Exploring AI-Human Triangulation for Research Reproducibility

Contents

Executive summary

Key project outcomes

The project set out to explore how generative AI could be used to teach students critical skills in evaluating the reproducibility of scientific research. The team developed a web-based application—the Reproducibility Analyser—which guides users through a structured evaluation of manuscripts using large language models (LLMs) and a checklist of reproducibility standards. The project aimed not just to automate this analysis but to embed human judgment and critique into each stage, fostering AI-assisted critical thinking.

Our outputs include the development of a modular, LLM-powered analysis pipeline, a React-based user interface, and an integrated PDF viewer to facilitate cross-checking of AI outputs. The tool performs stepwise analysis, including metadata extraction, applicability filtering, domain/item-level evaluation, and feedback generation. The Reproducibility Analyser is accessible at https://bit.ly/reproai (code: reproai2025).

We ran a successful codesign workshop with students in March 2025. Students could see the potential utility of the Analyser, highlighting the utility of splitting checklist items to facilitate checking and the transparency of the AI's reasoning. They proposed a number of concrete improvements—many of which have already been implemented. These included better error handling, more precise quote extraction, and scope clarification regarding single vs. multiple study designs within manuscripts.

Alongside teaching applications, the project generated new potential outputs. A meeting with Springer Nature confirmed that while the approach aligns with publishing priorities around reproducibility, no similar tool currently exists on their side—highlighting the project's novelty and potential for broader impact.

The project also laid foundations for related initiatives. These include the AI Evidence Analyzer, a system for structured EdTech evidence synthesis, and the DMP Analyzer, a tool for evaluating data management plans being adopted for national use by CSC – IT Centre for Science, Finland. Together, these tools demonstrate the strong potential of modular, AI-assisted, checklist-driven assessment beyond teaching alone.

Key learning for the team around using AI for developing teaching and learning

Working at the intersection of pedagogy and generative AI taught the team several key lessons. First, AI tools in educational settings must prioritize transparency and explainability. Students were quick to spot inconsistencies in

AI outputs and expressed hesitation when they couldn't trace claims back to source material. In response, the integration of a PDF viewer and structured justifications became not just a feature, but a pedagogical necessity.

Second, the project underscored the value of involving students as co-designers. Their feedback during the pilot workshop was not only detailed and technical—it often anticipated usability or ethical challenges before they became apparent to the development team. For example, their request to restrict quote extraction to methods sections or allow challenges to "non-applicable" classifications led to a restructuring of the AI pipeline logic.

Third, the team learned that rapid prototyping is valuable, but it must be supported by a simple user-interface if the goal is learning, not just testing. The HTML prototype was functional but flawed in keyways. The move to a React frontend was triggered by student feedback with the challenges they were finding in using the tool, and dramatically improved both trust in the tool and its learning value.

Finally, the project showed that integrating AI into education should not be about delegating tasks to machines, but about building reflective systems where AI supports decision-making, but does not make decisions. The most promising moments in the workshop were not when the AI was "right," but when students noticed it was wrong—and could say why.

This human-AI triangulation model holds promise not only for teaching reproducibility but for cultivating research literacy more broadly. As institutions wrestle with how to adopt AI in higher education, the Reproducibility Analyser offers a useful case for how AI can support students to make their own critical judgements.

Project introduction

Background and context

The reproducibility crisis has highlighted the need for better research practices and more transparent methodologies. As generative AI becomes increasingly embedded in academic practice, it presents both new risks and new opportunities for research reproducibility. This project explores how AI-human collaboration can be utilised in teaching to promote reproducibility as a core research skill. It aligns with growing institutional efforts to embed open science practices in training and supports students in becoming critical consumers and producers of research in the AI era.

Objectives

The project aimed to (1) equip students with skills to critically assess research reproducibility, (2) develop a generative AI tool that supports reproducibility analysis through human-AI collaboration, and (3) train students and staff to evaluate and improve AI outputs in academic contexts. Intended outcomes included a working prototype of an AI-assisted reproducibility tool, a structured evaluation template and checklist, and a collaborative repository of student-led analyses that can be iteratively expanded.

Scope

The project focused on evaluating the adherence of research papers to reproducibility standards such as method transparency, data/code availability, and pre-registration alignment. The scope was limited to selected aspects of reproducibility that are teachable and assessable within a short course or workshop. Although two student workshops were planned, only one was conducted during the funding period due to challenges in student participation during a busy Trinity exam term; the second will be delivered after summer. Importantly, development of the digital infrastructure progressed significantly beyond the original scope.

Tools and technologies

We developed a custom web-based platform to support AI-assisted analysis of research papers for reproducibility. The system uses a structured checklist framework to assess adherence to reproducibility standards—such as data and code availability and methodological transparency —through a transparent human-AI collaboration pipeline. The tool is accessible at https://bit.ly/reproai using passcode: reproai2025.

At the core of the system is a multi-step LLM-powered analysis pipeline. Uploaded PDFs are parsed into sentence-level data, and relevant sections are identified for further evaluation using a user-selected reproducibility checklist. The analysis pipeline incorporates chained LLM calls to perform sentence classification, checklist applicability filtering, and structured compliance evaluation. Outputs are justified with structured AI-generated text fields, which are traceable back to ‘key sentences’ in the source document via an integrated PDF viewer.

The backend is built with Python, FastAPI, and MongoDB, and the frontend is implemented in React. The platform supports versioned model comparisons, structured feedback capture, and lays the groundwork for iterative expansion through modular AI components.

This tool stands apart from static reproducibility checklists by enabling dynamic, user-involved evaluations and transparent AI outputs. It can be actively refined based on user feedback and supports our long-term goal of integrating reproducibility as a core research skill in teaching and assessment.

Collaboration

The project was implemented with a cross-institutional team from the University of Oxford and the University of Turku, Finland. Collaboration included technical support from the Competency Centre for AI in Teaching and Learning, academic input from open science experts, and pilot testing with students. The project also prepared the groundwork for possible future collaboration with a major academic publisher (Springer Nature).

Project outcomes and findings

Evaluation results

A student codesign workshop was held in March 2025 to evaluate the initial prototype of the AI-assisted reproducibility tool and allow students the opportunity to suggest new design features. Students engaged with the tool in a hands-on session and provided detailed feedback on both its educational value and technical performance.

Participants responded positively to the structured domain/item-level framework for assessing reproducibility. This checklist-based approach helped students critically evaluate research transparency and identify specific methodological strengths and weaknesses. The concept of using AI to support reproducibility assessments was well received, with students recognising its potential to streamline evaluation tasks that are typically manual, time-consuming, and inconsistently applied.

In parallel, the project team held a meeting with representatives from Springer Nature to explore alignment with publisher-side reproducibility initiatives. While Springer Nature confirmed that the project's aims were closely aligned with their strategic interests—particularly around reproducibility and AI in peer review—they noted that they did not have a comparable tool under development. This reinforced the novelty and relevance of our approach

Quantitative data (if collected)

Formal quantitative metrics were not collected.

Qualitative insights (if collected)

Key suggestions included clarifying the scope of analysis (e.g. single vs. multiple studies per paper), improving quote accuracy, and enabling better alignment between AI judgments and original text. These insights directly informed a major redesign of the tool.

The prototype has since been refactored into a more advanced React-based platform, with new features like a built-in PDF viewer that connects AI-generated outputs back to their source context—addressing transparency concerns and improving the learning experience. Further testing will follow in the second workshop scheduled after summer.

Lessons learned

Challenges

A key challenge was managing the tension between rapid prototyping and pedagogical usability. The initial version of the tool, built in HTML, functioned adequately for internal testing but lacked responsiveness and flexibility in a live teaching setting. Students encountered frustrating technical issues during the workshop, including crashes, submission errors, and confusing quote outputs. These were compounded by unclear handling of papers with multiple studies and by limitations in how AI judgments were presented and justified.

Another challenge was recruiting enough participants for the second planned workshop, which had to be postponed due to low sign-ups. Students expressed interest in attending, but were busy studying for exams. While this limited short-term evaluation, it also highlighted the importance of aligning project timelines with academic calendars and student availability.

Key takeaways

User feedback was not just useful—it was transformative. Students offered practical, specific, and sophisticated suggestions that directly shaped the next development phase. In response, the tool was refactored into a modern React-based application, resolving many of the earlier interface issues and adding a built-in PDF viewer to allow users to trace AI-generated quotes back to their source. This significantly improved transparency and trust in the tool.

Importantly, this project has contributed to the foundation for other AI-supported research assessment tools now under development, using the same underlying technology developed. One is the EduEvidence Analyzer, a parallel initiative focused on evaluating EdTech impact studies for evidence quality. Another is the DMP Analyzer, a tool for assessing data management

plans against institutional guidelines, currently being piloted for national implementation in Finland by CSC – IT Center for Science. These connected efforts demonstrate the scalability and cross-domain applicability of structured, transparent AI analysis in academic research.

Advice for teams

For future projects, we recommend testing early with target users—even if the tool is still underdeveloped. Plan for iteration, expect technical feedback, and prioritise explainability when using AI. Collaboration with external stakeholders (e.g. publishers) was also productive: our meeting with Springer Nature confirmed both the uniqueness and the potential relevance of our approach beyond teaching contexts. This opens up opportunities for future partnerships that bridge education, research, and publishing innovation.

Appendices

Digital artefacts

The Reproducibility Analyser is a modular web application developed to evaluate scientific manuscripts using Generative AI. It assesses research papers against a checklist of reproducibility standards and generates structured feedback for students and researchers.

A public instance is available at: https://bit.ly/reproai (passcode: reproai2025)

Prompts and Processing Modules

The system uses chained LLM prompts within a modular pipeline that includes:

1. Text Extractor (PDF parsing via Dockling API)
2. Checklist Applicability Module (identifies applicable criteria)
3. Metadata Extractor (detects study type, field, and research context)
4. Domain & Item Evaluators (LLM-based compliance checks)
5. Domain Reconciler & Summarizer (harmonizes and rates results)
6. Feedback Note Generator (constructs improvement guidance)

These prompt templates can be made available upon request to support reuse or adaptation.