The Problem of AI Hallucination and How to Solve It

Antonín Jančařík[1] and Ondřej Dušek[2]
[1]Charles University, Faculty of Education, Prague, Czech Republic
[2]Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic

antonín.jancarik@pedf.cuni.cz
odusek@ufal.mff.cuni.cz

Abstract: AI in education is a topic that has been researched for the last 70 years. However, the last two years have seen very significant changes. These changes relate to the introduction of OpenAI's ChatGPT chatbot in November 2022. The GPT (Generative Pre-trained Transformer) language model has dramatically influenced how the public approaches artificial intelligence. For many, generative language models have become synonymous with AI and have come uncritically viewed as a universal source of answers to most questions. However, it soon became apparent that even generative language models had their limits. Among the main problems that emerged was hallucination (providing answers containing false or misleading information), which is expected in all language models. The main problem of hallucination is that this information is difficult to distinguish from other information, and AI language models are very persuasive in presenting it. The risks of this phenomenon are much more substantial when using language modules to support learning, where the learner cannot distinguish correct information from incorrect information. The proposed paper focuses on the area of AI hallucination in mathematics education. It will first show how AI chatbots hallucinate in mathematics and then present one possible solution to counter this hallucination. The presented solution was created for the AI chatbot Edu-AI and designed to tutor students in mathematics. Usually, the problem is approached so that the system verifies the correctness of the output offered by the chatbot. Within the Edu-AI, checking responses is not implemented, but checking inputs is. If an input containing a factual query is recorded, it is redirected, and the answer is traced to authorised knowledge sources and study materials. If a relevant answer cannot be traced in these sources, a redirect to a natural person who will address the question is offered. In addition to describing the technical solution, the article includes concrete examples of how the system works. This solution has been developed for the educational domain but applies to all domains where users must be provided with relevant information.

Keywords: Chatbots, AI, Mathematics education, Hallucination

1. Introduction

Artificial intelligence and the possibilities of its use in teaching have been the subject of didactic research for decades (Chen et al., 2020; Zhai et al., 2021). Zawacki-Richter has explored the possibilities of studies focusing on the use of AI in higher education and lists four areas of research focus in 2019 (Zawacki-Richter et al., 2019):

- Profiling and prediction
- Assessment and evaluation
- Intelligent tutoring systems
- Adaptive systems and personalisation

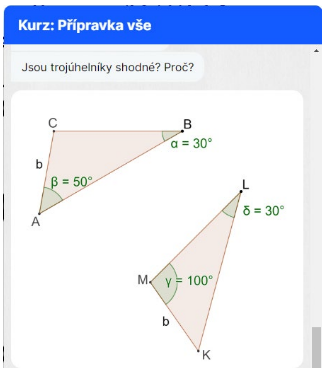
The situation was similar in mathematics teaching (Hwang & Tu, 2021). The focus of research has been on how AI can impact and enhance the performance of mathematics students in their teaching and learning process, with most research focusing on the effectiveness of AI use (Mohamed et al., 2022).

The professional and lay public's view of the possibilities of using AI in education was significantly changed by the publication of ChatGPT 3.0 in the fall of 2022. AI-driven large language models (LLMs) such as OpenAI's ChatGPT have become almost synonymous with AI. The new features offered by LLMs have taken the forefront of educational research (Zawacki-Richter et al., 2024; Crompton et al., 2024). Like all new technologies, LLMs bring benefits AND risks (Groza and Marginean, 2023). One of these risks is AI hallucination, i.e. providing incorrect or misleading results that AI models generate. Particularly around using LLMs to communicate directly with learners about educational content, AI hallucination can be a significant problem because LLMs provide incorrect answers with high fidelity, and learners may be unable to distinguish them from correct answers.

This paper presents one way to solve, or at least reduce, the AI hallucination problem. The solution was developed and used within the Edu-AI.eu project to create a chatbot for tutoring students in mathematics and the Czech language.

2. Chatbot Edu-AI

Chatbot Edu-AI is an application created to tutor primary and secondary school students. The application involves students interacting with an AI chatbot, which presents them with pre-prepared mathematical problems. It is important to note that the AI's role is not to generate or prepare mathematical content but rather to facilitate communication with the student beyond the assigned tasks (Jančařík et al., 2023). The application currently covers all the topics that appear in the entrance exams to secondary schools in the Czech Republic. The application presents students with tasks based on their choice and previous answers (see Jančařík et al., 2022). A sample of such a task is shown in Figure 1.

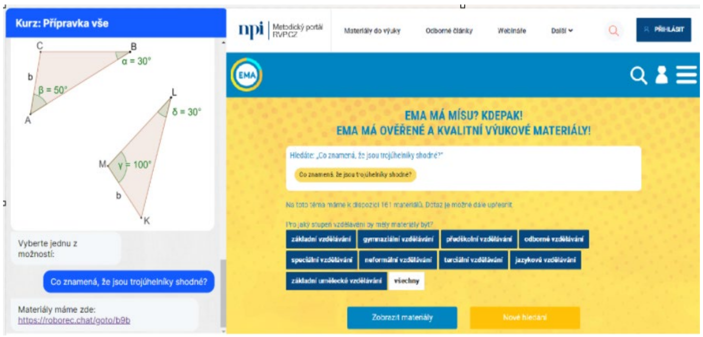


Kurz: Přípravka vše

Jsou trojúhelníky shodné? Proč?

Figure 1: Sample problem: Are the triangles congruent? Why?

Since this is a chat application, there may be a situation where a student does not know what it means that the triangles are congruent and asks this question in the chat. In this case, the app's creators wanted to make sure that the answer the student receives is correct. Therefore, the learner does not receive a direct answer but is offered a link to open a menu of learning materials that address the topic (see Figure 2).



Kurz: Přípravka vše

Vyberte jednu z možností:

Co znamená, že jsou trojúhelníky shodné?

Materiály máme zde: https://roborec.chat/goto/b9b

EMA MÁ MÍSU? KDEPAK!
EMA MÁ OVĚŘENÉ A KVALITNÍ VÝUKOVÉ MATERIÁLY!

Hledáte: „Co znamená, že jsou trojúhelníky shodné?"

Co znamená, že jsou trojúhelníky shodné?

Na toto téma máme k dispozici 161 materiálů. Dotaz je možné dále upřesnit.

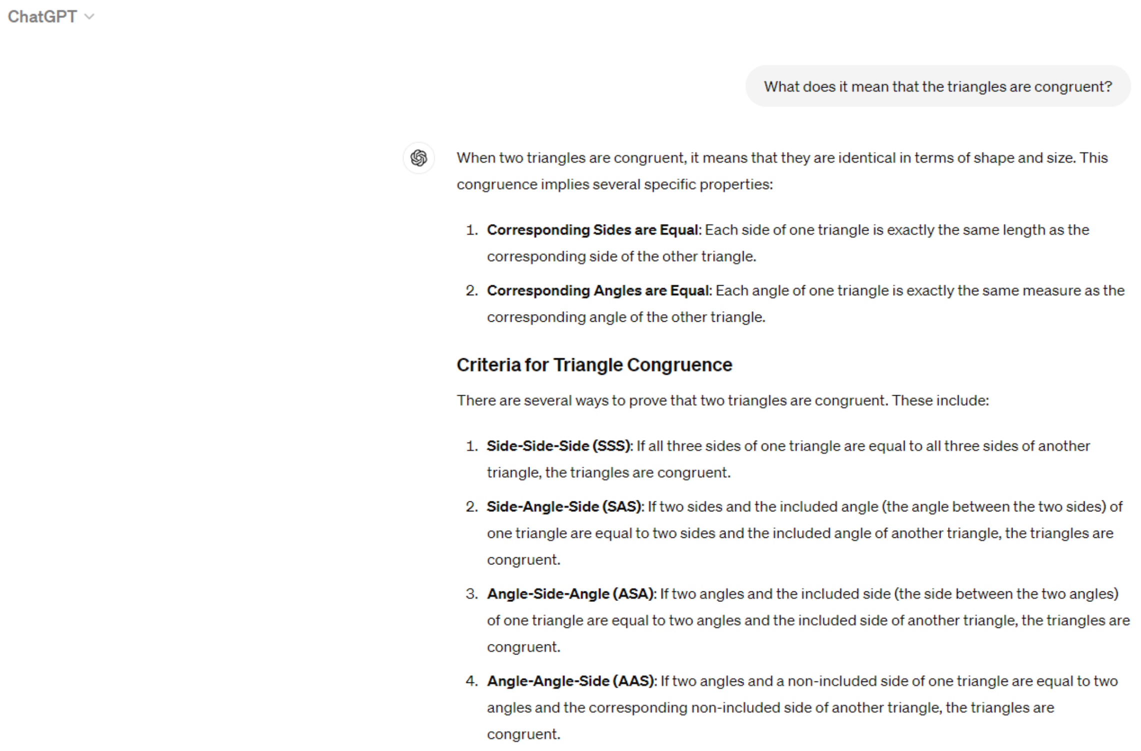
Pro jaký stupeň vzdělávání by měly materiály být?

Figure 2: Answer to the question: What does it mean that the triangles are congruent?

In this case, 161 teaching materials were found, so the student must specify which materials he/she is interested in; for example, he/she can choose teaching videos for primary school. All the resources referenced by the application in the case of a factual query have been peer-reviewed and can be considered validated.

3. AI Hallucination in Mathematics

This procedure is chosen because LLMs can sometimes hallucinate and provide misleading answers. This problem occurs in various areas but is often very pronounced in mathematics (Poola, 2023). We will demonstrate the problem with the previous example. We get the correct answer if we ask the same question to the ChatGPT application (see Figure 3).



What does it mean that the triangles are congruent?

When two triangles are congruent, it means that they are identical in terms of shape and size. This congruence implies several specific properties:

1. **Corresponding Sides are Equal**: Each side of one triangle is exactly the same length as the corresponding side of the other triangle.
2. **Corresponding Angles are Equal**: Each angle of one triangle is exactly the same measure as the corresponding angle of the other triangle.

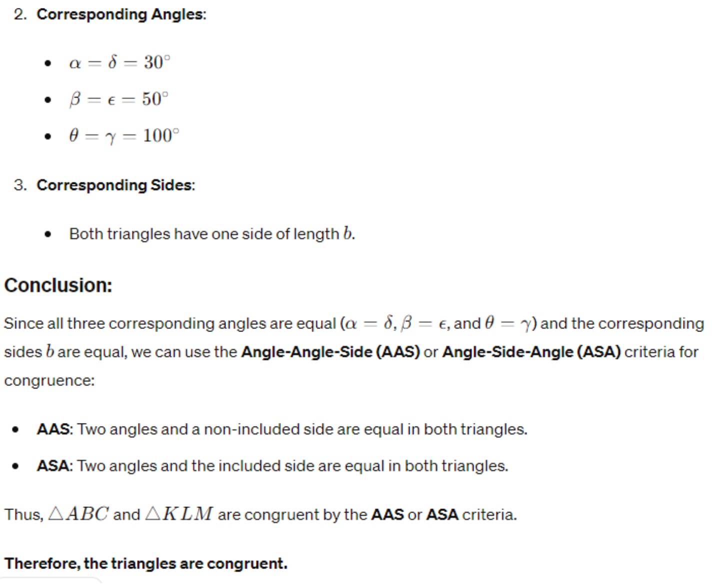
Criteria for Triangle Congruence

There are several ways to prove that two triangles are congruent. These include:

1. **Side-Side-Side (SSS)**: If all three sides of one triangle are equal to all three sides of another triangle, the triangles are congruent.
2. **Side-Angle-Side (SAS)**: If two sides and the included angle (the angle between the two sides) of one triangle are equal to two sides and the included angle of another triangle, the triangles are congruent.
3. **Angle-Side-Angle (ASA)**: If two angles and the included side (the side between the two angles) of one triangle are equal to two angles and the included side of another triangle, the triangles are congruent.
4. **Angle-Angle-Side (AAS)**: If two angles and a non-included side of one triangle are equal to two angles and the corresponding non-included side of another triangle, the triangles are congruent.

Figure 3: ChatGPT answer to the question: What does it mean that the triangles are congruent?

So, there does not seem to be a problem in this case. Copying the problem specification into the query is possible, and ChatGPT will solve the job correctly (see Figure 4).



2. **Corresponding Angles**:
   - $\alpha = \delta = 30^\circ$
   - $\beta = \epsilon = 50^\circ$
   - $\theta = \gamma = 100^\circ$
3. **Corresponding Sides**:
   - Both triangles have one side of length b.

Conclusion:

Since all three corresponding angles are equal ($\alpha = \delta, \beta = \epsilon$, and $\theta = \gamma$) and the corresponding sides b are equal, we can use the **Angle-Angle-Side (AAS)** or **Angle-Side-Angle (ASA)** criteria for congruence:

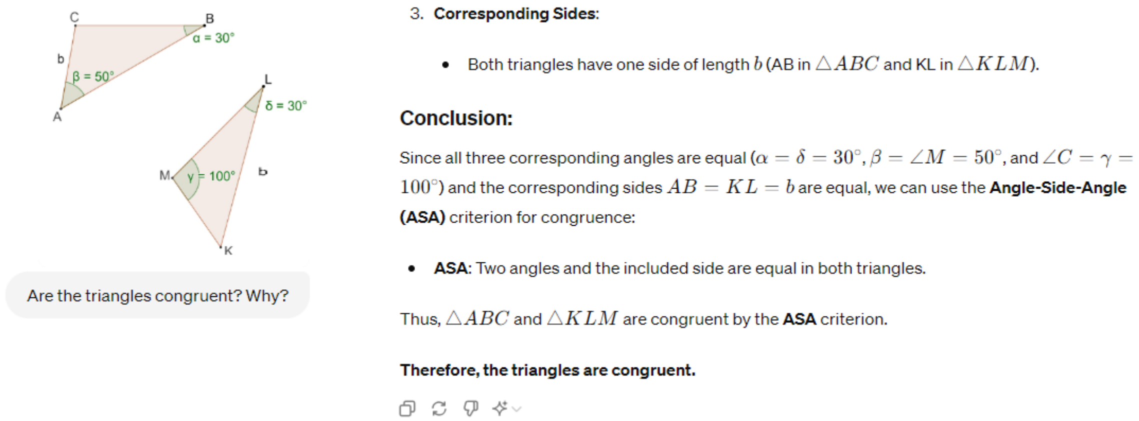
- **AAS**: Two angles and a non-included side are equal in both triangles.
- **ASA**: Two angles and the included side are equal in both triangles.

Thus, $\triangle ABC$ and $\triangle KLM$ are congruent by the **AAS** or **ASA** criteria.

Therefore, the triangles are congruent.

Figure 4: ChatGPT solution of the problem

However, the impression that ChatGPT is already capable of a similar task is illusory, and the hallucination problem is revealed in full if we present the same task to the application, with only a slight change in the assignment by moving the label of side b (see Figure 5). In this case, the answer is a detailed analysis of the task with the same conclusion that both triangles are identical. However, in this case, this answer is incorrect.



Are the triangles congruent? Why?

3. **Corresponding Sides**:

- Both triangles have one side of length b (AB in $\triangle ABC$ and KL in $\triangle KLM$).

Conclusion:

Since all three corresponding angles are equal ($\alpha = \delta = 30^\circ, \beta = \angle M = 50^\circ$, and $\angle C = \gamma = 100^\circ$) and the corresponding sides $AB = KL = b$ are equal, we can use the **Angle-Side-Angle (ASA)** criterion for congruence:

- **ASA**: Two angles and the included side are equal in both triangles.

Thus, $\triangle ABC$ and $\triangle KLM$ are congruent by the **ASA** criterion.

Therefore, the triangles are congruent.

Figure 5: ChatGPT incorrect solution of the problem

The indistinguishability of correct and incorrect answers led to the need to develop a solution that would avoid confusing students. We present a description of the implemented solution below.

4. Implemented Solution of AI Hallucination in Chatbot Edu-AI

The solution implemented within the Edu-AI application is that all inputs given by the learner are analysed by the LLMs (currently Chat GPT 4) before being processed by the exerciser. The chatbot consists of a rule-based in-domain module, where individual learning exercises are implemented, and an open-domain module, which handles queries not focused on carrying out the learning exercises. The open-domain module (Fig. 1) operates in the following fashion:

1. First, all queries undergo basic preprocessing (accented character correction; Richter et al., 2012)
2. The queries are passed through an intent classifier (Section 4.1). We use five general intents: simple queries (such as “What is the time?”) for handcrafted handling, returning to the in-domain module, searching for learning materials, factual questions, and chitchat.
3. A rule-based dialogue manager decides on further treatment of the input, simply depending on the intents. This includes handling the handcrafted intents and returning control to the in-domain module. Learning material search is redirected to the Czech school ministry’s EMA learning materials database.
4. Question intents are passed to two question-answering (QA) modules, one for domain-related questions and another for open-domain, based on Wikipedia (Sections 4.2 and 4.3).
5. The chitchat intent is forwarded to a chitchat module (Section 4.4).

The whole open-domain module runs as a simple HTTP server. A simple keywords-based filter is included to avoid responses involving age-inappropriate or potentially offensive topics.

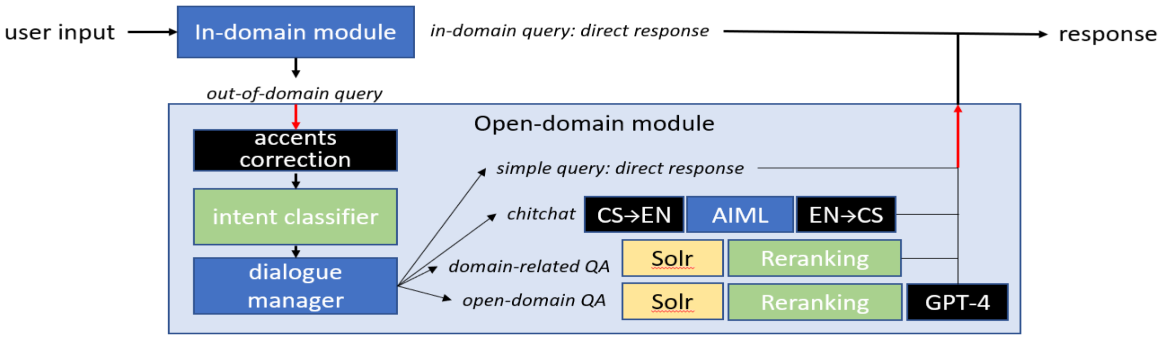


Figure 6: The overall architecture of the system.

4.1 Intent Classifier

Our intent classifier is based on the RobeCzech neural language model (Straka et al., 2021), which has been pretrained for masked language modelling (i.e., guessing masked out words from their context; Devlin et al., 2019) on vast amounts of Czech texts downloaded from the internet. We use this model as a starting point to provide us with vector representation of Czech sentences (user inputs) and finetune (train) it using a custom dataset of ca. 5,800 utterances manually annotated with our intent labels. We can achieve >80% intent classification accuracy on a held-out test portion of our data.

4.2 Domain-Related Question Answering

For the domain-related QA, we currently support two databases, which are switched based on the source of the HTTP request on the system, i.e., depending on the website the user is looking at. This means the same underlying chatbot can show different behaviour on different websites.

The QA starts with a full-text search in a database of frequently asked questions and corresponding answers related to the domain. We use Solr (https://solr.apache.org/) for data storage and full-text search. Before the search, the user query is filtered based on part-of-speech detection (Straková et al., 2014); only nouns, adjectives, verbs and adverbs are retained. The query is further lemmatised (converted to base word forms, removing plurals, verb tense, case endings, etc.), and standard abbreviations (chosen from a manually compiled list) are expanded to their complete forms.

The Solr full-text query returns the top 10 results, which are further re-ranked to choose the one that best matches the user query. Here, we use neural vector representations of sentences (embeddings, Bengio et al., 2003) and the cosine similarity of the user query to the retrieved questions as a matching criterion. We use the Seznam RetroMAE-small model (Bednář et al., 2023), which we selected based on evaluating multiple sentence embeddings on small test queries.

4.3 Open-Domain Question Answering

The open-domain one is more complex than the domain-related QA system. It uses the exact Solr full-text search and part-of-speech query filtering. However, this time, the user query is not compared to questions from a database but to Wikipedia article titles and excerpts (paragraphs). Article title matches are prioritised. In this case, the reranking uses a multilingual SBERT sentence embedding model (Reimers and Gurevych, 2019), as we found it more accurate in comparing the similarity (relatedness) of questions to text paragraphs.

Unlike the domain-related QA, where the database already contains ready-made answers, the Solr search and reranking result is a Wikipedia paragraph, which, in an ideal case, contains the reply. The reply must first be extracted and passed to the user. We use GPT4-turbo (OpenAI, 2023) for this task. We prompt the model to respond to the user question given the extracted paragraph and give it the option to indicate that the extracted paragraph is, in fact, not relevant. If the extracted paragraph is irrelevant, we pass GPT4 the question without additional data and ask for an answer. In this case, the chatbot output indicates lower confidence ("I am not sure, but I think…").

4.4 Chitchat

For chitchat, we use a pipeline based on machine translation into English and back and the AIML pattern-matching chatbot engine (Wallace, 2009). We opted to use a rule-based chatbot over a neural trainable system to retain complete control of the responses. We previously experimented with the BlenderBot neural chatbot (Roller et al., 2021) and found that it would often steer the conversation towards inappropriate topics. Having decided on AIML, we customised the patterns of the existing ALICE chatbot (Wallace, 2009) for our purposes. We opted for pattern matching in English as this avoids the morphological complexities of Czech, and we used machine translation (Popel et al., 2020) to achieve this. The translation of the AIML output back into Czech is post-processed by syntax-based rules built on top of the UDPipe syntactic parser (Straka, 2018) to ensure the chatbot shows a consistent gender in its responses.

5. Conclusion

We know that LLMs are undergoing a rapid evolution, including minimising AI hallucinations. The progress in this area is evident. Problems that we used to demonstrate problems with math tasks last year or the year before are now very often solved correctly. It is equally likely that the problem used as a demonstration in this paper will no longer be current by the time of the conference presentation. However, the hallucination problem is much deeper and is not likely to be solved anytime soon. In addition, there will always be a need to provide a link to specific information or relevant, pre-prepared resources instead of a general answer within the chat.

While the solution we present in this paper was developed for the needs of one Czech tutoring application, it is generally portable and applicable in a broader context. The procedure can be used whenever there are guaranteed sources of information. In our case, these were proven sources of educational material provided by one of the project partners. We are also aware that operator redirection can reduce system efficiency. We can consider replacing it with an automatic response that the chatbot cannot find a verified answer to the query.

Further research will focus on the solution's reliability and user feedback, including mapping how much users use the given links and how they perceive them compared to the text-based response.

Acknowledgements

This paper was supported by the project TAČR N. TL05000236 – AI assistant for pupils and teachers.

References

Bednář, J., Náplava, J., Barančíková, P. and Lisický, O. (2023) "Some Like It Small: Czech Semantic Embedding Models for Industry Applications", https://doi.org/10.48550/arXiv.2311.13921.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003) "A Neural Probabilistic Language Model", *Journal of Machine Learning Research*, Vol 3, pp1137–1155.

Chen, L., Chen, P. and Lin, Z. (2020) "Artificial intelligence in education: A review", *Ieee Access*, Vol 8, pp 75264–75278.

Crompton, H., Jones, M. V. and Burke, D. (2024) "Affordances and challenges of artificial intelligence in K-12 education: A systematic review", *Journal of Research on Technology in Education*, Vol 56, No. 3, pp 248–268.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" In: J. Burstein, C. Doran and T. Solorio, eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics. pp 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Groza, A. and Marginean, A. (2023) *Brave new world: Artificial Intelligence in teaching and learning*. arXiv preprint arXiv:2310.06856.

Hwang, G. J. and Tu, Y. F. (2021) "Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review". *Mathematics*, Vol 9, No. 6, Art. 584, https://www.mdpi.com/2227-7390/9/6/584.

Jančařík, A., Novotná, J. and Michal, J. (2022), "Artificial Intelligence Assistant for Mathematics Education". In *Proceedings of the 21st European Conference on e-Learning-ECEL,* pp 143–148.

Jančařík, A., Michal, J. and Novotná, J. (2023) "Using AI Chatbot for Math Tutoring", Journal of Education Culture and Society, Vol 14, No. 2, pp 285–296, https://doi.org/10.15503/jecs2023.2.285.296.

Metze, K., Morandin-Reis, R. C., Lorand-Metze, I. and Florindo, J. B. (2024) "Bibliographic research with ChatGPT may be misleading: the problem of hallucination" *Journal of Pediatric Surgery*, Vol *59, No.* 1, 158.

Mohamed, M. Z. B., Hidayat, R. and Mahmud, M. K. H. B. (2022) "Artificial Intelligence in Mathematics Education: A Systematic Literature Review", *International Electronic Journal of Mathematics Education*, Vol 17, No. 3, em0694, https://doi.org/10.29333/iejme/12132.

Poola, I. (2023) "Tuning ChatGPT mathematical reasoning limitations and failures with process supervision", *Novelty Journals*, Vol 10, No. 2, pp 55–66, https://doi.org/10.5281/zenodo.8296440.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O. and Žabokrtský, Z. (2020). "Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals", *Nature Communications*, Vol 11, No. 1, Art. 4381. https://doi.org/10.1038/s41467-020-18073-9.

Reimers, N. and Gurevych, I. (2019) "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", In: K. Inui, J. Jiang, V. Ng and X. Wan, eds. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* pp 3982–3992. https://doi.org/10.18653/v1/D19-1410.

Richter, M., Straňák, P. and Rosen, A. (2012) "Korektor – A System for Contextual Spell-Checking and Diacritics Completion", In: M. Kay and C. Boitet, eds. *Proceedings of COLING 2012: Posters*. [online] COLING 2012. Mumbai, India: The COLING 2012 Organizing Committee. Pp 1019–1028. https://aclanthology.org/C12-2099.

Straka, M. (2018) "UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task", In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. [online] CoNLL 2018. Brussels, Belgium: Association for Computational Linguistics. pp 197–207. https://doi.org/10.18653/v1/K18-2020.

Straková, J., Straka, M. and Hajič, J. (2014) "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition", In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. [online] Baltimore, Maryland: Association for Computational Linguistics. pp 13–18. https://doi.org/10.3115/v1/P14-5003.

Wallace, R.S. (2009) "The Anatomy of A.L.I.C.E", In: R. Epstein, G. Roberts and G. Beber, eds. *Parsing the Turing Test*. Dordrecht: Springer Netherlands. pp181–210. https://doi.org/10.1007/978-1-4020-6710-5_13.

Zawacki-Richter, O., Marín, V. I., Bond, M. and 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*, Vol 16, No. 1, pp 1–27. https://doi.org/10.1186/s41239-019-0171-0

Zawacki-Richter, O., Bai, J. Y., Lee, K., Slagter van Tryon, P. J. and Prinsloo, P. (2024) "New advances in artificial intelligence applications in higher education?" *International Journal of Educational Technology in Higher Education*, Vol 21, No. 1, Art. 32, https://doi.org/10.1186/s41239-024-00464-3.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... and Li, Y. (2021) "A Review of Artificial Intelligence (AI) in Education from 2010 to 2020", *Complexity*, Art. 8812542, pp 1–18, https://doi.org/10.1155/2021/8812542.