Not like that! Attempting to use GPT to generate test items in statistics

Inspired by Brigg's suggestions at AEA 2022 about learning progressions, a potential use for AI might be to help teachers to generate formative test items or graded worked examples which might assist students to construct models or schema.

Author

Imogen Casebourne

Affiliation

DEFI (Digital Education Futures Initiative), Hughes Hall, University of Cambridge.

ic407@hughes.cam.ac.uk

Introduction

Generative AI can be used to produce test items with humans in the loop, as with Duolingo DET. This poster explores the potential additional use of AI by teachers to create formative test items or graded examples, along with commentary on how to improve lower-scoring examples, offering insight not just into what is right, but also what is wrong and why. Students often only see perfect examples and miss out on learning from mistakes. However, this assumes that generative AI can reliably produce test items. This is not always the case!

Let's start with a T-test...

GPT 3.5

Produced a data set, explained the T-test and at first produced an apparently convincing marking rubric BUT data set too small for understanding effect size and it miscalculated the mean.

GPT 4 & Bing

Still predisposed to an overly small dataset, still can't add up. But now can visually represent equations. Bing's ability to search internet no help.

For Group A:

$$s_A^2 = \frac{\Sigma(x_i - \bar{x}_A)^2}{n_A - 1}$$

GPT 4 + Wolfram Alpha

Able to perform without basic arithmetic errors.

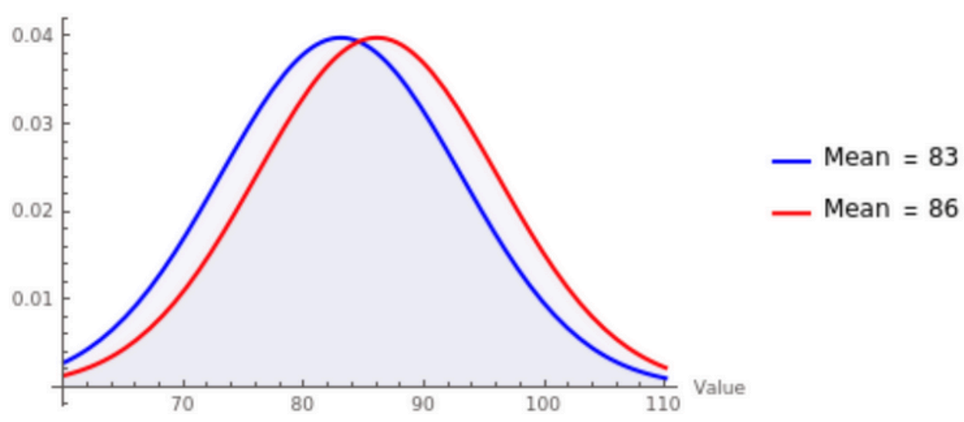
Analysis

Over 3 attempts with each system, with prompting, the Large Language Models (LLMs) got better at generating the text for worked examples and prompts, but I was not successful in prompting them to get better at maths.

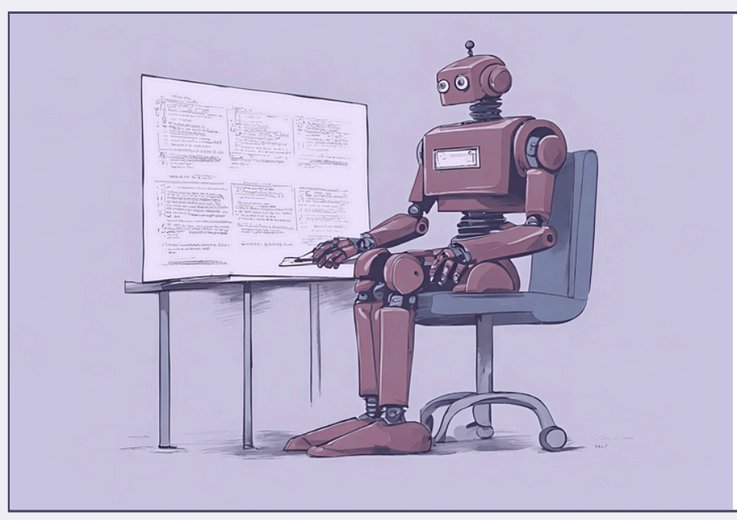
The calculation issues could stem from the models' training on text discussing statistics, but not on performing the calculations themselves. Additionally, higher creativity settings in Bing may lead to less predictable, and therefore more likely incorrect, answers in mathematical contexts. And I may not be the best prompter for statistics!

Wolfram Alpha is also an AI based model, but not a LLM and has a better representation of how to perform calculations. In partnership with Wolfram Alpha, the systems become more reliable.

Note that recent advances in LLMs means it is now possible to upload visual examples and have them work on them - which will be a problem for printed worksheets.



Indicative amount by which GPT can be off without assistance from WolframAlpha



Conclusion

For generative AI, it may be helpful to distinguish 'well-structured domains' (Jonassen, 1997), such as certain areas of maths, science, and engineering that require explicit principles and rules, from other types of knowledge. For well-structured domains, generalised LLMs fall short compared to more explicit GOFAI-based systems or systems such as Wolfram Alpha and need supplmenting.

Related Literature

Briggs, Derek, and Frederick Peck. (2015) 'Using Learning Progressions to Design Vertical Scales That Support Coherent Inferences about Student Growth'. Measurement: Interdisciplinary Research and Perspectives 13): 75–99. https://doi.org/10.1080/15366367.2015.1042814.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. Educational Technology Research and Development, 45(1), 65-94.

Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (2021). Enhancing students' critical thinking skills: is comparing correct and erroneous examples beneficial? Instructional Science, 49, 747-777. https://doi.org/10.1007/s11251-021-09559-0

Wodzak, S. (2022, April 6). Can a standardized test actually write itself? Duolingo Blog. Retrieved from [Duolingo Blog]