5

ACADEMIC STAFF AI LITERACY DEVELOPMENT THROUGH LLM PROMPT TRAINING

LOUISE DRUMM[a] AND ASHKAN SAMI[b]

[a]*Department of Learning and Teaching Enhancement, Edinburgh Napier University, Scotland*
[b]*School of Computing Engineering and the Built Environment, Edinburgh Napier University, Scotland*

ABSTRACT

A foundation in artificial intelligence (AI) literacy among all academic staff is essential for supporting students' AI literacy effectively. As tools like ChatGPT increasingly influence academic work, educators need to understand prompt engineering and reconsider assessment designs. However, many lack the necessary training or time to engage with courses, limiting their ability to design assessments that leverage these technologies while maintaining academic integrity. This project investigated the impact of prompt training on university academics' abilities to craft prompts and redesign assessments within a Scottish university. A two-hour workshop on prompt engineering was conducted for academic staff, during which participants graded AI-generated content before and after they received training. Results indicated a significant improvement in the quality of prompts crafted by participants post-training. Qualitative feedback revealed mixed

reactions, highlighting both the potential and limitations of AI in academic settings. The study demonstrated the need for ongoing staff development in AI literacy.

Keywords: AI literacy; training; academics; prompt engineering; assessment; faculty

INTRODUCTION AND BACKGROUND

In order for academics to support and develop students' AI literacy (Firat, 2023), they need support and training themselves in understanding AI and effective practices with these tools. As assessment often drives learning (Fischer et al., 2024) and generative AI is frequently seen as a means for students to circumnavigate that learning (Cotton et al., 2024), the design of assessments in relation to AI is a helpful lens to support academics in developing their own AI literacies. This project aimed to first assess the large language model (LLM) prompt literacy of academic staff and then evaluate the impact of prompt training on developing their thinking on module assessment redesign. Using real assessment briefs, academics were asked to grade the responses generated by ChatGPT twice during a workshop – once before they were trained on good practice in prompt engineering, and once afterwards – and they then were surveyed on their immediate response to the exercise and their perceptions of the implications for assessment design and redesign in the future.

The objective was to inform and support curriculum and assessment design decisions at the university and to empower educators to address AI literacy for their own learners within the context of their disciplines. Ethical approval was granted by the local university research integrity committee, and the workshop was designed so that data would be collected to answer the research questions:

RQ1: What impact does prompt writing training have on academics' competence in writing prompts?

RQ2: How does prompt writing training affect academics' thinking about the validity and suitability of assessment types?

OVERVIEW OF INTRODUCTION TO PROMPT ENGINEERING

Prompt engineering is the process of presenting and changing input queries to optimise the performance and outputs of LLMs like ChatGPT. This practice

is essential for guiding these models to generate accurate, relevant, and contextually appropriate responses.

When formulating prompts, several strategies can enhance the quality of the answers. Firstly, including specific details in queries reduces ambiguity and improves the relevance of responses (Lo, 2023). Secondly, requesting the model to adopt a specific persona can tailor responses to fit a particular style and ensure consistency. Additionally, dividing a prompt using text delimiters like quote marks separates the instructions from the context of the model. Step-by-step instructions and providing examples to illustrate the desired response have been found to be effective (Wei et al., 2023). Lastly, specifying the desired length of response helps manage the level of detail and conciseness in the response.

Academic research in the field of LLM and natural language processing supports the efficacy of these techniques (Liu et al., 2023), and they formed the basis of the training course described here.

COURSE OVERVIEW

The course consisted of a two-hour workshop designed for academic staff involved in assessing students in credit-bearing undergraduate or taught postgraduate programmes. The workshop was part of a suite of resources, events, and guidance provided to staff members on LLMs, which attendees could optionally access prior, or subsequent, to the event. The event was deliberately designed as stand-alone to facilitate access and engagement for busy academic staff during term time. The workshop aimed to train academics on best practices for designing effective prompts with text-generative AI tools to ensure quality outputs from ChatGPT and other similar large language model AI platforms. The workshop specifically looked at how ChatGPT might be used by students in current assessments and included open discussion with attendees on how assessments may need to be redesigned.

The workshop was delivered twice: face to face on campus (33 attendees), and in a hybrid format 9 attendees face to face and 107 online. Attendees were asked to undertake the following tasks in advance:

- Sign up for a free ChatGPT account with OpenAI
- Read and comply with OpenAI's terms and conditions, including being aware that anything they input may be used for future training of models
- Ensure they understand the university's guidance on the responsible use of AI, especially ensuring that they do not provide any personal or university business-sensitive data to OpenAI

- Identify two coursework assessments which they are comfortable being potentially used to train future AI models. The assessments could include essay titles, instructions, assessment briefs, or multiple-choice quiz questions.
- Ensure that they have a fully charged laptop with them (if attending on campus)

Before and at the start of the workshop, attendees were given the option to take part in this research project and links to consent forms were made available to all attendees.

DESIGN

Workshop Learning Objectives:

- Understand and apply LLM capabilities in assessment design.
- Enhance prompt writing skills among academic staff.

Workshop attendees who came in person were asked to sit with a partner from the same or closely related discipline to enable sense-checking of grades. All attendees were required to bring or have access to a laptop or mobile device in order to engage directly with ChatGPT for the hands-on tasks.

The workshop was structured and timed as follows:

1. *Pre-training*: Input the coursework details into ChatGPT (30 minutes):
    a. Collect responses from ChatGPT on two pieces of coursework
    b. Grade the responses from ChatGPT
        i. If sitting with a partner, ask them to independently grade the output from ChatGPT
2. Introduction to Prompt Engineering mini-lecture (30 minutes):
    a. Fundamentals of how large language models work
    b. Demonstration of effective prompting and how to refine them
    c. Chain of thought prompting
    d. Zero, one, and a few shot learning
    e. Guard rails in large language models and how to 'jailbreak' them

3. *Post-training*: Input the same coursework details into ChatGPT (30 minutes)
    a. Using refined prompts, collect responses from ChatGPT on the same two pieces of coursework
    b. Grade the responses from ChatGPT
        i. If sitting with a partner, ask them to independently grade the output from ChatGPT
4. General group discussion on ChatGPT's effectiveness in creating coursework (15 minutes)
5. Plenary and group reflection on implications for assessment design (15 minutes)

REFERENCE MATERIALS USED FOR PROMPT ENGINEERING WORKSHOP

While research on best practice in prompt engineering, and how to train in this skill, were still very much emergent at the time of the workshops, we drew upon a variety of resources from scholarship, the companies behind the platforms, and other educators. These materials included our own knowledge and experience using ChatGPT. One significant resource was OpenAI's online documentation on Prompt Engineering (OpenAI, 2023) and Chain-of-Thought Prompting (OpenAI Community, 2023). Additionally, several research papers, such as Liu et al. (2023), Lo (2023), and Wei et al. (2023), were particularly influential. Professor Mirella Lapata's talk at the Alan Turing Institute, titled 'What is Generative AI?' (Lapata, 2023), provided a clear overview of the fundamentals in simple terms, which we incorporated to explain the basics of LLMs to our colleagues. We also used some parts of online short courses, such as those by deeplearning.ai on ChatGPT Prompt Engineering for Developers (Fulford & Ng, 2023) in our workshop. These materials were gathered in the form of slides and handouts for workshop attendees.

METHODOLOGY

The data collection during the workshop utilised pre- and post-training assessment grading exercises and a qualitative self-reflection on the experience from

research participants. Additional data were also collected to ascertain the participants' prior confidence in using ChatGPT and their disciplinary context. The grading exercise was structured to yield quantitative data, allowing for analysis of 'before and after prompt training' patterns in academics' competencies in crafting prompts. The qualitative data collected were designed to reveal personal responses, feelings, and attitudes to how their experience of the workshop affected future assessment design; the plenary group discussion at the end of the workshop was designed to facilitate research participants in articulating these reflections though the plenary was not directly a source of data. These data were analysed for emergent themes on the future design of assessments at the university.

To allow for speedier grading, participants were asked to grade within a band (low fail/high fail/lower pass/middle pass/high pass/distinction) rather than a specific grade. The participants were asked to grade the content generated by ChatGPT in response to their prompts *as if it was coursework submitted by a student*.

FEEDBACK AND EVALUATION

Between both workshops, 149 people attended. Thirteen people in total chose to participate in the research data collection and completed the online form during or shortly after the workshop. The data presented here comprise 52 grading exercises, as each of the 13 research participants graded two ChatGPT-generated assessments twice, once before, once after prompt training. Participants shared that they found the workshop both helpful and interesting, providing verbal feedback during the session and through online chat. The discussion and plenary revealed a variety of reactions to the content produced by ChatGPT and thoughtful reflections on the implications for assessment design. Only the findings collected via the online survey tool are presented here, as this method received ethical approval for formal data collection; however, they are representative of the wider informal responses which participants shared during the workshop.

Most participants already felt moderately to very confident and experienced in using ChatGPT. However, even with the majority being familiar with the tool already, there was an average increase by nearly two grades after the participants had been trained in prompting (mean increase in grade band = 1.7).

Twelve out of 26 assessments graded had a higher grade after the participant training, with one assessment receiving a lower grade, and the rest (13) receiving the same grade. All but one of the assessments with an increased

grade band were undergraduate assessments, with one taught postgraduate assessment increased by a grade band. Participants who described themselves as under confident with AI were likely to give a higher grade to ChatGPT content after the training. However, the increased grade bands were spread among all confidence levels, with both confident and underconfident participants awarding large increases of 2, 3, or 4 grade bands to post-training content. In spite of the low numbers in the study, it can be speculated that prompt training can have a significant impact on staff competence in writing prompts, no matter their perceived competence.

The qualitative data revealed a mixed reaction to the variable results, with some participants viewing all assessments as potentially under threat to some specialised subject areas (e.g. Scots law) less likely to be convincingly portrayed in ChatGPT outputs. One participant found themselves torn:

> *This is a dangerous middle ground. It could be a potentially useful tool for planning assessment responses and research, but it isn't great for generating content with actual topic knowledge. However, it is also easy enough to use and learn that it offers a tempting shortcut to generating your own work.*

In terms of assessment design, essays, summaries, or reflections were identified by participants as no longer suitable as valid assessment types. Some were keen to continue to develop their own competences and the capabilities of paid-for AI models. Looking forward, all who responded indicated that future staff development and training were needed for both LLMs and assessment design in light of generative AI capabilities and developments.

TUTOR SELF-REFLECTION

During the plenary discussion, there was a significant variation in reactions from participants. Some described the shock of seeing their prompts elicit work which they would consider very good if submitted by students. On the other hand, a small number of participants felt their discipline required highly contextualised knowledge, which was less easily replicable with ChatGPT. While this level of difference could be expected, asking participants to discuss in multidisciplinary groups may have helped deepen understandings. There was also significant concern, given that we used the 'free' version of ChatGPT that participants would have liked to have had access to the paid-for versions, to see what responses the students who were accessing the premium services would get. It is also worth noting that the LLM model (ChatGPT

3.5 November to December 2023) has since been updated with a 'temporary chat' function, which means that users can opt out of their input being used for future training of models, and this would enable participants to provide more detailed information on the context of their assessments such as rubrics and university-specific instructions.

RECOMMENDATIONS

While this was a small study, it demonstrates a scalable approach to academic staff training and assessment design. Prompt literacy workshops can provide staff not just with competence but also confidence in discussing LLMs with students, with a view to incorporating AI literacy into the curriculum and assessments. Bringing together participants from a range of disciplines can present the opportunity for mutual knowledge sharing and a wider discussion about the impact of AI on higher education.

REFERENCES

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning and Teaching*, *6*(1), 57–63.

Fischer, J., Bearman, M., Boud, D., & Tai, J. (2024). How does assessment drive learning? A focus on students' development of evaluative judgement. *Assessment & Evaluation in Higher Education*, *49*(2), 233–245. https://doi.org/10.1080/02602938.2023.2206986

Fulford, I., & Ng, A. [DeepLearning.AI]. (2023). ChatGPT prompt engineering for developers: A short course from OpenAI and DeepLearning.AI [Video]. YouTube. https://youtu.be/H4YK_7MAckk?si=KhJIieOoGMu-pJ0y

Lapata, M. (2023, September 22). *What is generative AI?* [Video]. Alan Turing Institute. YouTube. https://youtu.be/fwaDtRbfioU?si=79j9YRglNNNq4jdN

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in

natural language processing. *ACM Computing Surveys*, *55*(9), Article 195, 35. https://doi.org/10.1145/3560815

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720

OpenAI. (2023). *Prompt engineering guide*. https://platform.openai.com/docs/guides/prompt-engineering

OpenAI Community. (2023). *How does chain-of-thought prompting work?* https://community.openai.com/t/how-does-chain-of-thought-prompting-work/198558

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv.Org. https://doi.org/10.48550/arxiv.2201.11903