Research Question

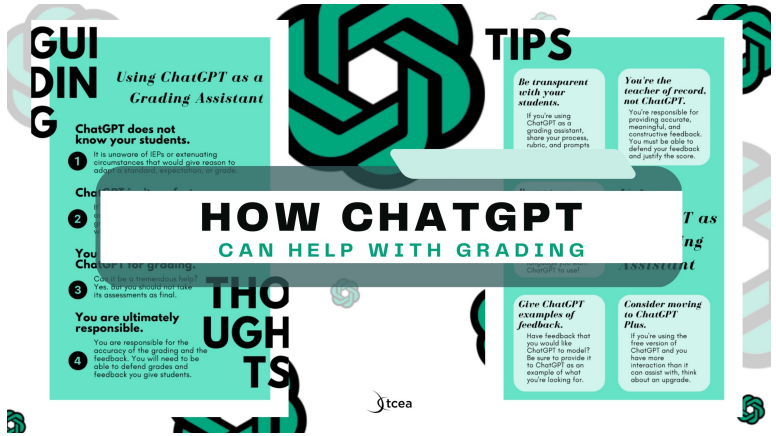
HOW MIGHT CHATGPT (LLM) WORK AS A TOOL TO SUPPORT IN THE GENERATION OF FEEDBACK FOR TEXT-BASED ASSIGNMENTS FOR MASTERS LEVEL STUDENTS AT UAL

SAM BARBER

ACTION RESEARCH PROJECT 2023

Why this topic?
Use of Generative AI is happening already in Higher Education

"Organisations that are trying to block people, or who say they are not ready are going to find their staff are doing it anyway - *but without the governance, ethics, security and intellectual property controls they might have had if they'd "allowed" them*!"[2]



Can large language models provide useful feedback on research papers? A large-scale empirical analysis

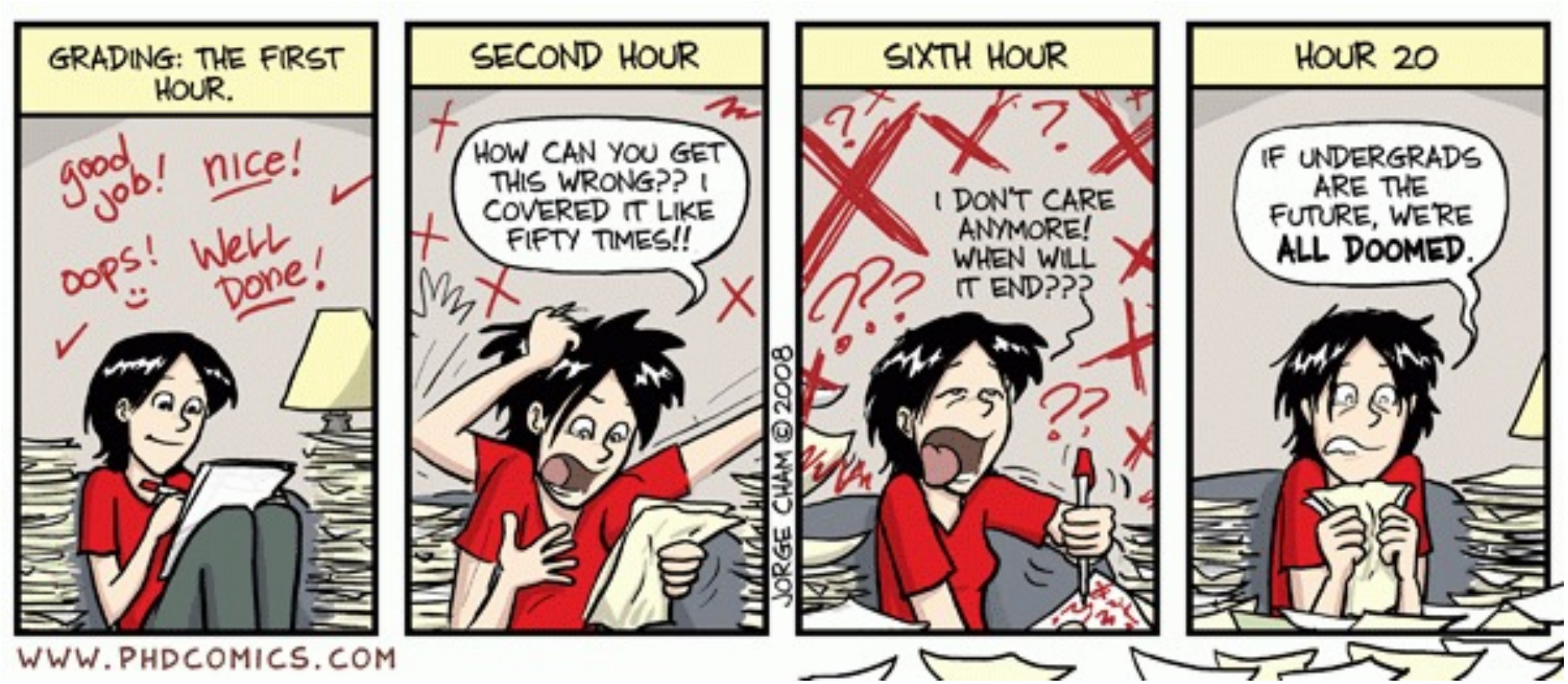
Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, James Zou

Expert feedback lays the foundation of rigorous research. However, the rapid growth of scholarly production and intricate knowledge specialization challenge the conventional scientific feedback mechanisms. High-quality peer reviews are increasingly difficult to obtain. Researchers who are more junior or from under-resourced settings have especially hard times getting timely feedback. With the breakthrough of large language models (LLM) such as GPT-4, there is growing interest in using LLMs to generate scientific feedback on research manuscripts. However, the utility of LLM-generated feedback has not been systematically studied. To address this gap, we created an automated pipeline using GPT-4 to provide comments on the full PDFs of scientific papers. We evaluated the quality of GPT-4's feedback through two large-scale studies. We first quantitatively compared GPT-4's generated feedback with human peer reviewer feedback in 15 Nature family journals (3,096 papers in total) and the ICLR machine learning conference (1,709 papers). The overlap in the points raised by GPT-4 and by human reviewers (average overlap 30.85% for Nature journals, 39.23% for ICLR) is comparable to the overlap between two human reviewers (average overlap 28.58% for Nature journals, 35.25% for ICLR). The overlap between GPT-4 and human reviewers is larger for the weaker papers. We then conducted a prospective user study with 308 researchers from 110 US institutions in the field of AI and computational biology to understand how researchers perceive feedback generated by our GPT-4 system on their own papers. Overall, more than half (57.4%) of the users found GPT-4 generated feedback helpful/very helpful and 82.4% found it more beneficial than feedback from at least some human reviewers. While our findings show that LLM-generated feedback can help researchers, we also identify several limitations.

Universities will support students and staff to become AI-literate.
(Russell Group Five Principles – No1)

[1] Kumar, R. (2023a) 'Faculty members' use of artificial intelligence to grade student papers: A case of implications', International Journal for Educational Integrity, 19(1). doi:10.1007/s40979-023-00130-7.

The social justice lens on feedback and marking



Tutors workload and capability to generate quality feedback can impact on student outcomes

- Students: Feedback quality[1]
  - Personalised to student and the course
  - Consistency across markers
  - Signposts where and how to improve
  - Motivates students
- Tutors: workload and capability. Can AI improve staff workload or capability in feedback generation

[1] *What makes good feedback* (no date) *What Makes Good Feedback | Learning and Teaching @ Newcastle | Newcastle University*. Available at: https://www.ncl.ac.uk/learning-and-teaching/effective-practice/assessment/good-feedback/ (Accessed: 14 January 2024).

The intervention

Qualitative Approach
Comparative marking exercise supported by semi-structured interviews

Comparative research techniques

Comparative thinking is 'one of our first and most natural modes of thought' [1]

Flexible and exploratory and enables movement within a topic.

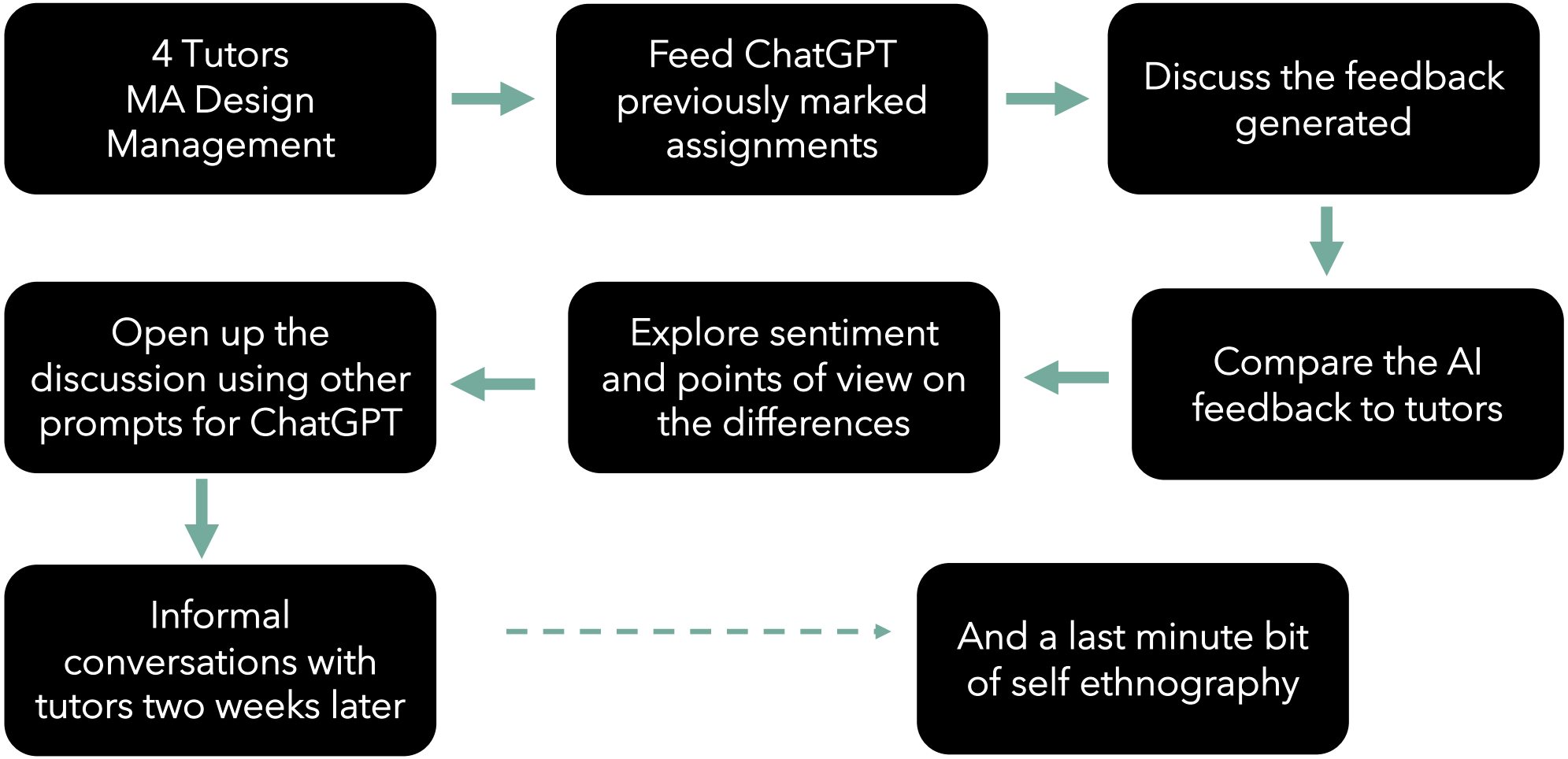
Semi-structured interviews

"the open-ended nature of the question defines the topic under investigation, but also provides opportunities for the interviewer and interviewee to discuss some topics in more detail"[2]

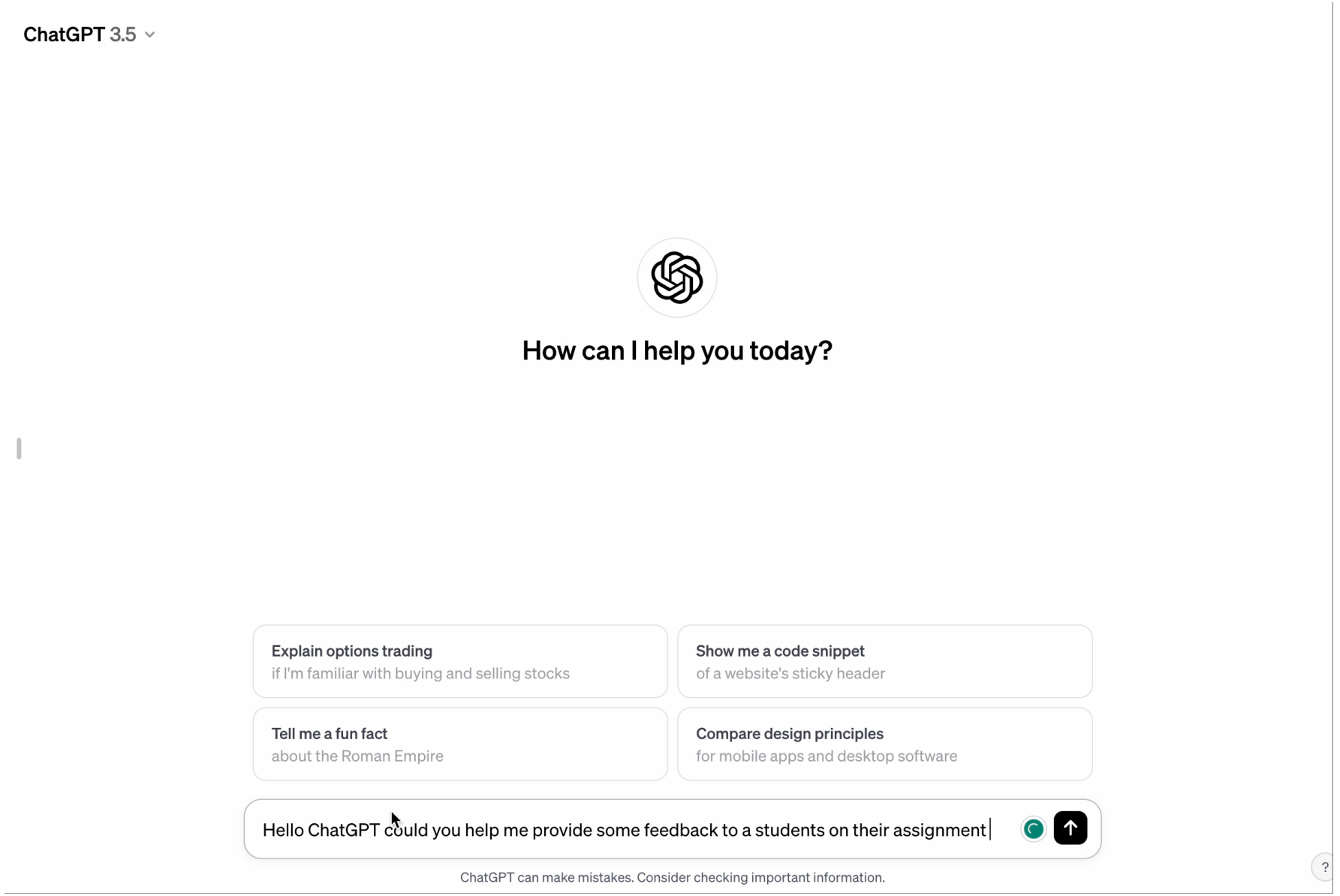
1. Silver, H.F. (2010) *Compare & Contrast: Teaching comparative thinking to strengthen student learning*. Alexandria, VA: Association for Supervision and Curriculum Development
2. Mathers, N.J., Fox, N.J. and Hunn, A. (1998) *Using interviews in a research project*. NHS Executive, Trent.

.

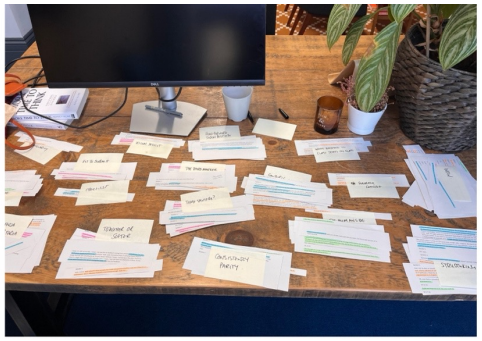
The Process

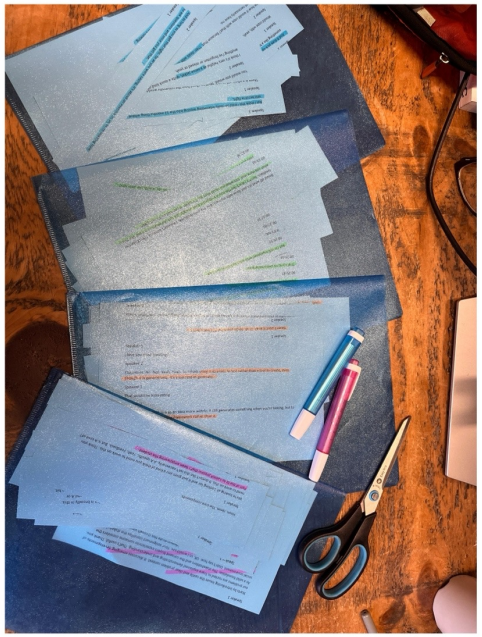

4 Tutors
MA Design
Management
Feed ChatGPT previously marked assignments
Discuss the feedback generated
Compare the AI feedback to tutors
Explore sentiment and points of view on the differences
Open up the discussion using other prompts for ChatGPT
Informal conversations with tutors two weeks later
And a last minute bit of self ethnography

How this worked in practice



Analysis Approach





Record interviews using App

Transcribe via Word on the web

Print out

Read and highlight key areas of insight

Cluster by theme

Look for differences and clusters within the themes

Interview 1 - Pink

Interview 2 - Blue

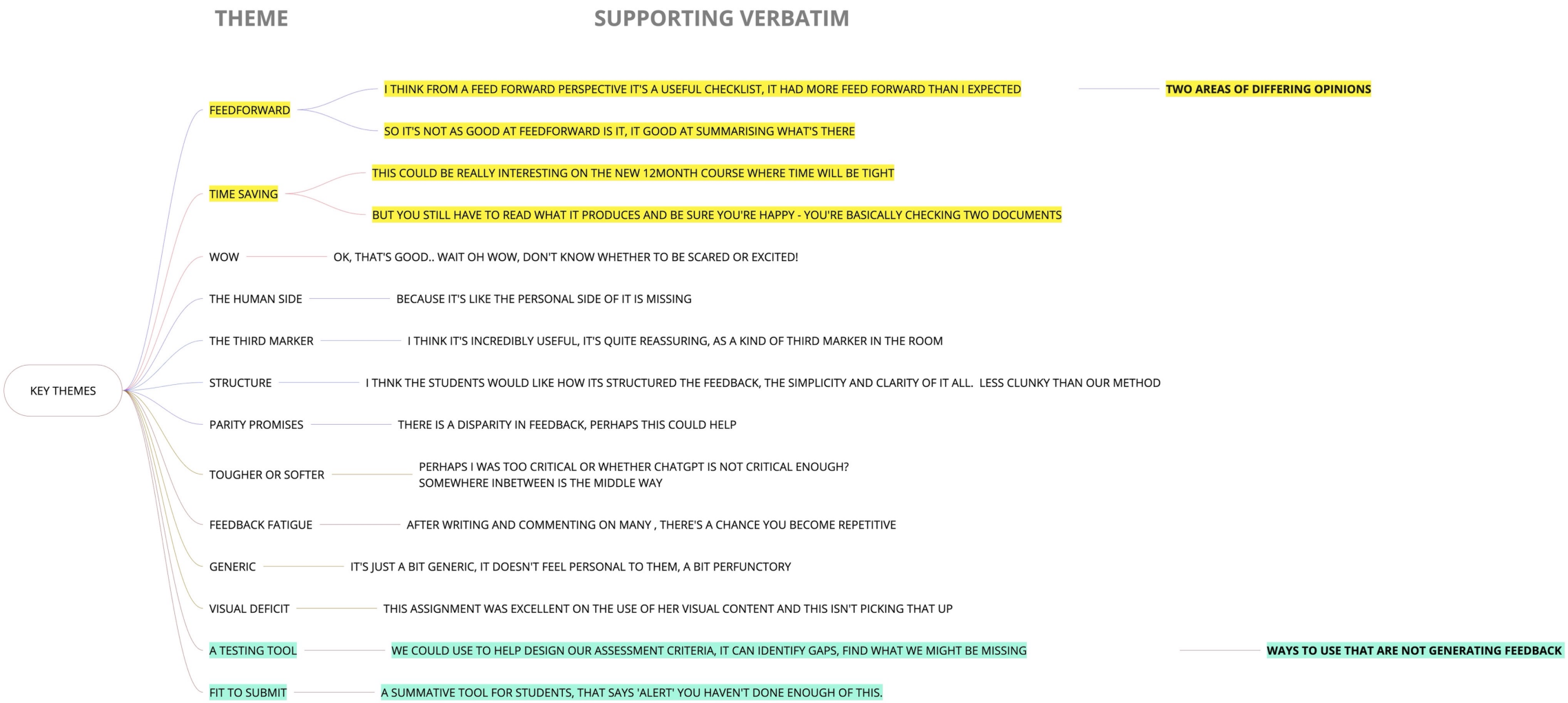
Interview 3 - Green

Interview 4 - Orange

Clusters and Findings

13 topics for further investigation

Mind-map to cluster findings supported by quotes



Looking for different angles on the findings Restructured in terms of positive to negative

Thoughts and observations….in addition to the themes identified

Teaching staff want to explore and test new technologies "I stopped at Grammarly"

AI may be able to play a role in building tutor skills and capabilities around feedback generation

Could this comparative marking intervention help to up-skill teaching staff's ability and familiarity with AI

AI may be able to play a role in testing our assessment criteria and helping to identify gaps – the third marker in the room

The messy nature of research

Action Research
A sense of the approach