15

Automatic Short Answer Grading in Health Sciences with ChatGPT

PhD Nuria Padros-Flores
Universidad Miguel Hernández, Spain
npadros@umh.es
https://orcid.org/0000-0001-5206-8857

Ivan Gadea Sáez
Universidad de Alicante, Spain
ivan.gadea@ua.es
https://orcid.org/0009-0009-2290-3220

PhD Carolina Alonso-Montero
Universidad Miguel Hernández, Spain
c.alonso@umh.es
https://orcid.org/0000-0002-8856-1907

Abstract
Artificial Intelligence (AI) has emerged as a transformative tool in education, notably in facilitating automated exam grading. This study focuses on Automatic Short Answer Grading (ASAG) via ChatGPT-4, a widely accessible and versatile general-purpose generative AI model. We compare the grading outcomes from ChatGPT with those adjudicated by human evaluators within the health science domain. An evaluative framework was deployed to gauge the GPT-4 model's concordance with an expert educator's scoring. Human scores were compared to those offered by ChatGPT with different versions of prompts, specifically with 10 examples, 25 examples, and a grading rubric, employing a scoring metric that spans from 0 to 10 points, allowing for decimal values, without any model fine-tuning or parameter modulation. Our findings show that rubrics markedly enhance score alignment with an educator's evaluative benchmarks, registering intraclass correlation coefficients surpassing 0.8, thus nearly mirroring human judgment. These results suggest that there is ample scope for increasing the effectiveness of ASAG using Large Language Models

(LLM) such as ChatGPT. However, it is imperative to recognize that the operability of these systems is not yet fully reliable and stable, making human supervision necessary. The integration of expert supervision ensures both the accuracy and pedagogical validity of these automated tools.

Keywords: AI, ASAG, ChatGPT, health sciences, large language models.

15.1. Introduction

The emergence of ChatGPT as a universally accessible tool has popularized terms such as Large Language Model (LLM) and generative Artificial Intelligence (AI), enhancing public familiarity with these technologies (Leiter et al., 2023; Taecharungroj, 2023). LLMs are advanced AI systems capable of understanding and generating human-like text from the vast datasets on which they have been trained. This subset of generative AI technologies specializes in the production of coherent and contextually relevant content. In educational contexts, these models provide innovative approaches for the generation of dynamic learning materials and the delivery of personalized feedback. Other AI applications that are not aimed at content generation are focused on data analytics, predictive modelling and automation of task execution. Collectively, these diverse roles significantly contribute to the enhancement of teaching and learning experiences (Chen et al., 2020).

Recognizing the transformative potential of LLM in educational contexts, it is critical to address the dual-sided nature of their integration. Concerns such as preserving human-centric learning experiences, ensuring academic integrity, and managing copyright issues present significant challenges in an AI-enhanced learning environment (Ifenthaler & Schumacher, 2023; Preiksaitis & Rose, 2023). However, the unique capabilities of LLMs to generate contextually relevant and coherent text provide unprecedented opportunities for personalizing learning experiences, developing educational content, and providing automated feedback to students (Zawacki-Richter et al., 2019). Additionally, AI's incorporation into education promises to spur pedagogical innovation and enhance access to learning opportunities, particularly in geographically isolated or socioeconomically disadvantaged areas (Pacchiega, 2021).

The release of the GPT-3 model (Generative Pre-trained Transformer) in 2020 marked a significant advancement in AI research, although it was ChatGPT 3.5, launched towards the end of 2022, that really caught the public's attention with its greater accessibility and user-friendly interface. The development of GPT-3 incorporated reinforcement learning with human feedback, facilitating the creation of a powerful chatbot capable of understanding and generating responses to natural language prompts with unprecedented ease (Wu et al., 2023). Research on prompts soon began, uncovering that certain prompts work better than others in achieving specific responses (Cain, 2024; Henrickson & Meroño-Peñuela, 2023; Lee et al., 2023). The introduction of ChatGPT-4 further advanced the field by incorporating the ability to generate and analyze images. This enhancement established ChatGPT-4 as a leading chatbot with multimodal capabilities, pushing the boundaries towards achieving artificial general intelligence (AGI) (Wu et al., 2023).

The ease of use of these new generative AI tools has raised concerns among educators, particularly regarding the ease with which students can generate texts. Conversely, these technologies also present new opportunities for the automated assessment of exams and assignments. In educational settings, teachers often rely on various question types to evaluate student understanding, from multiple-choice questions, which can be automatically graded by specialized hardware, to short open-ended questions and essays that require more nuanced assessment. Specifically, Automatic Short Answer Grading (ASAG) is a field that has been of interest since the 1960s (Burrows et al., 2015).

Existing research in ASAG faces notable challenges. One limitation of current experiments is the use of limited evaluation categories, ranging from binary ("correct"/"incorrect") to more nuanced five-level scales ("very good" to "very bad"). To our knowledge, there is no ASAG model that numerically evaluates responses. This is understandable given that current linguistic models perform better with hierarchical labeling than with numerical ratings due to their text-based training. There are also certain cultural implications in this aspect. In addition, these models often overlook nuanced assessment styles unique to individual educators, which can undermine the unique assessment

perspectives they bring to their roles. Another major obstacle is the requirement for training examples for model effectiveness, posing a challenge when teachers wish to assess novel questions, requiring the labor-intensive creation of new training examples, comparable in effort to manual grading.

Among the most advanced ASAG models, the one proposed by Schneider et al. (2023) stands out. This model is based on multilingual transformers (BERT and LaBSE), which have been trained on a substantial dataset comprising approximately 10 million question-answer pairs across two classes. A notable feature of its contribution is its capacity for modulating the system's error tolerance –false positives and false negatives– delegating to the educators the correction of the items that pose the most doubts to the model. In contrast, the model introduced by Ormerod et al. (2023) is characterized by an ensemble of deep neural networks alongside a Latent Semantic Analysis-based model. In this model, holistic 2-point and 3-point rubrics were used, and special emphasis was placed on mitigating the biases inherent in machine learning models. In the domain of reading comprehension questions, Henkel et al. (2023) claim to be the first authors to announce an ASAG model, which matches or exceeds human evaluative performance. This model leverages the ChatGPT Application Programming Interface (API) and employs grading scales of 2 and 3 points.

The datasets currently available for ASAG research are not without their limitations. A primary constraint is the reliance on categorical rather than numerical grading, which is common to the aforementioned ASAG models. Moreover, the public nature of these datasets raises questions about their possible inclusion in GPT model training materials, a detail that the model developer has not publicly disclosed. Therefore, to safeguard the validity of our ChatGPT experiments, we decided to employ a novel, unpublished dataset, despite the resultant limitation in data quantity.

The search for reliable ASAG models is particularly relevant in the context of teaching overload and pursuit of more objective, consistent assessment methodologies. This quest takes on even greater importance in the field of distance education and is particularly crucial in the burgeoning context of Massive Open Online Courses (MOOCs), as highlighted by Y. Wang & Song

(2022). Manual grading, especially in courses with a large number of students, is often a laborious task prone to subjective bias (Campbell, 2015). This study seeks to examine the efficacy of LLMs to perform coherent grading aligned with teacher standards. Specifically, our intention is to test the capability of ChatGPT as an ASAG tool using a numerical rating and using the web interface. The rationale for employing ChatGPT-4's web interface in this investigation stems from its broad accessibility, user-friendliness, absence of additional model training prerequisites, and cost efficiency as an AI tool. While the API of ChatGPT-4 offers capabilities for fine-tuning certain parameters, such as the model's creativity tendency or "temperature"—a feature recommended to be set to 0 in this type of experiments by OpenAI, the corporation responsible for developing this model (Henkel et al., 2023)— this mode of operation requires programming knowledge, thereby limiting its accessibility. Since this kind of technical manipulation is beyond the reach of most teachers, this study opts for the more accessible web interface approach.

Advances in automated assessment systems have important implications for both operational efficiency and equity in the education sector, as they present a viable answer to a long-standing problem in pedagogy: providing rapid, comprehensive, accurate and equitable assessments.

15.2. Objective and Methods

The aim of this study is to evaluate the efficacy of ChatGPT-4 in numerically grading short open-ended questions within a specific field of Health Sciences, adhering to the assessment standards established by a subject matter expert. This involves comparing the grading outcomes of ChatGPT-4 with those determined by an educational expert in the discipline, across various types of input prompts. Such comparative analysis is instrumental in understanding the applicability and preparation of Large Language Models (LLMs) like ChatGPT for specialized grading tasks.

This research employs a mixed-methods comparative analysis to explore the congruence between ChatGPT's grading capabilities and those of an expert educator within the domain of Physi-

cal Podiatry. The participant cohort consisted of 62 Spanish undergraduate students, with all participants attempting the first question (Q1) and 59 addressing the second question (Q2). The teacher's grades for each of the questions were compared with 3 different prompts: one incorporating 10 examples, a second featuring 25 examples, and a third guided by a detailed marking rubric. The prompts had the following format:

> ACT AS AN EXPERT + TASK STEP BY STEP + EXPECTED OUTPUT + EXAMPLES OR EVALUATION CRITERIA + QUESTION TO EVALUATE

Furthermore, to augment the study's robustness, an external educator, not specialized in Physical Podiatry, was also asked to grade the two questions using the same rubric, offering an additional comparative perspective on the grading alignment. The assessments were conducted in January 2024 using ChatGPT-4, with responses graded on a 0 to 10 scale, allowing for decimal values, without any model fine-tuning or parameter modulation.

15.3. Results

To evaluate the congruence between measurements, we employed the Intraclass Correlation Coefficient (ICC) utilizing a two-way random effects mixed model, which assumes absolute agreement and single measurement by the rater. Additionally, we calculated the Quadratic Weighted Kappa (QWK) to facilitate comparison with other studies. It is important to note that the QWK must be applied to categorical data, requiring discretization of the continuous variables in our study to ensure its applicability. This dual approach (Table 15.1) allows for a detailed evaluation of ChatGPT's accuracy in performing ASAG tasks. This not only helps to elucidate the concordance among diverse grading methodologies but also establishes a solid framework for comparison with methodologies previously established in the literature.

Table 15.1. Comparison of the performance of different evaluators vs. the subject teacher

	Evaluator	Intraclass correlation coefficient Lower limit	95% Confidence interval		QWK
			Upper Limit		
Q1	GPT 10X	0.563	0.345	0.719	0.540
	GPT 25X	0.20	0.440	0.753	0.569
	GPT rubric	0.868	0.759	0.925	0.862
	Human	0.941	0.904	0.964	0.931
Q2	GPT 10X	0.697	0.539	0.808	0.698
	GPT 25X	0.621	0.438	0.756	0.616
	GPT rubric	0.828	0.621	0.913	0.829
	Human	0.859	0.679	0.929	0.861

Source: developed by autor.

15.4. Discussion

The findings of the present study provide empirical evidence of ChatGPT's ability to match educators' evaluation criteria in ASAG scenarios. This competence is not only apparent through the presentation of concrete examples but is more evident when the grading rubrics are clearly shown to the model at the prompt. Observations revealed varying levels of concordance between the assessments rendered by the expert educator and those generated by the GPT models. Utilization of a correction rubric in the prompts facilitated the achievement of elevated ICC values, registering 0.868 for Q1 and 0.828 for Q2, suggesting a significant congruence between the expert's evaluations and those proffered by ChatGPT. Although prompts based on examples yielded more modest outcomes, the outcomes remained robust.

Comparatively, the ICC values for the two questions graded by ChatGPT using a rubric (0.868 and 0.828) juxtaposed against the grades of a secondary human evaluator (0.941 and 0.859) demonstrate ChatGPT's proximity to mirroring the evaluative precision of an educator. This is in line with Henkel et al. (2023), who were the first to report a model capable of matching or ex-

ceeding human performance on ASAG tasks in reading comprehension contexts at elementary and middle school levels, also employing ChatGPT. Our outcomes are marginally inferior, which was predictable given the domain of the questions, aimed at a university level and outside the linguistic context for which the large language models have been trained and therefore perform better. Furthermore, it is important to consider that Henkel et al. (2023) designed their study using the ChatGPT API, thereby enabling control over certain variables to enhance model stability. The results obtained (0.89 and 0.92) in grading 2- or 3-class responses are very similar to those of our study using a continuous variable and rubrics (0.862 and 0.829), but superior to our experiments with examples, all below 0.7 QWK.

Recent research, such as that conducted by Ormerod et al. (2023), who implemented mixed models with specific training and rubrics, did not reach such high QWK coefficient values observed in our study, around 0.7. Nevertheless, the analysis revealed that the assessments produced by the model surpassed those executed by human evaluators using the identical dataset. Conversely, Schneider et al. (2023) report a maximal accuracy rate of 86.5% in binary grading (categorized as "correct" or "incorrect") using a model refined through training on millions of question-and-answer pairs, which would also be in line with our results.

A key observation from our study is that it is much more effective to teach the model our evaluative criteria rather than supplying it with examples for autonomous learning. Although this outcome was anticipated, the substantial magnitude of this effect was beyond our initial expectations. Indeed, the prompt designed for the correction of Q1 and Q2 provided with explicit instructions on the correction criteria, exhibited significantly superior performance (0.868 and 0.828), compared to the prompts incorporating either 10 examples (0.563 and 0.697) or 25 examples (0.620 and 0.621). Generally, the time investment required to generate 10 response examples exceeds that required to clearly define the correction criteria or to develop a rubric, and the results, as observed, are significantly better.

Furthermore, it was observed that the prompt with 10 examples for Q2 (ICC of 0.697) outperformed the prompt with 25 examples (ICC of 0.621). This suggests that there is a limit to the

number of examples that ChatGPT can effectively consider, and that exceeding this limit could deteriorate the overall performance of the model. To verify this hypothesis, it would be necessary to conduct a study specifically addressing this issue.

A qualitative review of the data generated by ChatGPT during the grading process revealed that, although the AI correctly reasoned the rationale for each assigned rating, discrepancies sometimes arose between its justifications and the resulting ratings. For instance, we identified situations where the model argued that a given response was superior to a certain example graded with a 3, yet lacked the comprehensive detail of other examples graded at 7. However, following this accurate argumentation, it awarded a grade of 3.5, closer to 3 than to 7, without observing that other examples rated with a 5 were more similar to the evaluated response. We also observed that, in the process of evaluating responses via the rubric, ChatGPT demonstrated computational inaccuracies on several occasions. Specifically, when segmenting the student responses to assign partial scores, we found errors in the addition or division operations required to obtain the final grade. In certain scenarios, the model generated and executed a small internal program for mathematical calculations, achieving accurate results thereafter. Despite these computational discrepancies, we adhered to a policy of non-intervention, upholding the model's final assigned grade, even in the presence of arithmetic errors.

Regarding the documented computational issues of ChatGPT with mathematics (Borji, 2023; Shakarian et al., 2023), the decision to implement a continuous scale from 0 to 10 for grading may have negatively impacted the model's performance. An assessment of ChatGPT's efficacy in grading complex university-level responses on a categorical rather than a numerical scale could facilitate a more congruent comparison with extant literature. Nonetheless, the aim of our study was to evaluate the model's capability to accurately process numerical data. We base this on the assumption that correct numerical data handling would likely enhance its performance in categorization tasks.

An additional limitation relates to the linguistic context of the assessment materials; the questions, answers, and prompts were presented in Spanish. Although the model can interpret and generate text in this language, the majority of its training corpus is in

English, which suggests that the results could improve if this language were used. To date, in our literature review we have not found any studies that specifically address the comparative performance of ChatGPT across languages. Given the global application of large linguistic models and the inherent linguistic diversity of users, understanding how ChatGPT's effectiveness varies by language is vitally important. This gap in the existing body of research presents a great opportunity for future research. Such studies would not only enrich our understanding of the linguistic capabilities of the model, but would also provide strategies for its optimization and application in multilingual contexts. Accordingly, we advocate the initiation of research aimed at evaluating ChatGPT's performance across a broad spectrum of languages, which would provide information of great value to the academic and technology communities.

In the course of our investigation, we identified specific instances where student responses resulted in an overestimation of grades by ChatGPT. For example, responses featuring extensive lists of technical terms—regardless of their accuracy—tended to be awarded higher grades compared to concise, error-free submissions. This phenomenon is consistent with findings from prior research, which has documented the susceptibility of LLMs to adversarial inputs that exploit model vulnerabilities (Filighera et al., 2020; J. Wang et al., 2023). Despite concerted efforts within the field, a robust solution to mitigate these types of adversarial attacks remains elusive.

Contrary to findings reported in other studies, our analysis did not reveal any biases in the text generated by the LLM (Acerbi & Stubbersfield, 2023), which may be attributable to the specific nature of the task assigned to ChatGPT and the evaluation context.

The systematic observation of ChatGPT to align with the grading standards of educators, even with a limited number of examples or a simple rubric, across both evaluated questions (Q1 and Q2), not only substantiates the methodological approach employed but also highlights the potential of LLMs as versatile and effective tools for educational assessment. This is especially pertinent in educational contexts, where the demand for efficiency is ever-increasing, and educators frequently face substantial workload challenges.

The significance of these findings extends beyond merely facilitating a reduction in educators' workload through the deployment of accessible and economically viable technological solutions. It also encompasses the enhancement of grading uniformity. Although the initial development of prompts with rubrics or multiple examples may incur substantial effort, this investment is marginal compared to the labor-intensive process of evaluating numerous student responses. Importantly, this approach fosters educational equity by mitigating the variability introduced by human assessors' fatigue, which can lead to inconsistent grading over time (Klein & El, 2003).

Importantly, ASAG models need not entirely replace educators in assessing student performance. Instead, AI can complement and support instructional efforts by offering alternative assessments, identifying grading inconsistencies, or preliminarily sorting responses to expedite the evaluation process. The results obtained suggest that such implementations could be applied in a wide range of educational contexts, providing scalable support to educators. This, in turn, could free up valuable time to focus on other aspects of teaching and allow for quicker and more personalized feedback for students.

The initial outcomes are indeed encouraging, but it is necessary to solidify them by conducting further comprehensive research across various academic disciplines and among different educator demographics. Future research should also focus on the mechanisms through which AI models interpret and apply grading criteria, examining these processes in light of existing evaluation theories and practices.

15.5. Conclusions

This research corroborates the hypothesis that LLMs, and particularly GPT series, represent the most promising approach in the ASAG field. These large language models are highly versatile and are capable of undertaking classification and grading tasks without needing specific prior training.

A significant finding of our research is that, to align with the teacher's grading style, a prompt with a rubric or a good description of the objectives sought by the teacher proves more effective

than providing the model with many examples. This approach not only simplifies and speeds up the process but also improves the outcomes. ChatGPT's ability to adapt to different evaluation styles underscores its potential as a transformative tool in educational assessment.

However, ChatGPT used via its web interface and without specific controls, can lean towards overly creative responses, yielding arbitrary grades, thereby constraining its utility as a universally applicable, unsupervised ASAG tool. It is also highly susceptible to mathematical calculation errors and adversarial attacks. Despite these challenges, its competence in grading complex health science answers at a human-equivalent level is remarkable. Future research should focus on how to effectively control this model to ensure uniform assessments.

In conclusion, the findings of this study, along with those of similar recent research, suggest that the way forward is the use of large language models with fine-tuning to achieve more accurate and stable grades.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, *120*(44), e2313790120. https://doi.org/10.1073/pnas.2313790120

Borji, A. (2023). *A Categorical Archive of ChatGPT Failures* (arXiv:2302.03494). arXiv. https://doi.org/10.48550/arXiv.2302.03494

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, *68*(1), 47-57. https://doi.org/10.1007/s11528-023-00896-0

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy*, *44*(3), 517–547. https://doi.org/10.1017/S0047279415000227

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in education: A review. *IEEE Access*, *8*, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

Filighera, A., Steuer, T., & Rensing, C. (2020). Fooling automatic short answer grading systems. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.). *Artificial Intelligence in Education* (pp. 177-190). Springer International. https://doi.org/10.1007/978-3-030-52237-7_15

Henkel, O., Hills, L., Roberts, B., & McGrane, J. (2023). *Can LLMs Grade Short-answer Reading Comprehension Questions: Foundational Literacy Assessment in LMICs* (arXiv:2310.18373). arXiv. https://doi.org/10.48550/arXiv.2310.18373

Henrickson, L., & Meroño-Peñuela, A. (2023). Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01752-8

Ifenthaler, D., & Schumacher, C. (2023). Reciprocal issues of artificial and human intelligence in education. *Journal of Research on Technology in Education*, *55*(1), 1-6. https://doi.org/10.1080/15391523.2022.2154511

Klein, J., & El, L. P. (2003). Impairment of teacher efficiency during extended sessions of test correction. *European Journal of Teacher Education*,*26*(3),379-392.https://doi.org/10.1080/0261976032000128201

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12249-8

Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., & Eger, S. (2023). *ChatGPT: A Meta-Analysis after 2.5 Months* (arXiv:2302.13795). arXiv. https://doi.org/10.48550/arXiv.2302.13795

Ormerod, C., Lottridge, S., Harris, A. E., Patel, M., van Wamelen, P., Kodeswaran, B., Woolf, S., & Young, M. (2023). Automated short answer scoring using an ensemble of neural networks and latent semantic analysis classifiers. *International Journal of Artificial Intelligence in Education*, *33*(3), 467-496. https://doi.org/10.1007/s40593-022-00294-2

Pacchiega, C. (2021). How can education use artificial Intelligence? A brief history of ai, its usages, its successes, and its problems when applied to education. In G. Panconesi, & M. Guida (Eds.). *Handbook*

of Research on Teaching with Virtual Environments and AI (pp. 558-590). IGI Global. https://doi.org/10.4018/978-1-7998-7638-0.ch024

Preiksaitis, C., & Rose, C. (2023). Opportunities, challenges, and future directions of generative Artificial Intelligence in medical education: Scoping review. *JMIR Medical Education*, *9*(1), e48785. https://doi.org/10.2196/48785

Schneider, J., Richner, R., & Riser, M. (2023). Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, *33*(1), 88-118. https://doi.org/10.1007/s40593-022-00289-z

Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). An independent evaluation of ChatGPT on Mathematical Word Problems (MWP): *AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering*, AAAI-MAKE 2023. CEUR Workshop Proceedings, 3433. http://www.scopus.com/inward/record.url?scp=85166472786&partnerID=8YFLogxK

Taecharungroj, V. (2023). "What can chatgpt do?" Analyzing early reactions to the innovative AI chatbot on Twitter. *Big Data and Cognitive Computing*, *7*(1), Article 1. https://doi.org/10.3390/bdcc7010035

Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., Jiao, B., Zhang, Y., & Xie, X. (2023). *On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective* (arXiv:2302.12095). arXiv. https://doi.org/10.48550/arXiv.2302.12095

Wang, Y., & Song, J. (2022). The success of Massive Open Online Courses (MOOCs): An investigation on course relevance. *Communications of the Association for Information Systems*, *51*(1). https://doi.org/10.17705/1CAIS.05131

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on Artificial Intelligence applications in higher education. Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0