A critical analysis of the methodologies of two research papers investigating the impact of large language models in an educational context

Nicholas Johnson

January 2024

1 Introduction

Since the advent of the Transformer Architecture in 2017, Large Language Models (LLMs) have grown enormously in power and scope [1]. Unlike an RNN, which is inherently sequential [2], a Transformer has a parallel internal architecture that can be trained quickly using a modern GPU [3]. Given that a transformer receives its entire context window at once rather than sequentially, transformers do not suffer from the vanishing gradient problem that previously limited the power of RNNs [4, 5].

It is hard to overstate the impact that deep learning could have on all areas of human activity [6]. Significant advances have recently been made in the fields of biochemistry [7], materials science [8], image and video generation [9, 10], medicine [11], and robot control [12], to name a few.

It seems plausible that language models such as ChatGPT will continue to push this frontier. Evidence points to the potential for massive increases in worker productivity, on a par with the Industrial Revolution [6]. The societal impacts of this increase remain to be seen.

In this report, we will analyse and contrast two research papers that seek to investigate the impact of ChatGPT specifically on academia. We will investigate their methods and methodologies, and how these methodologies support their hypotheses. We will conclude with a discussion of qualitative and quantitative methodologies, making reference to both Mertens [13] and Allwood [14].

2 Task 1 - Evaluation of Both Papers

Two papers are presented. "ChatGPT - Boon or Bane" (paper one) [15] conducts a simple quantitive analysis correlating awareness of ChatGPT amongst students to field of study and gender. In contrast, "Exploring Students' Perceptions of ChatGPT" (paper two) [16] adopts a more sophisticated two-stage mixed methods analysis.

Both these papers seek to investigate student engagement with ChatGPT, but each adopts a different methodology and has different research goals. We will compare and contrast these two papers, their respective hypotheses, methodologies and methods, then propose an extension for paper one. We shall conclude with a discussion of qualitative vs. quantitative methodologies in the context of research.

Paper one proposes the following research goals:

1. "Empirically analyse the awareness regarding ChatGPT among the students."

2. "Highlighting the opportunities and challenges of ChatGPT." [15]

To support these goals, the paper proposes the following null hypotheses:

1. "There is no significant difference in the awareness level of ChatGPT based on gender."

2. "There is no significant difference in the awareness level of ChatGPT based on the field of study." [15]

Both these hypotheses support the first research goal. No hypotheses are proposed to support the second research goal, and this could be considered a weakness of the paper.

These hypotheses are interesting in that they propose that there is no correlation between the various pieces of data that were collected in the survey. In other words, all genders and fields of study have a similar awareness of ChatGPT.

To test these hypotheses, the authors used a simple quantitative methodology. They distributed an online survey via email, looking for correlations between demographics and the awareness of ChatGPT among students.

The method used in paper one aligns well with its first research goal, and also with its hypotheses. This quantitative approach is an appropriate way of

gauging awareness levels among a large group of students. The method allows for the collection of measurable data that can be statistically analysed [17].

Independent sample t-tests are suitable for evaluating hypothesis one, that there is no correlation between awareness of ChatGPT and on gender and field of study [18].

However, the study's reliance on self-reported data from a single method could limit the depth of understanding. The results might be influenced by self-selection bias, as the respondents chose to participate; perhaps students already engaged with ChatGPT might be more interested in responding. The study's findings might not be generalisable beyond the specific demographic of college students who participated.

The limited scope of the questionnaire curtails the types of conclusions that can be drawn from it; the results are limited to discussions of gender and field of study. The paper also makes very little attempt to tackle the second research question it proposes, relying exclusively on secondary sources. No primary data is presented to answer the headline question "boon or bane" and this also could be considered a weakness.

In contrast, paper two adopts a mixed methodology. A two-phased approach is employed to gather both quantitative and qualitative data.

In phase one, the author conducts a thematic analysis. Qualitative data, collected through an open-ended question posed to students after completing a learning activity, was analysed using Taguette [19] to identify patterns or themes. These themes were used to understand students' perceptions of ChatGPT.

In phase two, the author conducts a follow-up survey. A questionnaire was developed based on the themes discovered during phase one. Students were required to indicate their level of agreement with various statements related to ChatGPT. This survey aimed to produce quantitative data to complement the earlier qualitative stage.

The paper proposes two research questions:

1. "How do students perceive ChatGPT in the context of learning?"
2. "What are ChatGPT's pros and cons from students' perspectives?" [16]

The paper does not explicitly propose a hypothesis. However, reading between the lines, a hypothesis could be something like:

1. Students' perceptions of ChatGPT in educational settings are predominantly positive.

2. Students have notable concerns regarding the accuracy, ethical implications, and dependency on AI for educational purposes.

This two-stage method effectively addresses these hypotheses. Initially, the thematic analysis of students' written responses provides qualitative insights into their perceptions of ChatGPT, both positive and negative. This initial approach allows the follow-up study to be tailored specifically to the subject's lived experiences [13].

The quantitative data gathered in the second stage offers a more systematic view of student opinions and enables the statistical identification of correlations. This combination of qualitative and quantitative methodologies ensures the questions asked conform to the subject's experiences. The qualitative phase informs and directs the quantitative phase [20].

These two papers tackle similar themes but in different ways. Both papers use primary sources, but paper one adopts a simple quantitative methodology compared to paper two's mixed methods.

Paper one is somewhat limited in scope; two pieces of demographic data are collected and then correlated against usage of ChatGPT. On the other hand, paper two's usage of an initial open-ended question allows for the development of a more interesting second stage that better reflects the experiences of the experimental participants. This allows for a more nuanced analysis. For this reason, the approach adopted in paper two could be considered more robust and mature.

3 Task 2 - Recommendations for one paper

As mentioned above, paper one sets out two objectives:

1. "Empirically analyse the awareness regarding ChatGPT among the students."

2. "Highlighting the opportunities and challenges of ChatGPT." [16]

The paper investigates objective one with two hypotheses (see above) and a survey but fails to tackle objective two with any primary research.

An interesting way to tackle the second objective might be an impact analysis combined with a thematic longitudinal study.

An academic test could be carried out before and after the introduction of an LLM into an organisation. Quantitative metrics could be collected, which might include grades or the ability to complete a particular task without the use of the LLM. The delta between these two tests could be correlated with total engagement with the technology.

This analysis could be complemented with a longitudinal study that would attempt to extract qualitative data about the specific ways students interact with the LLM. This qualitative data could be subjected to thematic analysis [19], allowing further insights to be gleaned as to the specific ways in which large language models might be beneficial or harmful.

3.1 Refined Research Question

A suitable research question for this extension might be:

"How does the integration of ChatGPT into educational curricula affect the academic performance and engagement levels of students in higher education institutions?"

This question builds upon the second objective of paper one but narrows the scope specifically to performance and engagement in education. Rather than looking for "opportunities and challenges" [15], we are looking specifically at performance and engagement, which are directly measurable.

3.2 Refined Hypothesis

A refined hypothesis for this research question could be:

"The integration of ChatGPT into educational curricula significantly improves the academic performance and engagement levels of students in higher education institutions."

This hypothesis is directly testable and sets a clear expectation: that the use of ChatGPT in education positively affects student outcomes. It can be validated or refuted through empirical research and statistical analysis. No hypothesis for objective two was presented in the original paper.

3.3 Experimental Design

A four-stage mixed methodology experimental design is proposed:

1. Pre-Test - an academic test carried out without the use of an LLM.

2. Implementation of ChatGPT.

3. Longitudinal Monitoring over one academic year.

4. Post-Test - At the end of the period, conduct the same assessment as in the pre-test.

Participants will be divided into a test group and a control group. A pre-test and post-test will be administered one year apart. This will seek to provide quantitative data showing change in academic performance.

In addition, qualitative data will be collected during the experiment by means of an open-ended survey. This will be subject to thematic analysis that can be correlated with changes in academic performance.

3.4 Subject Selection

In the proposed study, participants will be selected to form two distinct groups: an experimental group that will engage with ChatGPT and a control group that will not. The selection process will employ stratified sampling techniques to ensure equal representation of relevant demographics [13].

All subjects will be selected from the same year group and academic subject. This will control for variables such as differences in academic assessment techniques or teaching methods and allow the pre and post-tests to be standardised, producing cleaner quantitative data. This will somewhat limit the scope of the study, and further work may be required at a later date.

Participant loss may be an issue, specifically amongst the control group.

3.5 Pre-Test

Prior to the intervention, a comprehensive pre-test will be conducted. This preliminary assessment will involve standardised academic performance metrics. These instruments will be administered to both the experimental and control groups, ensuring comparability of data.

3.6 Implementation phase

During the implementation phase, ChatGPT will be introduced to the experimental group. This phase will involve a structured rollout of the technology.

Participants will be encouraged to integrate ChatGPT into various aspects of their learning, such as research, writing, and problem-solving.

3.7 Longitudinal monitoring

During the monitoring phase, spanning one academic year, data will be collected to track the experiences of the experimental group. Monitoring will consist of an open-ended survey, administered once each academic term, to gauge the types of interactions each subject has with the technology.

This qualitative data will be subjected to later thematic analysis via Taguette [19].

3.8 Post-test

At the conclusion of the academic year, a post-test assessment will be conducted. This phase will mirror the pre-test, employing the same academic performance metrics to evaluate changes.

At this time, raw data will be gathered from the ChatGPT API, including total tokens generated per subject.

3.9 Analysis

Paired t-tests will be used to compare the pre and post-intervention data within and between groups. This analysis will ascertain whether significant changes in academic performance and engagement are attributable to the use of ChatGPT.

In addition, qualitative data from the longitudinal study and interviews will be subjected to thematic analysis to extract underlying themes about each subject's use of the technology. The combination of quantitative and qualitative data will provide a holistic understanding of ChatGPT's impact [17].

By correlating ChatGPT usage stats against the delta in academic performance, it may be possible to determine whether ChatGPT is beneficial or harmful to student success.

By correlating delta academic performance against the thematic analysis of the longitudinal study, it may be possible to determine precisely which patterns of usage are beneficial and which are harmful. Since this data is qualitative, it may yield unexpected surprises.

The comparison of pre and post-test results will provide quantitative data to assess the impact of ChatGPT, while qualitative feedback gathered through

open-ended survey questions may provide insights into which specific patterns of use create the largest delta.

4 Task 3: Discussion of key characteristics

Merriam defines qualitative research as a process of discovering the meaning that subjects find in the world. Whereas quantitative research imposes the researcher's ideas onto the subject, channelling the subject's responses down specific predefined paths, qualitative research inverts the direction of control, allowing the subject to guide the researcher towards questions that more closely match the subject's worldview [20].

Qualitative research is investigative. It seeks to ask open-ended questions. It brings the researcher towards the subject, rather than channelling the subject into a framework defined by the researcher. It represents a split between Positivism, where reality is objective, and Interpretivism, where reality is internally experienced and socially constructed [21]. As Lather points out, "There is a context from which I speak" [22].

As Mertens says, the divide between quantitative and qualitative can have implications in terms of social justice. The subject's individualised ontology is respected [13].

However, this presents significant challenges when interpreting the results. Data collected in this way is not standardised. Different subjects may approach questions in incompatible ways. Statistical tools like ANOVA are unlikely to apply [20].

Queros identifies seven types of qualitative research including observation, ethnography, field research, focus groups, case studies, structured interviews and in depth interviews [23]. Other examples of qualitative research methods might include open-ended surveys, and even non-traditional methods such as painting or drawing [20].

Qualitative techniques are particularly appropriate for introductory exploration of the research question, when it might not be clear exactly what specific data to collect [20]. They are particularly strong when dealing with diverse subjects who may not share the same worldview as the experimenter, or when conducting social science [13]. They may not be as appropriate as a tool to produce hard data. Although hard data can be extracted from them, this may not be of sufficient quality to be subjected to statistical analysis [14].

To give a concrete example, in "Programming Languages, improvements,

popularity and the need of the future", Norlin employs an initial qualitative interview phase to get a sense of the potentially unknown reasons developers have for preferring one language over another. This complements a later quantitative phase [24].

In contrast, quantitative research is standardised. Variables are controlled for. Data generated using quantitative methodologies is hard data [14], compatible with various statistical techniques. However, it may be the case that the questions the researcher asks are not the questions the subject would choose to respond to [13].

We see this clearly in paper one, where the researchers attempted to correlate gender and field of study against awareness of LLMs and found no correlation. The researchers decided on their metrics in advance and picked a metric that yielded no particularly interesting results [15].

By contrast, in paper two, the author employs an initial qualitative phase to inform the direction of a secondary quantitative phase, uncovering correlations that may not have been obvious without this first undirected phase. In this case, mixed methods provide the best of both worlds [16].

Queros identifies five types of quantitative research: Field experiments, simulation, surveys, correlation studies and multivariate analysis [23]. Anything that yields a hard numeric or Boolean response could be considered quantitative [14].

As we see in paper two, thematic analysis can effectively convert qualitative data into quantitative data, discovering common themes in free-form responses. A tool such as Taguette [19] can be used to tag portions of a text response. The frequency of themes and correlations between themes can be treated as quantitative. However, this is typically a manual process that involves a good deal of subjectivity [14].

The distinction between methods is not always clear-cut. Allwood highlights the heterogeneity of qualitative approaches in research. He points out that these methods exhibit significant variation in aspects such as their generalizability, epistemological perspectives, structural characteristics, the extent of interest in discovering regularities, use of quantification, and approaches to causal explanation. This could certainly be considered a weakness. This diverse nature of qualitative methods challenges the notion of a clear-cut distinction between qualitative and quantitative research, underscoring the complexity and overlap inherent in various research methodologies [14].

Given a bucket of water, it is possible to measure the exact amount of water in that bucket; its temperature; the composition of the metal. However, that bucket may have different meanings to different people in different contexts. It

may be a drink; a burden; a leaking roof.

Referring back to the original papers [15,16], it should be clear that both these viewpoints are valid. When seeking to answer the question "boon or bane" [15], one must determine not only the objective effects of the technology but also how users subjectively define those words. This ontological dichotomy is perhaps not a dichotomy at all.

5 Conclusion

In this report, we have critically analysed two papers with similar objectives but very different methodologies. We have seen how the mixed methodology employed by paper two allows for deeper insights into opinions and usage patterns amongst students.

We have offered an alternative research design that builds upon the questions posed by paper one, incorporating a mixed methodology that incorporates some of the strengths of paper two, while providing controlled quantitative data in the form of an academic test.

Finally, we have discussed the differences between qualitative and quantitative data, specifically looking at how qualitative data can relate the experiment to the subject's individualised worldview [13], and discussing the heterogeneous nature of qualitative data [14].

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020.

[2] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[4] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, pp. 107–116, 04 1998.

[5] O. Kuchaiev and B. Ginsburg, "Factorization tricks for lstm networks," 2018.

[6] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187–192, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.adh2586

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021. [Online]. Available: https://doi.org/10.1038/s41586-021-03819-2

[8] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, "Scaling deep learning for materials discovery," *Nature*, vol. 624, no. 7990, pp. 80–85, Dec 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06735-9

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[10] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," 2023.

[11] F. Wong, E. J. Zheng, J. A. Valeri, N. M. Donghia, M. N. Anahtar, S. Omori, A. Li, A. Cubillos-Ruiz, A. Krishnan, W. Jin, A. L. Manson, J. Friedrichs, R. Helbig, B. Hajian, D. K. Fiejtek, F. F. Wagner, H. H. Soutter, A. M. Earl, J. M. Stokes, L. D. Renner, and J. J. Collins, "Discovery of a structural class of antibiotics with explainable deep learning," *Nature*, 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06887-8

[12] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch, R. Diethelm, S. Bachmann, A. Melzer, and M. Hoepflinger, "Anymal - a highly mobile and dynamic quadrupedal robot," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 38–44.

[13] D. Mertens, "Transformative paradigm: Mixed methods and social justice," *Journal of Mixed Methods Research*, vol. 1, pp. 212–225, 07 2007.

[14] C. M. Allwood, "The distinction between qualitative and quantitative research methods is problematic," *Quality & Quantity - QUAL QUANT*, vol. 46, pp. 1–13, 08 2011.

[15] S. Wagholikar, A. Chandani, R. Atiq, M. Pathak, and O. Wagholikar, "Chatgpt -boon or bane: A study from students perspective," in *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 2023, pp. 207–212.

[16] A. Shoufan, "Exploring students' perceptions of chatgpt: Thematic analysis and follow-up survey," *IEEE Access*, vol. 11, pp. 38 805–38 818, 2023.

[17] J. W. Creswell, *Research Design - Qualitative, Quantitative, and Mixed Methods Approaches*. Los Angeles: sage, 2022.

[18] H.-Y. Kim, "Statistical notes for clinical researchers: the independent samples t-test," *Restor Dent Endod*, vol. 44, no. 3, p. e26, Jul. 2019.

[19] R. Rampin and V. Rampin, "Taguette: open-source qualitative data analysis," *Journal of Open Source Software*, vol. 6, p. 3522, 12 2021.

[20] S. B. Merriam and E. J. Tisdell, *Qualitative Research: A Guide to Design and Implementation, 4th Edition*. San Francisco: Jossey-Bass, 2015.

[21] M. Tombs and L. Pugsley, "How to: Understand research philosophies and paradigms in medical education," 2020.

[22] P. Lather, "Issues of validity in openly ideological research: Between a rock and a soft place," *Interchange*, vol. 17, pp. 63–84, 12 1986.

[23] F. Almeida, D. Faria, and A. Queirós, "Strengths and limitations of qualitative and quantitative research methods," *European Journal of Education Studies*, vol. 3, pp. 369–387, 09 2017.

[24] N. Philip and W. Valentin, "Programming languages : Improvements, popularity, and the need of the future," p. 40, 2018.