Scaling Course Evaluations with Large Language Models: Semester-level Digestible Student Feedback for Program Leaders

Zhang, Mike; Lindsay, Euan; Quitzau, Maj-Britt; Bjerva, Johannes

Publication date:
2025

Document Version
Early version, also known as pre-print

Scaling Course Evaluations with Large Language Models: Semester-level Digestible Student Feedback for Program Leaders

M Zhang [a], E D Lindsay [b,1], M-B Quitzau [c], J Bjerva [d]

[a] Aalborg University, Copenhagen, Denmark, https://orcid.org/0000-0003-1218-5201

[b] Aalborg University, Aalborg, Denmark, https://orcid.org/0000-0003-3266-164X

[c] Aalborg University, Copenhagen, Denmark, https://orcid.org/0000-0002-9907-8224

[d] Aalborg University, Copenhagen, Denmark, https://orcid.org/0000-0002-9512-0739

***Conference Key Areas**: Digital tools and AI in engineering education, Building the capacity and strengthening the educational competencies of engineering educators*
***Keywords**: Course Evaluations, Automated Feedback, Large Language Models*

ABSTRACT

End of semester student evaluations represent the primary feedback mechanism for academics' teaching practices. However, at the department or semester level, the sheer volume of feedback renders traditional analysis methods impractical. This paper addresses a gap in previous work where only course-level synthesis is explored using open-source generative AI for creating factual, actionable, and appropriate summaries of student feedback across an entire department. Instead, our study analyses 28 semester-level evaluation reports with student comments—with approximately 25,000 words and 170,000 characters—spanning the department, the model produces insights on several levels, namely degree-level, semester-level, year-level, and department-level. Through structured prompting, we developed a methodology that meets the specific needs of study board chairs who previously faced high workload from manually reviewing evaluations twice yearly. Our prompts allow the model to systematically checks for predetermined themes, while also identifying emergent patterns across courses. This approach enables targeted professional development initiatives at the departmental scale. Our contribution demonstrates that generative AI can effectively synthesize student feedback at a large organizational level, providing a cost-effective mechanism to support educational development and quality improvement across an entire academic unit.

1. INTRODUCTION

Feedback is essential for improving educational practices, benefiting both students and educators by highlighting areas for development (Hattie 2008; Narciss 2008). While considerable effort is devoted to providing meaningful feedback to students, the mechanisms through which academics receive feedback on their teaching are less robust. Typically, academics rely heavily on end-of-semester student

evaluations, combining quantitative Likert-scale items with open-ended responses. However, as student numbers and feedback volume grow, effectively synthesizing these evaluations becomes increasingly challenging. Consequently, departments often default to simplistic interpretations, primarily responding only to extremely high or low ratings, potentially overlooking nuanced insights contained in qualitative comments.

Recent advances in natural language processing (NLP), particularly through large language models (LLMs), have opened new possibilities for handling extensive textual data across many domains (Wahle et al. 2023; Kasneci et al. 2023). Previous research leveraging LLMs for educational feedback has primarily focused on peer learning, student performance, formative assessment, coding assistance, or student feedback (Bauer et al. 2023; Botelho et al. 2023; Guerraoui et al. 2023; Liang et al. 2023; Pankiewicz and Baker 2023; Katz, Gerhardt, and Soledad 2024; Zhang et al. 2025) and the responsible development thereof (Lindsay et al. 2025). Additionally, automated tools have been explored for monitoring student engagement and providing instructor feedback based on classroom discourse (Samei et al. 2014; Kelly et al. 2018; Jensen et al. 2020; Schwarz et al. 2018; Aslan et al. 2019; Alrajhi et al. 2021; Demszky and Liu 2023; Demszky et al. 2023; Wang and Demszky 2023). However, little attention has been given to systematically synthesizing student course evaluations to inform teaching practices at departmental or institutional scales.

Previous work also demonstrated the feasibility of applying LLMs to generate meaningful feedback summaries at the individual course level (Zhang et al. 2025). Yet, the true potential of this technology emerges at scale—when synthesizing evaluations across an entire department or study board becomes impractical or impossible manually, causing significant strain on faculty tasked with this process. Thus, a key research question emerges:

RQ: How can LLMs be applied to automatically synthesise large volumes student feedback in a manner that is factual, actionable, appropriate, and generates value for academic leadership at a program or department scale?

This work investigates to apply this to the semester-level at the Department of Sustainability and Planning covering the 2023–2024 academic year. The dataset comprises evaluations from 116 courses across 3 programs spanning 10 academic semesters, totalling student responses of total 25,000 words provided in both Danish and English. Data validation incorporates study board minutes, course summaries, and insights from stakeholders such as semester coordinators and study board chairs.

To guide our analysis, we adopt three critical criteria drawn from previous literature (Wang and Demszky 2023; Wang et al. 2023; Guo et al. 2023; Chang et al. 2023; Zhang et al. 2024):

1. Factuality: Ensuring generated feedback accurately represents student evaluations without introducing inaccuracies or irrelevant content.
2. Actionability: Providing actionable insights rather than mere summaries of evaluation data.
3. Appropriateness: Focusing exclusively on pedagogical matters, filtering out inappropriate or non-instructional commentary.

Our goal is to demonstrate that simple prompting strategies and careful evaluation of LLM outputs can significantly enhance the usability and impact of student feedback

at an institutional scale, informing professional development and improving teaching quality.

2. Our Approach

2.1 Data Pre-processing

Our dataset is the complete set of course evaluation documents for the 2023/24 academic year at the Department of Sustainability and Planning of Aalborg University. The data is initially in PDF-format containing quantitative metrics such as course scores and free text qualitative comments from students. We are mainly interested in the qualitative comments. We use the Marker library (Paruchuri 2025), which converts PDF documents to several possible formats. We transform the data to simple markdown and apply post-processing to extract the text of *only the student comments* from the PDF. This paper reports on analysis of three programs, with the results partitioned across the ten semesters of these programs.

2.2 Selecting a Format for the Output

In previous work (Zhang et al. 2025), we allowed for the formatting of outputs to emerge organically from the model. In this work, we deliberately guide the format of the output to better support the users of the synthesis (i.e., a study board). In doing so we have moved away from asking the models for suggestions; out-of-the-box language models can provide a synthesis, but they lack the context familiarity to provide actionable advice.

Stakeholder discussions revealed that there are already structures in place for analysing student feedback responses, with specific categories for analysis already identified, such as “academic”, “practical”, “well-being” remarks. The model was therefore explicitly prompted to organise its synthesis of the responses using those categories (see Section 2.3). In addition, we also prompted the model to identify a miscellaneous category to allow for the emergence of other themes in the data that may not be visible a priori.

This simplistic approach does not yet allow us to control for how common a theme should be before it is reported. In some instances (such as inappropriate behaviour), every instance is meaningful and should appear in the synthesis. For other themes (e.g., pace of delivery) only an emergent consensus should appear in the synthesis. The extent to which an untrained model can manage this challenge forms part of the evaluation in this study.

2.3 The Model

For processing the text with an LLM, we used Qwen2.5 (7B; Qwen-Team et al. 2025) as the model to synthesise the student feedback. This model originates from end 2024. Qwen2.5 is an open-source 7-billion-parameter language model based on the Transformer architecture (Vaswani et al. 2017). We directly prompt the model with the following prompt:

Process the following student evaluations in Danish or English and transform them into a JSON array. Each element of the array should be an object with the following keys:

- Module Name: Usually the semester in general and several course modules;
- Semester Coordinator: if not applicable, put N/A;
- Academic Remarks;
- Practical Remarks;
- Well-being Remarks;
- Harassment;
- Internal Comment: Admin or policy notes.
  - Feedback: Summary for the coordinator, highlighting issues to address.
  - Miscellaneous: Anything that is not mentioned yet.

Guidelines:

- Do not include extra keys.
- Use "N/A" for missing data.
- If the evaluation is not module-specific, use "General Semester Feedback" as the Module Name.

Follow the format of the following output example:

```
[ {
    "Module Name": "",
    "Semester Coordinator": "",
    "Academic Remarks": "",
    "Practical Remarks": "",
    "Well-being Remarks": "",
    "Harassment": "",
    "Internal Comment": "",
    "Feedback": "",
    "Miscellaneous": ""
} ]
```

Note there can be more than one dictionary in the list. Only return JSON and nothing else.

In the prompt, we state the format the model should follow when generating text. In this case, we require JavaScript Object Notation (JSON), contemporary LLMs are potent in generating this format and others, e.g., the Extensible Markup Language (XML). We chose the specific columns of the output to be following the original format of the study boards from Aalborg University.

The model does not require any training, nor does it need any specific examples. We decided not to fine-tune the model as it is first non-trivial on what should be trained on. Furthermore, having a prompting strategy allows for a simpler and faster implementation process. It also allows for a transferable implementation. In doing so, however, we forego the improved performance that tailoring the model to our context would provide.

In previous work (Zhang et al. 2025), the bottleneck of was that the LLM used (Llama2) was not able to handle context longer than 4K tokens. Instead, Qwen2.5 can handle long-context support up to 128K tokens and can generate up to 8K tokens.

3. RESULTS

In this section, we present illustrative examples of the LLM outputs and summarize our key observations. To align outputs with study board preferences, we structured the responses from the LLM strictly in JSON format, as detailed in Section 2.3.

*Table 1. **Illustrative Excerpts.** We show examples of the generated output of the LLM. We omit the JSON structure for readability.*

Generated Output of the Model
(Example 1; *Semester 1*)
"**Module Name**": "General Semester Feedback", "**Semester Coordinator**": [REDACTED], "**Academic Remarks**": "The semester plan was not provided in advance, making it difficult to plan work, holidays, etc., in good time. The exam schedule was also announced late, which caused difficulties. There was a lack of learning on sustainability, particularly regarding the three elements of sustainability. The practical project on user-oriented product design was poorly organized, with students being expected to build a complete playground equipment during winter when there were few children available for testing. There was much confusion due to conflicting messages from different lecturers.", "**Practical Remarks**": "The [REDACTED] component was mixed with the clinic project, leading to a less satisfactory exam experience. The project lacked clear guidance and support, especially in terms of testing the playground equipment in unfavorable conditions. The deadline was too tight, and there was a lack of clarity about the expectations and support provided.", "Well-being Remarks": "There was a lack of a study café for mechanics, which meant that students lacked support. The university was under construction, causing disruptions during lectures. The university did not foster a sense of

community and did not provide a space where students could gather and socialize.",
"**Harassment**": "N/A",
"**Internal Comment**": "N/A",
"**Feedback**": "Improve the clarity of communication and provide a clear schedule well in advance. Enhance the support for sustainability-related courses and practical projects, especially in terms of testing and guidance. Create a study café for mechanics and improve the learning environment by addressing construction noise during lectures. Foster a stronger sense of community and provide spaces for students to gather and socialize.",
"**Miscellaneous**": "Student engagement with tutors was positive, but there was a need for better coordination and clear communication from the university administration."

(Example 2; *Semester 5*)

"**Module Name**": "General Semester Feedback",
"**Semester Coordinator**": [REDACTED],
"**Academic Remarks**": "The exam form for all three courses (excluding the project) is unrealistic. It's unreasonable to expect students to complete a task over 20 pages long for a 5 ECTS course and have two oral exams without proper preparation.",
"**Practical Remarks**": "There has been a lack of suitable group rooms, especially this semester. The available rooms were not conducive to working and were spread across the entire campus, making it difficult to create a good sense of community and see each other.",
"**Well-being Remarks**": "N/A",
"**Harassment**": "N/A",
"**Internal Comment**": "N/A",
"**Feedback**": "Address the issue of unrealistic exam forms and lack of adequate group workspaces. Consider ways to improve social cohesion among students.",
"**Miscellaneous**": "Students feel the exam workload is too heavy and the group work environment is inadequate."

(Example 3; *Semester 10*)

"**Module Name**": "General Semester Feedback",
"**Semester Coordinator**": [REDACTED],
"**Academic Remarks**": "Opportunity to discuss projects with other students and teachers was positive. Communication about exam dates could be stronger and earlier.",
"**Practical Remarks**": "Additional lecture on data collection and its impact on problem formulation would be beneficial. Study spaces were not provided, causing issues with scheduling and workspace for group projects.",

"**Well-being Remarks**": "The study programme's closure and teacher departures have negatively impacted students' sense of future and well-being.",
"**Harassment**": "N/A",
"**Internal Comment**": "Review the decision regarding study spaces allocation and ensure transparency in such decisions. Consider providing digital resources and support for students.",

"**Feedback**": "Improve communication about exam dates, provide additional lectures on specific topics, and ensure adequate study spaces for collaborative work. Address the closure of the study programme and provide support for affected students.",
"**Miscellaneous**": "Study spaces were not allocated to 10th semester students, leading to logistical issues. Digital resources and support should be provided."

Table 1 outlines the findings across different semesters within the same degree program, highlighting distinctive feedback patterns from students in their first, fifth, and tenth (final) semesters. Notably, students in their first semester frequently articulated a clear need for transitional support into the program. Conversely, students approaching program completion in their tenth semester emphasized the importance of greater autonomy and opportunities for improvement. These nuanced differences underscore the potential of leveraging an LLM to customize support services to distinct student cohorts, while simultaneously identifying overarching insights from aggregated data.

On a qualitative level, we observed a consistent pattern wherein the model organizes its output into clearly defined sections: “General Semester Feedback”, followed by individual course feedback labelled sequentially (e.g., “Course_1”, “Course_2”, “Course_n”). Interestingly, the model maintains this structured approach even in scenarios where a course received minimal or no feedback, explicitly noting the absence of specific remarks.

The model also demonstrated the ability to capture rare but significant incidents effectively. An illustrative case in our dataset involved an infrequent mention of harassment—one of only two such incidents captured by the LLM—highlighting its sensitivity in identifying critical issues from otherwise routine feedback.

Finally, we identified instances of language code-mixing when processing evaluation reports originally written in Danish, indicating the LLM's responsiveness to multilingual inputs and its nuanced handling of mixed-language data.

4. Evaluation of the Outputs

The outcomes of the model were validated anecdotally by the authors, including the chair of the Study Board for the degree programs included in the dataset. Outcomes were compared with study board minutes, contemporaneous course summary documents, and the recollections of the Study Board Chair.

The strongest finding was that the models bring a different lens to the process of synthesising student evaluations at a larger scale. Unlike the humans whose task is interpreting the data, the model bring neither memory nor context to the process. Our study board chair noted that they are usually looking for specific things in the student evaluations, guided by the previous offerings of a course, by other sources of feedback and interaction with students throughout the semester, or by an expectation of change, such as from a new academic teaching the course. The model, however, simply synthesises the student responses, which can lead to a difference in emphasis when it comes to producing summaries.

We also observed that for some of our identified categories, the model had differing interpretations – particular with regards to the “academic” and “practical” issues. While there is a clear distinction in the mind of the study board as to the difference between these categories, this nuance is not represented by the untrained model and its relatively simple prompt. This could be addressed through a more sophisticated training and/or prompting in the future.

The model was able to deal with summarising low volume responses. Unlike our previous iteration that had a tendency to add to the summary details not present in the dataset, this model was able to acknowledge that there was insufficient data and respond with a blank or “N/A” response.

Overall, the model rated well on appropriateness. While the underlying dataset itself has little in the way of inappropriate raw data, the synthesis from the model does not convey inappropriate tone or messaging.

The outcomes of the model also performed well with regard to actionability. The summaries produced by the model did provide insights that could be used to inform future practice, particularly at the larger scale (eg semester or overall). The themes that emerge in the overall synthesis of multiple courses are the themes that most warrant some kind of action on the part of the teaching team, and so the approach appears to naturally tend towards providing actionable summaries.

One observation regarding the summaries is that the model is not able to identified problems that have in fact already been resolved. When students complete their evaluations they reflect upon their experience throughout the whole semester, which can include frustrations with early-semester difficulties (such as a teacher illness) that have in fact already been addressed. While capturing these issues is factual, emphasising them in a summary actually serves to make it less actionable, because action has already been taken.

5. Conclusion

In this work, we present a proof-of-concept leveraging open-source large language models to automatically synthesise course evaluations into a digestible format for academic leaders. We prompted an out-of-the-box open-source LLM to summarise course evaluations and synthesise student responses across multiple courses within a single semester.

The inherent context-free nature of the LLM-based synthesis allowed the model to provide a different perspective to academic leaders. This lack of context, however, also serves to limit somewhat the actionability of the summaries that are provided. Further work to develop more sophisticated prompting, as well as to incorporate context into the synthesis will serve to improve the value of the summaries.

Despite the simplicity of our model, our findings suggest that the resulting synthesis is largely representative of the overall student evaluations, as previously captured by other channels in the student evaluation framework. This suggests that such tools offer the potential for simply and inexpensively offering a big picture view of student evaluations where previously it was either extremely labour intensive or outright impossible.

6. ACKNOWLEDGEMENTS

This work was supported by a research grant (VIL57392) from VILLUM FONDEN. We acknowledge the assistance of our colleagues in validating the outputs of our models against their recollections of the academic year in question.
This work was approved by the Aalborg University's Human Research Ethics committee, with approval number 2024-505-00440.

REFERENCES

Alrajhi, Lina, Amirah Alamri, Francisco D. Pereira, and Alexandra I. Cristea. 2021. "Urgency analysis of learners' comments: An automated intervention priority model for MOOC." Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings. Springer International Publishing, 148–160.

Aslan, Senem, Nihan Alyuz, Cihan Tanriover, S. Ertan Mete, Erhan Okur, Sidney K. D'Mello, and Asli Arslan Esme. 2019. "Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms." Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 1–12.

Bauer, Eva, Maria Greisel, Ilya Kuznetsov, Michael Berndt, Ingo Kollar, Markus Dresel, ..., and Frank Fischer. 2023. "Using natural language processing to support peer-feedback in the age of artificial intelligence: a cross-disciplinary framework and a research agenda." British Journal of Educational Technology 54 (5): 1222–1245.

Botelho, Amber, Suchismita Baral, Justin A. Erickson, Prachetas Benachamardi, and Neil T. Heffernan. 2023. "Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics." Journal of Computer Assisted Learning 39 (3): 823–840.

Chang, Yi, Xisen Wang, Jing Wang, Yichao Wu, Lingfei Yang, Kaixiang Zhu, ..., and Xing Xie. 2023. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology.

Demszky, Dorottya, and Jing Liu. 2023. "M-powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes." Proceedings of the Tenth ACM Conference on Learning@ Scale. 59–69.

Demszky, Dorottya, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2023. "Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course." Educational Evaluation and Policy Analysis.

Guerraoui, Rachid, Philipp Reisert, Naoya Inoue, Firoj Alam S. Mim, Keshav Singh, Jinho D. Choi, ..., and Kentaro Inui. 2023. "Teach Me How to Argue: A Survey on NLP Feedback Systems in Argumentation." Proceedings of the 10th Workshop on Argument Mining. 19–34.

Guo, Zhengyang, Ruixiang Jin, Chuang Liu, Yefeng Huang, Dapeng Shi, Li Yu, ..., and Deyi Xiong. 2023. "Evaluating large language models: A comprehensive

survey.” arXiv preprint arXiv:2310.19736.

Hattie, John. 2008. Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge.

Jensen, Emily, Margaret Dale, Patrick J. Donnelly, Christopher Stone, Scott Kelly, Amanda Godley, and Sidney K. D’Mello. 2020. “Toward automated feedback on teacher discourse to enhance teacher learning.” Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 1–13.

Kasneci, Enkelejda, Kai Seßler, Stefan Küchemann, Maria Bannert, Daria Dementieva, Frank Fischer, ..., and Gjergji Kasneci. 2023. “ChatGPT for good? On opportunities and challenges of large language models for education.” Learning and Individual Differences 103:102274.

Katz, Adina, Michael Gerhardt, and Marisa Soledad. 2024. “Using Generative Text Models to Create Qualitative Codebooks for Student Evaluations of Teaching.” arXiv preprint arXiv:2403.11984.

Kelly, Scott, Andrew M. Olney, Patrick Donnelly, Martin Nystrand, and Sidney K. D’Mello. 2018. “Automatically measuring question authenticity in real-world classrooms.” Educational Researcher 47 (7): 451–464.

Liang, Weixin, Yuhui Zhang, Han Cao, Bingzhe Wang, Danni Ding, Xiaochuang Yang, ..., and James Zou. 2023. “Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis.” arXiv preprint arXiv:2310.01783.

Narciss, Susanne. 2008. “Feedback strategies for interactive learning tasks.” In Handbook of research on educational communications and technology, edited by J Michael Spector, M David Merrill, Jeroen van Merriënboer, and Marcy P Driscoll, 125–143. Routledge.

Lindsay, E.D., Zhang, M., Johri, A. and Bjerva, J., 2025. The Responsible Development of Automated Student Feedback with Generative AI. *2025 IEEE Global Engineering Education Conference (EDUCON)*

Pankiewicz, Michal, and Ryan S. Baker. 2023. “Large Language Models (GPT) for automating feedback on programming assignments.” arXiv preprint arXiv:2307.00150.

Paruchuri, Vik. 2025. Marker: Convert PDF to Markdown and JSON Quickly with High Accuracy. https://github.com/VikParuchuri/marker. Accessed: 2025-03-11.

Qwen-Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, and Xingzhang Ren an. 2025. Qwen2.5 Technical Report.

Samei, Behnaz, Andrew M. Olney, Scott Kelly, Martin Nystrand, Sidney D'Mello, Nicolas Blanchard, ..., and Arthur Graesser. 2014. "Domain Independent Assessment of Dialogic Properties of Classroom Discourse." Grantee Submission.

Schwarz, Baruch B., Naama Prusak, Osama Swidan, Avivit Livny, Kinneret Gal, and Annelise Segal. 2018. "Orchestrating the emergence of conceptual learning: A case study in a geometry class." International Journal of Computer-Supported Collaborative Learning 13:189–211.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ..., and Illia Polosukhin. 2017. "Attention is all you need." Advances in Neural Information Processing Systems, Volume 30.

Wahle, Jan Philip, Terry Ruas, Muhammad Abdalla, Bela Gipp, and Saif Mohammad. 2023. "We are Who We Cite: Bridges of Influence Between Natural Language Processing and Other Academic Fields." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 12896–12913.

Wang, Ruihan, and Dorottya Demszky. 2023. "Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction." Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). 626–667.

Wang, Ruihan, Qian Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. "Step-by-Step Remediation of Students' Mathematical Mistakes." arXiv preprint arXiv:2310.10648.

Zhang, M., E. Lindsay, F. B. Thorbensen, D. B. Poulsen, and J. Bjerva. 2024. "Leveraging Large Language Models for Actionable Course Evaluation Student Feedback to Lecturers." Proceedings of the 52nd Annual Conference of SEFI. Lausanne, Switzerland.

Zhang, Mike, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D Lindsay, and Johannes Bjerva. 2025. "SEFL: Harnessing Large Language Model Agents to Improve Educational Feedback Systems." arXiv preprint arXiv:2502.12927.