CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

The performance of transformer-based auto-markers on science content: a scoping review

Conference Abstract

Frank Morley
Emma Walland
Carmen Vidal Rodeiro

AEA-Europe conference, 6-9 November 2024

Author contact details:

Frank Morley, Emma Walland & Carmen Vidal Rodeiro
Assessment Research and Development
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

frank.morley@cambridge.org
emma.walland@cambridge.org
carmen.vidalrodeiro@cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division
If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Morley, F., Walland, E. & Vidal Rodeiro, C.L. (2024, November 6-9). *The performance of transformer-based auto-markers on science content: a scoping review*. [Conference presentation]. AEA-Europe 2024, Cyprus. https://2024.aea-europe.net/

Abstract

Background

Automatic marking of student responses in science has traditionally been difficult due to the "myriad and sometimes unconventional ways in which credit-worthy answers [are] expressed" (Sukkarieh, Pulman, and Raikes, 2005, p.19). Recent developments since a seminal paper on artificial intelligence in 2017, which outlined a new method called 'the transformer' (Vaswani et al., 2017), have led to a variety of new 'large language models' (LLMs), promising state-of-the-art performance in natural language processing.

A review in 2022 found that BERT had the highest performance of any auto-marker on a widely used dataset (Kusuma et al., 2022). Since November 2022, when ChatGPT was released, there has been a surge of interest in LLMs. However, criticisms have been made of modern auto-markers: most do not assess domain knowledge (Ramesh and Sanampudi, 2022), neither do many prioritise explainability (Huawei and Aryadoust, 2022). This review investigated these gaps, by providing up-to-date insights into transformer-based auto-marking, focusing on domain knowledge and evaluation using additional metrics such as explainability.

Research aim

Our review summarised recent research on auto-markers for science questions (excluding fully objective items such as multiple choice). We focused on assessing subject content rather than writing quality.

We evaluated the performance of auto-markers against both quantitative and qualitative metrics. Quantitative metrics included Quadratic Weighted Kappa, Cohen's Kappa and mean mark error. Qualitative metrics included feedback, transparency, explainability, and ethics.

Theoretical underpinning and policy implications

The review is underpinned by theories comparing how examiners and auto-markers arrive at their mark. How these two forms of marking differ, and our ability (or lack thereof) to peer inside the 'black box' of transformer-based auto-markers, has real policy implications if these systems are implemented.

Method

We followed a rigorous scoping review methodology (Tricco, 2016) to search for, evaluate and summarise relevant literature. We used MAXQDA to support the coding and organisation of the literature. We focused on peer-reviewed research (including conference papers), obtained from Scopus, whilst also including grey literature with caution. We performed reference list scanning on all articles. Our search terms were: (auto*) AND (essay OR writing OR "short answer" OR item) AND (grading OR scoring OR evaluation OR marking) AND (GPT* OR BERT* OR transformer* OR LLM* OR "large language model"). We restricted the results to 2017 and later, and articles written in English. The initial abstracts were screened against our inclusion and exclusion criteria, which were as follows:

- Criteria 1: Auto-marking in an educational assessment context.
- Criteria 2: Natural language responses (i.e., not fully objective items).
- Criteria 3: English language (i.e., not other languages).

- Criteria 4: Non-language domain knowledge (i.e., subject-related knowledge and understanding rather than the quality of writing).
- Criteria 5: Primary study (i.e., not reviews).
- Criteria 6: Science subjects (i.e., natural sciences).

We used the PRISMA checklist to ensure a rigorous review process was followed and reported (Page et al., 2021).

Results

Our initial Scopus search yielded 252 abstracts, leading to 20 studies being reviewed in depth. We evaluated the studies according to areas such as region, educational setting, item type, technologies used, training datasets, and auto-marker performance.

Conclusion

Our review provides a resource for researchers, data scientists and assessment practitioners, particularly in the science domain, about state-of-the-art developments in auto-marking. Our work provides an overview to inform subsequent research and identifies challenges and opportunities associated with the latest auto-marking technologies. While there have been major advances in auto-marking, there are still challenges associated with marking science questions in educational assessment, which our review highlights.

References

Huawei, S., & Aryadoust, V. (2022). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, (pp. 771–795).

Kusuma, J. S., Halim, K., Pranoto, E. J., Kanigoro, B., & Irwansyah, E. (2022). Automated Essay Scoring Using Machine Learning. *4th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-5). Prapat, Indonesia: Institute of Electrical and Electronics Engineers (IEEE).

Page, M. J., McKenzie, J. E., Bossuyt, P.M., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10, 89.

Ramesh, D., & Sanampudi, S. K. (2022). Automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, (pp. 2495–2527).

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2005). Automatic marking of short, free text responses. In *Research Matters: Issue 1* (pp. 19-22).

Tricco, A. C. (2016). Scoping reviews: What are they and how you can do them. Retrieved from https://training.cochrane.org/sites/training.cochrane.org/files/public/uploads/resources/downloadable_resources/English/Scoping%20reviews%20webinar%20Andrea%20Tricco%20PDF.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008).