share of neutral responses indicates variability in individual workflows (e.g., those already using detailed checklists perceived smaller time gains). Most instructors agreed the report provided valid inputs for judging defense readiness, while reaffirming that final determinations remained with the committee.

Fig. 3. Instructor perceptions of the GPT reviewer (n = 7)



Artifact audit

A convenience subset of thesis drafts (drawn from Thesis I and Thesis II) was reviewed to observe document-level changes associated with GPT-assisted iterations. Across cases, we observed (a) more complete structure (presence and ordering of mandatory sections), (b) improved APA author–date citations and reference formatting, and (c) clearer alignment between the stated problem/objectives and the described method. Typical revisions following GPT feedback included: adding or tightening problem statements and objectives, clarifying method steps and evidence, consolidating duplicated background content, and correcting citation/reference inconsistencies. Because this was a pilot without controlled pre/post measures or external grading, we treat these as indicative improvements rather than causal effects; they informed the next iteration of prompts and course guidance.

Illustrative cases

Case A (Thesis I – early framing). A student submitted an initial draft to the instructor. The instructor generated a GPT report as a first-pass diagnostic and shared it with the student, using it to structure the feedback meeting. The report flagged misalignment (overbroad problem vs. narrow method), identified missing elements (operational definitions, inclusion/exclusion criteria), and proposed a three-step action plan (tighten the problem → refactor objectives to one general and two specific → map each objective to a method subsection). Crucially, the instructor also provided their own disciplinary feedback—for example, on method feasibility and evidentiary adequacy—explicitly positioning the GPT comments as complementary, not substitutive. The subsequent student draft integrated both sources of feedback, illustrating a human-in-the-loop workflow in which the GPT standardizes the first pass (structure/format) while the instructor retains evaluative judgment.

Case B (Thesis II – results and conclusions). An instructor requested a GPT report on a near-final draft to prepare for a meeting. The GPT surfaced issues in results–discussion coherence (claims not supported by presented data) and APA consistency (mismatched in-text citations and reference list). The instructor used the report as a checklist to drive a targeted session; the student corrected inconsistencies and added traceable links between findings, limitations, and conclusions before the committee review. In the committee pre-defense phase, members used the GPT report to derive questions about validity threats and scope.

Taken together, the adoption patterns and perception data indicate that a rubric-bound GPT can standardize the first pass of thesis review, accelerate feedback cycles, and improve document hygiene (structure, APA, basic alignment) while keeping expert human judgment at the center. The mixed views on fairness and exact coherence with instructor emphasis underscore the importance of ongoing calibration and feedback-literacy scaffolds so students can translate AI suggestions into substantive improvements.

Qualitative insights from open-ended responses

Students. The most frequently cited benefit was speed—students emphasized rapid turnaround and reduced waiting time (e.g., "rapid responses," "fast review"). A second theme was diagnostic specificity, noting that the GPT "pinpointed my weaker parts" and "covered things often overlooked." Several comments valued the critical yet constructive tone and the inclusion of possible solutions.

Asked about improvements, many students reported no changes needed; the suggestions that did appear clustered around pedagogical scaffolding and tone—requests for "a more educational version for students" and to avoid an overly harsh/over-critical feel. Overall experience terms were strongly positive ("excellent," "very good," "future," "cooperative"), reinforcing the acceptability signals seen in the Likert items.

Instructors. Open responses described the report as a useful triage layer that saves review time and structures feedback, with repeated mentions of detecting cohesion/coherence and chapter-to-chapter alignment, and of the report as a systematic starting point before meetings. Several highlighted that the GPT helped emphasize key points during live feedback and increased rubric consistency across the course.

Perceived limitations/risks focused on: prompt dependence ("good prompts matter"), risk of over-automation or blind trust, occasional mixing of external information or context drift, ethics/integrity concerns, a possible tendency toward grade inflation, and insufficient contextualization to the specific thesis.

Suggested improvements centered on process integration (use at planned milestones—seminar, end of Thesis I, end of Thesis II), tighter rubric alignment per program, faculty training for critical use, clear protocols on how to combine AI output with human evaluation, limited browsing to focus strictly on the submitted thesis, and clearer context windows.

4. Discussion

Interpretation of findings. The pilot indicates that a rubric-bound GPT can meaningfully shift feedback earlier in the thesis process, converting waiting time into learning time through immediate, elaborated comments. Student and instructor perceptions converged on clarity, rubric alignment, and usefulness, consistent with evidence that timely, information-rich feedback improves performance when it specifies what to change and how [1–4]. Open-ended responses reinforced these patterns: students repeatedly highlighted speed and diagnostic specificity ("pinpointing weaker parts"), while instructors emphasized time savings, cohesion/coherence checks, and the value of a systematic starting point for meetings. These patterns align with a division of cognitive labor: the GPT standardizes the first pass; instructors retain judgment on method quality, evidence, and claims.

Positioning within the literature. The results echo meta-analytic findings that elaborated feedback outperforms correctness signals and that speed matters for formative cycles [1–4]. At the same time, mixed views on fairness and coherence with instructor emphasis, together with student requests for a more pedagogical tone, reveal the importance of feedback literacy and shared standards [5,6]. Students need support to interpret comments and translate them into revisions; instructors need calibration so prompts mirror local rubrics. The human-in-the-loop design recommended by reviews of AI in higher education is visible here: the GPT supplies rapid scaffolds while instructors orchestrate their use and arbitrate high-stakes decisions [7,8].

What the pilot adds. Prior reports often describe general-purpose writing assistants; this study contributes evidence from an authentic semester with a domain-specific, rubric-bound configuration. Three features appear salient for adoption: (i) explicit rubric mapping (outcomes → criteria → observable evidence) embedded in the prompt; (ii) section-level outputs that mirror thesis structure and are immediately actionable; and (iii) governance guardrails (no grading, transparency of limits, privacy) that signal the complementary role of AI.

Tensions and mitigations. Reservations around fairness likely reflect perceived weighting of criteria and the diversity of instructor styles. We mitigated this through shared exemplars and periodic prompt

updates; future iterations should incorporate calibration workshops and versioned prompts linked to rubric changes. Qualitative comments also flagged risks of over-automation or blind trust, occasional context drift, and tone that can feel overly harsh. Mitigations include student-mode prompts that explain why and how to revise (softer, instructional language), bounded context (disable browsing by default; require thesis excerpts), explicit protocols that instructors always deliver their own feedback, and faculty training to model critical use.

Implications for programs. For capstone courses facing large cohorts and tight timelines, a GPT reviewer can function as a standardization and triage layer, improving document hygiene (structure, APA, traceability between objectives, methods, and claims) and freeing instructor time for conceptual critique. Programs should treat the GPT as part of an assessment ecosystem: integrate it into defined milestones (seminar, end of Thesis I, end of Thesis II), specify that instructors run the tool and share the report with students, teach students how to use feedback, and collect routine data (e.g., review minutes per draft) to monitor benefits and unintended effects.

Scalability and transferability. The workflow is portable to other professional programs that rely on structured reports (engineering, health, applied sciences). Scaling should prioritize (a) local rubric alignment, (b) faculty calibration, and (c) data protection practices suited to institutional policies. As cohorts grow, future research should move beyond perceptions to objective outcomes (revision quality, time-on-task, pass rates) and examine reliability (ordinal α/ω) and dimensionality of perception instruments.

5. Conclusions

This semester-long pilot shows that a rubric-bound GPT can serve as a credible first-pass reviewer in Construction Engineering theses. Students and instructors converged on high perceptions of clarity, rubric alignment, usefulness, and (for students) timeliness, while instructors emphasized diagnostic precision for structuring subsequent human feedback. Open-ended comments converged on the same value proposition—speed and diagnostic specificity—while also pointing to improvements in pedagogical tone and governance. A light

artifact audit suggested improvements in document hygiene (complete structure, APA compliance) and clearer alignment between problems, objectives, and methods. At no point did the GPT replace academic judgment; rather, it standardized and accelerated the preliminary review so instructors could invest scarce time in disciplinary depth.

Two caveats temper these findings: the small, self-selected samples and the absence of controlled pre/post outcomes preclude causal claims; and fairness/coherence perceptions varied, underscoring the need for calibration and feedback-literacy supports. Program-level adoption should therefore pair rubric-bound prompts with instructor calibration, bounded context (e.g., limited browsing), and explicit protocols that safeguard human oversight. Practically, programs can integrate a GPT reviewer as a standardization and triage layer—embedded in milestones, paired with faculty calibration sessions, and accompanied by guidance that helps students translate comments into revisions.

Future work should scale to larger cohorts and report objective outcomes (revision quality, time-on-task, pass rates), alongside reliability and dimensionality analyses of perception instruments. With appropriate governance, GPT-assisted review can convert waiting time into learning time while safeguarding the primacy of human expertise.

6. References

[1] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

[2] Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

[3] Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, 3087. https://doi.org/10.3389/fpsyg.2019.03087

[4] Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. Review of Educational Research, 85(4), 475–511. https://doi.org/10.3102/0034654314564881

[5] Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

[6] Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. Assessment & Evaluation in Higher Education, 38(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

[7] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[8] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? International Journal of Educational Technology in Higher Education, 16, 39. https://doi.org/10.1186/s41239-019-0171-0

[9] Sepúlveda I, Alarcón LF, Barkokebas B, Ebensperger A. Developing a GenAI methodology for data analysis in industrialized construction: A Lean view. In: Seppänen O, Koskela L, Murata K, editors. Proceedings of the 33rd Annual Conference of the International Group for Lean Construction (IGLC33); 2025 Jun 2–8; Osaka and Kyoto, Japan. p. 965–975. https://doi.org/10.24928/2025/0231