3. Меретукова, М.М. О малых жанровых формах английской рождественской литературы / М.М. Меретукова // Вестник Адыгейского государственного университета. – 2018. – № 3(222). – С. 156–162.
4. Растрепенин, А.В. Страх и ужас в английских готических романах XVIII–XIX вв / А.В. Растрепенин // Littera terra : Материалы X международной конференции молодых ученых. – 2022. – № 16. – С. 14–23.
5. Dickens, Ch. The Haunted House / Ch. Dickens – Global Grey, 2018. – 136 p.
6. Dickens, Ch. The Signal–Man / Ch. Dickens – Short Story America, 1866. – 11 p.
7. Dickens, Ch. The Trial For Murder / Ch. Dickens – Holland : Dreamscape Media, 2018. – 34 p.
8. Grose, F. A Provincial Glossary; with a collection of local proverbs, and popular superstitions / F. Grose. – London : E. Jeffery, 1811. – 124 p.

Науч. рук.: Баева Н.А., к.филол.н, доц.

А.В. Зайда, Я.А. Согуляк

Национальный исследовательский
Томский политехнический университет

Educational AI technologies: from useful to problematic

This paper identifies and explores newly popular AI technologies that have found use in education and academia. AI tools were categorized by their potential usefulness for education. Specifically, AI detection tools were analyzed more closely. Conducted experiment suggested that the accuracy of such tools varies a lot and is often insufficient to be used in judgment of text authenticity.

Keywords: education, artificial intelligence, machine learning, educational technologies, language learning.

Recently, there was an uproar in academia and education following the wide availability of access to Large Language Models (LLM) in the form of chat bots. The emerging cluster of technologies called Artificial Intelligence (AI) quickly gained notoriety for all the uses it has, that are linked to academic dishonesty. On the bright side, this situation also pushed the introduction of numerous applications under the very broad definition of AI into various fields. Nowadays, there is a considerable variety of AI solutions marketed particularly towards modern academia and education [2]. However, the quality and utility of said technologies differs substantially, from genuinely useful to troubled execution and nonsensical in the context of the operating principles of AI.

This paper aims to explore available AI tools in academic and educational context in order to identify those more worthwhile and practical at the current stage of development. Simultaneously, prospective and unreliable options will also be identified. Consequently, the following tasks were put forward:

– Clear up what constitutes AI;
– Explore fundamental of AI technologies;
– Discover useful educational AI tools;
– Identify promising yet troubled applications of AI;
– Outline problematic solutions that contradict the theory of AI.

The term AI has been haphazardly thrown around majority of new software and services, becoming a true buzzword, foregoing any initial meaning it had. In the actual scientific field of Artificial Intelligence, the term is no more specific. Overall, AI encapsulates any machine, algorithm or program that achieves human-level performance on a task, which is difficult to formalize [4]. Under such definition, a calculator does not constitute AI, since the task of mathematical computation follows very rigid and formal rules. On the other hand, image recognition software would fall under AI category, since it is impossible to describe visual features of any object in an exhaustive way, that is invariant to the perspective.

More importantly, the current advancements in relation to AI can really be attributed to the specific subfield of machine learning (ML) and even its subfields, like the deep learning (DL). Similarly, machine learning has a disputed definition, yet it is generally accepted that ML specifies formal algorithms that allow specialized machines or software attain a human level performance measured according to a concrete metric, at some tasks after observing examples the task being performed. As such, an insight into the well-defined algorithms and principles of ML is instrumental to judge which AI applications make sense and which are yet not possible. Even though the reasons why most powerful ML and DL methods, namely artificial neural networks, work can hardly be fully understood and interpreted, the limitations of the technology are still apparent, considering their design.

The specifics of ML algorithms vary considerably from one subject area to another, for example, image recognition and text generation employ quite different techniques that have substantially dissimilar motivations and justifications. In the end, all ML methods are stochastic and statistical by nature: they generalize knowledge from observing limited datasets to model data distributions. For a simple example, a large language model (LLM) observes text data to model how often specific «words» appear in the context of other «words». Consequently, the resulting model describes the most generic type of text across different styles, genres and even cultures. For most applications, that model is further shifted with state-of-the-art methods to align more closely with desired text type. Still, the resulting «word» distribution remains statistical, not logical, therefore, inherently prone to errors. Anecdotally, AI-powered calculator is right 99.99 % of the time, while a regular one is always correct.

Then, more successful AI educational technologies were considered. The first of such applications is automatic transcription. Powered by AI, it offers a convenient way to convert audio or video recordings into written text. Transcripts of lectures, discussions, and presentations provide access to information for students with hearing impairments or learning disabilities. This can level the playing field and create a more inclusive learning environment. In addition, reading a text is usually faster than listening to it. Transcripts can also be used for reviewing material, identifying key points, and searching for specific information. This can lead to a deeper understanding of the subject matter. Moreover, this technology is well-studied and sequence-to-sequence model helps processing audio signal stably and accurately. However, current AI transcription tools are not perfect. While improving, they often require human intervention for potential error correction. Summarizing: automatic transcription using AI offers significant advantages in terms of speed, cost, and scalability, but it is important to acknowledge its limitations. The best approach is often a hybrid system that combines AI technology providing speed of operation with human review to ensure the highest level of accuracy and quality.

Secondly, there is such a thing as computer vision, a field of artificial intelligence that empowers computers to interpret images and videos in a manner similar to humans. This capability extends to various tasks, ranging from basic object detection to complex scene understanding and action recognition. It can be particularly useful in subjects such as biology, chemistry and physics. For example, computer vision systems can identify different plant and animal species in images, helping students learn about their characteristics and classification. Modern computer vision algorithms exhibit high accuracy and integrating such algorithms into educational processes opens up new avenues for personalized learning, optimizing educational resources, and improving efficiency. However, computer vision also faces limitations. Training these algorithms requires enormous amounts of data, making data acquisition and labelling time-consuming and resource-intensive. Despite these challenges, computer vision continues to evolve, holding immense potential for various applications, including education.

Moving on to more nuanced tools, AI exam proctoring was explored. AI proctoring is a promising application that may help alleviate some of the risks associated with online education and the use of other informational technologies in the classrooms. The idea is to use computer vision and audio processing to supervise exams that are conducted digitally and remotely. Any suspicious movements seen on web camera and unrelated human speech caught by the microphone may be detected automatically, alerting the examiners to intervene. While the proposal is technically sound, since both subject area fields of audio and video recognition are well formulated and developed, the implementations have faced considerable amount of trouble. When such systems get deployed in a classroom setting with multiple students taking the exam in the same room simultaneously, students occasionally become falsely flagged for infractions committed by others exam-takers. Whispers or messages passed by surrounding people can occasionally trigger the response against closely seated innocent students. Admittedly, the problem can be solved if students could be placed into separate rooms or were allowed to take the test from home, yet such decisions hurt the ability of the examiner to resolve mounting suspicions around specific students. Overall, the AI proctoring finds itself in a predicament where it is needed because of cheating and still its use becomes increasingly frustrating when a cheater is present in the exam room. Hopefully, such tools will be streamlined after a prolonged use period under human supervision, making it more appealing to employ widely in educational setting.

Similarly, AI Driven question creation is a relatively new development, aiming to leverage the same generative capabilities of LLM used to cheat in examinations to build more tests. In principle, this method works by prompting the LLM to create questions for some extensive source material that is provided. Generally, such algorithms do demonstrate the emergent characteristic of text comprehension, enabling them to execute arbitrary instruction like question generation, in a way that is not entirely understood. The example of such implementation can be seen [1]. Still, the generated questions occasionally exhibit the common generative AI problems, such as catastrophic forgetting and general hallucinations. As a result, additional human attention is needed to review generated questions, with special care being paid to factual accuracy of created tasks with respect to the provided source material, as questions are sometimes created concerning the same topic but using details taken from a different source material.

One of the less useful technologies turned out to be AI writing assistance. This idea dates back to early 2000s when first simple ML models were introduced in more sophisticated mobile phones with the purpose of spell-checking texts. Then, by mid 2010s lots of smartphones already came with predictive text generation that helped compose messages, one word at a time. Finally, last couple of years have seen rapid growth of LLM that are able to generate extensive and cohesive texts from a miniscule prompt. While spelling proofing can be arguably useful in education, full generative capabilities only open up a way to academic dishonesty. Moreover, it is difficult to ensure that even automated grammatical correction will be used by the student in a learning manner. In the end, such tools remain open to abuse, while contributing little to academic value.

Next family of tools, generative AI assessments, represent a new way to evaluate learning, using large language models to create personalized and adaptable tests. This approach offers potential benefits like automating tasks and providing immediate feedback to students. However, relying solely on AI for assessments has limitations. AI models cannot fully replace human judgment, especially in areas requiring creativity and critical thinking. AI models can struggle to understand nuances of language and culture, leading to misinterpretations and potentially offensive or insensitive assessments. Additionally, AI models can produce incorrect or unrealistic outputs, especially when dealing with complex concepts. Their performance is also heavily dependent on the quality of the data they are trained on, which can lead to biased and inaccurate assessments. Ultimately, without careful development and deployment, AI-based assessments could become a tool that measures conformity rather than actual skills, potentially leading to unfair results. If students skillfully use prompts, they can bypass AI assessor altogether, tricking it to obtain a passing grade. Assessment based on AI work does not provide an understanding of whether the student actually understood the topic and was able to complete the assignment independently.

Equally troubled is the application of LLMs for discussion practice. AI systems can generate realistic conversations, mimicking real-life interactions, allowing learners to practice various communication skills in a diverse range of contexts. This also provides immediate feedback on mistakes, helping learners quickly improve their grammar, and vocabulary. However, while AI offers significant benefits, it cannot fully replace human interaction in the learning process. Human teachers possess a nuanced understanding of language and can tailor their instruction to individual learners' needs. Additionally, AI ability to fully replicate human intelligence and the natural flow of conversation remains limited, sometimes leading to artificial or stilted interactions. Overall, AI is a viable tool for dialogue training, but it is crucial to remember that it is not a substitute for human interaction. While AI can offer a somewhat structured approach to language learning, the human element remains essential for fostering genuine communication skills and a deeper understanding of language nuances.

Finally, AI plagiarism detectors are also notable. Such tools promise to identify unoriginal work, leveraging their ability to process vast amounts of text data quickly and efficiently. However, these tools have significant limitations. They often effectively compare text against a statistical average of the generated examples. This can lead to misidentification of legitimate content as plagiarized, resulting in false accusations [3]. Additionally, they primarily focus on textual matching, overlooking the originality of ideas and concepts. This means they can't detect instances where someone reuses another's ideas without proper attribution. Furthermore, current AI detection tools face a fundamental challenge: lack of accuracy. They struggle to distinguish between human-written and AI-generated text, leading to high rates of both false positives and false negatives. These tools can provide an estimated probability of AI-generated text, but this probability may not make any sense because usually there is nothing objective behind it. This creates a false sense of security, as superficial markers like word choices, grammar, and logical flow are easily manipulated by advanced AI, rendering these detectors ineffective. The main problem is that no clear giveaways reliably differentiate human writing from sophisticated AI output. In conclusion, while promising, AI detectors rely on statistical analysis and superficial cues, their inability to discern original ideas, and their struggle to differentiate human writing from advanced AI output makes them fall short of their intended academic purpose.

Therefore, it was decided to demonstrate the inaccuracies of human text recognition by AI detectors on a set of 5 extracts from the articles written by us between 2021 and 2023. For an experiment, a set of 12 freely accessible detectors were chosen from across the web. Across 5 article pieces, the responses of the detection services were recorded as the provided percentage of accuracy where applicable. Only the percentage of human-written text was recorded, as it is complementary to AI-written portion. Similarly, the control group of 4 AI generated article pieces was also evaluated. The articles were generated with character limit of 5000 and a topic of educational technologies. The results of the experiment are presented in the tables 1 and 2.

Table 1

AI detection experiment for human-written articles

	2021 article	2022 article, part 1	2022 article, part 2	2023 article, part 1	2023 article, part 2
Copyleaks	100.00	100.00	100.00	100.00	100.00
Undetectable.ai	100.00	100.00	100.00	100.00	100.00
Quillbot	100.00	100.00	100.00	100.00	100.00
Scribbr	100.00	100.00	100.00	100.00	100.00
Zerogpt	100.00	95.12	100.00	94.38	100.00
Writer.com ai detector	100.00	100.00	100.00	90.00	95.00
App.gptzero.me	99.00	98.00	98.00	70.00	23.00
Detecting-ai.com	66.30	68.40	89.60	87.70	69.00
Contentdetector.ai	65.12	78.89	81.03	89.71	63.75
Smodin.io	100.00	100.00	100.00	88.00	76.00
Neuralwriter.com	70.00	75.00	85.00	70.00	60.00
Ai-detector.info	28.16	28.32	29.18	28.97	25.47
Statistics					
Expected value	85.72	86.98	90.23	84.90	76.02
Standard deviation	22.30	20.82	19.51	19.67	27.47

Table 2

AI detection experiment for AI-written articles

	AI article 1	AI article 2	AI article 3	AI article 4
Copyleaks	0.00	0.00	0.00	0.00
Undetectable.ai	0.00	50.00	0.00	100.00
Quillbot	0.00	32.00	0.00	63.00
Scribbr	0.00	32.00	0.00	63.00
Zerogpt	1.29	6.14	0.62	33.03
Writer.com ai detector	28.00	86.00	70.00	85.00
App.gptzero.me	0.00	71.00	0.00	83.00
Detecting-ai.com	25.40	32.70	39.70	36.90
Contentdetector.ai	22.73	35.42	21.15	17.39
Smodin.io	45.70	45.00	20.00	33.00
Neuralwriter.com	90.00	60.00	90.00	95.00
Ai-detector.info	22.27	24.62	17.99	22.06
Statistics				
Expected value	19.62	39.57	21.62	52.62
Standard deviation	25.84	23.78	29.07	31.87

As evident from the data, AI detectors are very inconsistent. Same text may both be recognized as predominantly AI generated and mostly human written seemingly arbitrarily by different software. In addition, apparent biases of the detecting models may be observed: a single detecting service often places similar scores to various texts. Consistent near certain scores of human authenticity also raise the suspicion, that some services do little actual processing and assign scores in unexplainable fashion. Statistical evidence does not point towards random continuous distribution of scores in a closed interval [0, 100] with expected value at around 50 percent, although the standard deviations of experiment data are close to or somewhat exceeding that of random distribution. As such, low accuracy of individual assessments by AI detectors is evident, making them less than trustworthy.

Based on the overviews, the aforementioned technologies were grouped into 3 categories: useful, prospective and problematic. Useful ones are ready to be utilized as is and offer some degree of educational value. At the same time, prospective tools usually require some additional work or are yet to be proven in academic world. Lastly, problematic technologies are the ones with significant drawbacks. The resulting categorization along with additional notes is presented in table 3. Notes in parenthesis of cells include clues as to how useful and prospective tools may be tested.

Table 3

Categorization of AI technologies

Useful	Prospective	Problematic
Automatic transcription (AudioPen, some video conferencing software)	AI proctoring (mostly provided on a case-by-case basis for whole institutions)	Generative AI assessments
Computer vision (Google Lens)	AI driven question creation (can be employed with any LLM chatbot)	AI interview
	AI writing assistance (Grammarly)	AI plagiarism detection

In conclusion, this paper has identified eight AI powered technologies that are related to education. Automatic transcription and computer vision were found to be mature enough to be used as is, while AI applications in proctoring, question creation and writing assistance require some additional human control over the results produced by AI. Evaluated instances of AI assessment, AI-driven interviews and plagiarism detection were found inconsistent and too unpredictable.

References

1. Аксёнова, Н.В. Применение инструментов генеративного искусственного интеллекта при создании обучающих материалов по английскому языку / Н.В. Аксёнова, Д.В. Шепетовский // Открытое и дистанционное образование. – 2023. – № 2(84). – С. 47–52. – DOI 10.17223/16095944/84/8. – EDN WXYUYX.
2. Assessments, E. Top 31 AI EdTech tools that are changing education in 2024 // Online Skill Assessment. Online Exam Software. Eklavvya.com. Splashgain Technology Solutions Pvt Ltd. – 2023. – URL: https://www.eklavvya.com/blog/ai-edtech-tools/ (date of access: 10.10.2024).
3. Edwards B. Why AI writing detectors don't work // Ars Technica. – 2023. – URL: https://arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/ (date of access: 11.10.2024).
4. Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. / I. Goodfellow, Y. Bengio, A. Courville – London, England : MIT Press, 2016. – 781 p.

Scientific supervisor: Aksenova N.V., PhD in Literature, docent.