https://doi.org/10.53656/math2024-3-4-imp

USING AI TO IMPROVE ANSWER EVALUATION IN AUTOMATED EXAMS

Georgi Cholakov, Asya Stoyanova-Doycheva
University of Plovdiv "Paisii Hilendarski" (Bulgaria)

Abstract. The objective of the research is to enhance the functionality of the Evaluator software agent within the Distributed eLearning Center (DeLC), a platform providing extensive support for e-learning activities. This paper is focused on presenting the latest step in the evolution of one software agent (Evaluator Agent), which helps teachers during evaluating students' exams, being responsible for evaluation of short free text answers, provided by the students. To accomplish its job, the agent has its own dictionary, created from a set of words and expressions, provided by test creators for each question of that type, and from the words of each student's answer, marked as successful by the teacher. Now this agent's effectiveness could be increased by extending its dictionary with much larger knowledge base or so-called Large Language Models, made accessible through third party AI, as ChatGPT. Experiments and results are provided to compare the change in agent's behavior after the integration with the AI system. The results presented in this work are from first experiments and are deliberately limited in number to enable manual verification.

Keywords: e-learning; software agents; automated evaluation; LLMs; ChatGPT

1. Introduction

With the widespread adoption of e-learning as a standard for distance education, nearly every school and university endeavors to enhance their educational processes through various platforms. Popular choices include Google Classroom[1], Moodle[2], Blackboard Learn[3], and others. Selecting the right platform involves considering a multitude of factors such as price, functionalities, support efforts, interface user-friendliness, and more. However, when the desire for partial or overall control over developed components arises, investing in a custom solution becomes increasingly appealing. Opting for a custom solution provides the flexibility to implement only the necessary functionality in a way that aligns with the institution's specific needs, without

requiring the institution to adapt its processes to fit the business logic of a third party provided system. These considerations played a significant role in the decision-making process for many institutions, including the Faculty of Mathematics and Informatics at Plovdiv University "Paisii Hilendarski", Bulgaria.

The concept was developed years ago, and the established system has been fulfilling our needs for over a decade. Originating as the Distributed e-Learning Center[4] (Doychev 2013), it commenced as a research initiative focused on crafting a novel, context-oriented, and adaptive architecture. Its primary objectives revolved around meeting our requirements for distance learning, exams conduction, and various educational and organizational activities. A significant emphasis was placed on delving into the development and experimentation of diverse prototypes within the e-learning domain.

Through a series of iterations, a hybrid service and an agent-oriented environment were fashioned to deliver educational materials and electronic services. Creating an in-house customized solution provided the added advantage of granting access to the code base for researchers within the institution. This accessibility facilitated the continuous development, reengineering, and enhancement of many features. It allowed internal exploration of the system's workflows, stored data, and the execution of analytical processes to extract valuable information and knowledge.

The main goal of this system was to elevate the educational process's quality by offering personalized and interactive services that encourage creative thinking among students. This initiative was prompted by the increasing standards in university education, advancements in technology, and the heightened expectations of students regarding the excellence of their education. The system aimed to involve students in a self-education approach that is personalized, creative, and adaptable, thereby fostering their activity and collaboration. Additionally, the system was intentionally designed to be open for extensions and experimentation with prototypes, such as service and agent-oriented architectures. This aligns it with the interactive, reactive, and proactive educational processes observed in other systems (Goyal & Krishnamurthi 2019; Farzaneh et al. 2012; Rani & Vyas 2015).

Throughout its life cycle, the architecture experienced expansion through the incorporation of various subsystems, including IntelliDeLC. This addition ensures the establishment of a personalized e-learning environment characterized by both reactive and proactive behavior. Proactivity, which enhances usability and friendliness, is accomplished by reinforcing the service-oriented architecture with intelligent components – essentially, software agents. This extension based on agent-oriented principles establishes an environment housing these continually developed and improved software agents. Details regard-

ing their functionalities, behavior, and the latest advancements can be found in (Cholakov 2020). Recently, amid the era of pervasive artificial intelligence (AI), an exploration was initiated to assess how these agents could benefit from the integration of AI systems, already provided by third parties. For the current study's purpose, such system would be useful to provide easy access to underlying large language models (LLMs), so the idea is not primarily using AI methods for adding new intelligence to our agents but use it as a mediator to essentially get structured and narrowed data from its LLMs, avoiding direct interaction with these large models, which would introduce new unwanted complexity in the agents.

2. Related works

The latest advancements indicate that incorporating AI in education, especially in e-learning, has emerged as a prevailing trend – a contemporary and essential approach. It has become a must-have solution for institutions that aim to stay abreast of new developments. This approach seems to be widely adopted, with various fields seeking to leverage its benefits. While precise statistics are lacking, a general assumption suggests that AI is primarily employed to assist learners on their educational journey, although its applications extend beyond this singular focus. Amongst the areas where AI is used in e-learning are:

- in the context of enhancing the learning process, such as enabling adaptability (Benkhalfallah & Laouar 2023);
- utilizing chatbots to expedite the learning process (Benachouret al. 2023) and subsequently assessing the efficacy of each functionality (Raghavendrachar et al. 2023);
- enhancing the accessibility of e-learning and fostering academic connectivity (Sinha et al. 2021);
- customized learning trajectories (Tapalova et al. 2022) and adaptive evaluation methods (Tanjga 2023);
- for employing conversational agents in a classroom setting (Alfehaid & Hammami 2023);
- developing virtual assistants/educators, even exploring the possibility of substituting teachers in instances of shortages (Muzurura et al. 2023);
- and more, including the automation of educational processes through the generation of teaching materials, curriculum development, training,

and the assessment of student performance (Udayakumar et al. 2022), among other aspects.

Direct usage of LLMs for educational purposes and estimation/evaluation also seems a popular idea. Since the creation of tests and quizzes for students could be time-consuming, cognitively exhausting, and complex, in help comes automatic generation of such self-assessment quizzes based on lecture material using a large language model (LLM) to reduce lecturers' workload (Meißner et al. 2024).

Fagbohun et al. 2024 present a potential use of LLMs in assessment to achieve consistency in grading – as variability may occur not only across different evaluators but also within the assessments of a single instructor over time, undermining the reliability and fairness of the grading system; to avoid bias – it could come from preconceived notions about student capabilities or unconscious preferences; get personalized feedback – it plays a crucial role in the educational process by providing students with specific guidance tailored to their individual needs.

Application of automated evaluation with the help of LLMs is presented in (Kostic et al. 2024) for the automated evaluation of student essays, exploring the effectiveness of LLMs in assessing German-language student transfer assignments, presenting the difference in their performance with traditional evaluations by human lecturers. The research shows the gap between the capabilities of LLMs and the nuanced requirements of student essay evaluation, pointing out the necessity for ongoing research and development in the area of LLM technology to improve the accuracy, reliability, and consistency of automated essay assessments in educational contexts.

Kashi et al. have emphasized the potential of using large language models (LLMs) like ChatGPT, enhanced with domain-specific expertise, to achieve more precise scoring. LLMs distinguish themselves from traditional AI models through their extensive knowledge base, adaptability, and context awareness. They have shown exceptional potential in evaluating and scoring complex text-based responses, a task that has historically been difficult for automated systems due to the richness and variability of human expression.

Certainly, some features listed above may seamlessly integrate into one system, but may not align well with another. This holds true for the DeLC portal as well. While some of the mentioned areas are under consideration for integration into DeLC, the current study is directed to a potentially valuable application of AI's access to LLMs, thus enhancing automated assessment functionality of students' exams. Automated systems for assessment are not rarely seen among e-learning systems and infrastructures. Consequently, DeLC boasts an internal implementation tailored to its requirements, devel-

oped in the form of a software agent, whose improvement is described in this article.

3. Materials and methods

The experiment is focused on enriching the existing functionality for automated evaluation – that's the job Evaluator agent is currently doing. This agent relies on its own knowledge base or dictionary, formed by keywords (also synonyms) and expressions, provided by the test creator for each question – detailed description of its algorithms is presented in (Cholakov 2013). By utilizing third party AI, we leverage from extensive knowledge base and capabilities, enhancing the skill set of this agent and evaluating its performance thereafter. Anticipated outcomes include a more precise evaluation of the answers – since students often use words and phrases that are not quite technical, yet true, if the test creator of the question has provided only technical keywords and expressions, the evaluation value tends to be under the one that a human would give. Comparing student's answer with the one generated by AI, such as ChatGPT[5], could improve the quality of evaluation for technically non-precise answers.

The Evaluator Agent (EA) is intended to assist teachers in evaluating students' electronic tests. DeLC's test engine has a built-in system service for automated assessment of "multiple choice questions", e.g., answers with radio buttons and checkboxes. EA specializes in analyzing responses to short free-text questions. It assigns a rating to each answer (integer points), deferring the final decision to the teacher – this rating is in the range from 0 to maximum points that are given for a particular question's answer (every question has a maximum number of points that could bring to the student and is set by the test creator, usually depending on the complexity of the question).

The workflow goes in the following manner - when external assessment for short free-text answers is required, the test engine sends a request for expert assistance to EA, which then utilizes its knowledge base to search for matches generated from keywords and phrases associated with each test question, typically provided by the test creator. The precision of these keywords directly impacts the quality of the agent's results, underscoring the importance of effectively "educating" the agent. We rely on the competence of the test creator (usually the teacher) to set the appropriate keywords and phrases for each question, currently there is no automated mechanism for validating this data. In the knowledge base, keywords carry no priorities; they are treated equally in searches.

Currently, EA employs two distinct algorithms for calculating points, as mentioned above. Throughout the evaluation process, EA also takes into consideration the points previously assigned by the teacher for the answers to

that specific question in prior exam runs. This approach allows EA to refine its estimation based on the teacher's style and approach. After evaluation, EA stores data about each answer, including the awarded points. Recent details on the work of this agent can be found in (Cholakov 2020).

Now the idea and the goal of this study is to try to benefit from using third-party AI and utilizing its much larger data dictionary for comparing answers to the results from it. It would help enriching the knowledge base of Evaluator Agent and make it more "educated".

4. Enhancement of functionality for estimation in Evaluator Agent

The quality of estimation in EA depends currently on the keywords/synonyms and phrases, provided by the test creator for each question. But it turns out that this approach is far from perfection. Well, it was a good start at the beginning years ago, but now we are looking for something more sophisticated and precise. The main reason is that a human couldn't easily cover all potential words and terms used by the students in their answers – of course, it's common case that one question maps to more than one correct answer. And which terms the student will use to answer is hard to predict, as our experience found out.

Trying to find a generic approach our research came to the idea to compare the answers with such provided by third-party AI, e.g., ChatGPT which is our current choice, searching for matches. Generally, the idea is to extend EA's functionality in a way to be able to ask ChatGPT the same question, answered by the student, and compare result with the answer for similarities – based on this comparison EA could estimate final points. For the sake of clarity, we must state here that in the current implementation the comparison is only lexical rather than semantical; this is one of the reasons the final score to be set by a human (teacher). The comparison between student's answer and ChatGPT's answer is done using the same two algorithms already implemented in Evaluator Agent's functionality. Expanding this agent in such a manner would provide indirect entry to a vast knowledge repository. This is due to the large volume of data utilized by ChatGPT, which is significantly incomparable to the current data scale used by the agent.

As a result, the anticipated precision of estimation would see a notable increase. In more details, EA agent will have to send request to AI for each question and store the response. Every student's answer to this question will be searched for similarities with the response from AI, apart from current matching with keywords and phrases added by the test creator. Storing the response helps to avoid negative impact on the estimation performance. Furthermore, the correct answer from AI is relatively static, at least it seems so

for couple of months, and making request for each question on every estimation would bring a significant overhead. The AI response for each question is planned to be refreshed on configurable interval of time – e.g., three months, which is approximately the duration of one semester in our university.

During research phase not only ChatGPT was considered as a potential solution. Among the most popular solutions tested were also Perplexity[6] and Google Gemini[7] (known as Bard) – many comparisons are present in Internet, among good ones are (Horsey 2023; Java 2023). Here are some arguments supporting our choice:

- Google Gemini is currently undergoing development, and in the foreseeable future, it has the potential to serve as a compelling alternative. It generates additional information, although at present, this doesn't contribute extra value to our experiment. Nevertheless, we will monitor its progress closely, especially since it excels in providing answers with real-time access to Google. In contrast, ChatGPT (with GPT 3.x) occasionally provides outdated information. Additionally, Gemini is currently free, whereas the new version of GPT 4 comes with a cost. Another noteworthy feature is Gemini's ability to produce multiple answers or variations for a single question. This feature is particularly intriguing for experiments in our field, as it allows us to explore various responses in a single roundtrip, potentially enhancing performance and yielding more results for further analysis.

- Perplexity stands out as an excellent option, boasting an advanced answer engine that considers the entire conversation history. Leveraging predictive text algorithms, it efficiently produces concise responses from various sources. This approach proves beneficial for generating answers that are closely tied to a specific context. Similar to Gemini, Perplexity offers real-time information from multiple sources, distinguishing itself from ChatGPT in this aspect. Opting for Perplexity as our second choice in the experiment is likely to yield more topic-oriented results, aligning with our objectives;

- ChatGPT endeavors to emulate human conversation, with its training methodology centered on learning from human feedback. Engaging with ChatGPT provides the opportunity to tailor the search for answers, allowing for a spectrum ranging from more deterministic to more creative responses. This adaptability proved particularly intriguing for our experiment, leading us to prioritize ChatGPT due to its simple communication interface, strong overall support, and recent advancements.

A drawback is that the new version requires a paid subscription, unlike Gemini and Perplexity. However, considering the affordability of the subscription and our primary focus on result accuracy, this factor does not significantly impact our decision.

Integration architecture is depicted on fig. 1. The communication between Evaluator Agent and ChatGPT relies on REST calls, since the existing ChatGPT API exposes its functions in this way. Trying to minimize the changes in already working components in the architecture, and thus avoid new bugs, the communication was implemented in a new software component, which plays the role of adapter – it's responsible for all details in communication between EA and the third-party solution, ChatGPT. Hence, any modifications to technical details in REST calls or the replacement of ChatGPT with another AI provider will not impact the EA. The adapter is designed to seamlessly accommodate such implementation changes, ensuring the continued functionality of the agent despite alterations to the underlying technical components. Moreover, this decoupling would enable the utilization of more than one AI provider as a source for answers comparison in the future, making the implementation transparent to EA, as illustrated in fig. 1.

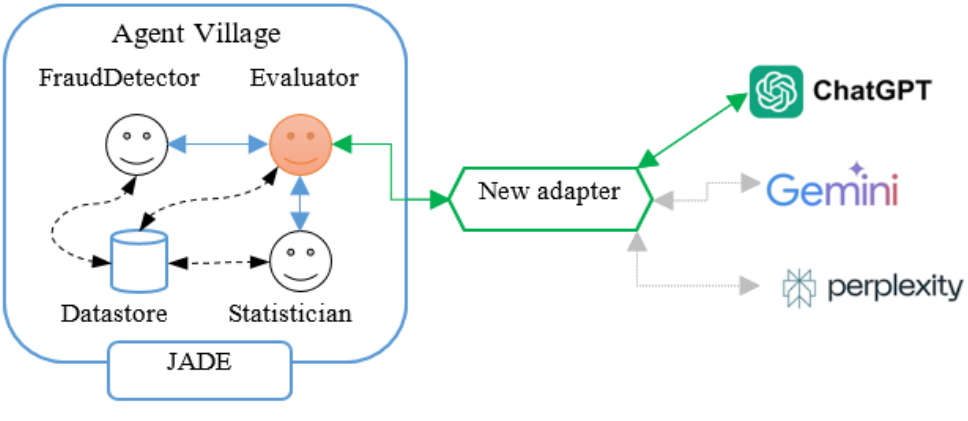

Figure 1. Improved architecture with access to AI providers, e.g. ChatGPT, Gemini, Perplexity

Currently, the integration with multiple systems is not in focus, the implementation is specifically concentrated on integrating with ChatGPT only. Given the need to fine-tune numerous parameters for achieving reliable results, the possible process of adding multiple integrations will be carried out

iteratively. This approach allows for maintaining control over the incremental complexity of the system.

Fig. 2 illustrates that both the request and response are straightforward and even human readable. The results generated are indeed precise. We could pay more attention to temperature attribute of the request. This attribute varies from 0 to 2, with higher temperatures leading to more random outcomes and lower temperatures yielding more predictable results. For more targeted and consistent answers, we figured out empirically that keeping the temperature below 0.7 generates results with more terms, which is applicable for the goal of estimation; otherwise, the responses might become too verbose, and this could lead to distraction of our agent – it works best with smaller set of keywords. On the other hand, lowest values of this attribute are producing sometimes formulas, which most of the time is not applicable for test questions requiring short, free-text answers.

Figure 2. Raw result from ChatGPT API to a particular question

Ultimately, the optimal temperature setting is context-dependent and requires extensive experimentation and statistical analysis, which is an ongoing process. The results on fig. 3 summarize the tests for choosing the appropriate value of temperature parameter, iterating over 89 questions, sending each one to ChatGPT with all possible values for the temperature with step

of 0.1. As indicated by the graph, the most accurate results for our purposes were achieved around the middle of the range, leading us to set the temperature parameter to 0.7. This value may be adjusted in the future, based on real-world results, to empirically compare different behaviors and outcomes.

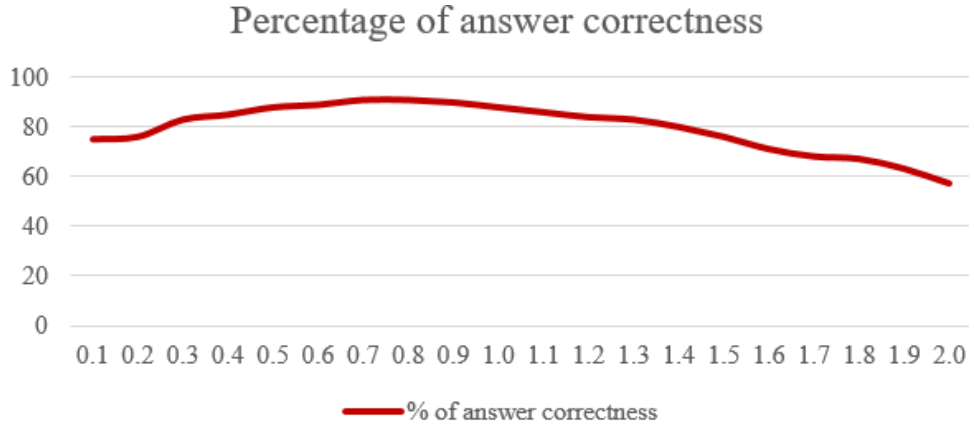

Figure 3. Results accuracy of the answers by ChatGPT

5. Results

The results of the simulation would provide clarity on whether it is worth it or not to go ahead with the implementation of the enhancement in production environment. The experiments and testing of this enhancement were conducted using a dataset from previous exams runs from the past academic years for the subject "Database Management Systems", which contains questions that were answered with short free text by students, the answers, the points estimated by EA, and the points given by the teacher. The number of questions was 89, answered by 235 students, with 864 answers in total (up to 4 answers per student). The amount of answers was limited to minimize the manual checking of responses generated by ChatGPT, because we needed to analyze what is returned and is it related to the question, and eventually reliable. On the other hand, it should be enough to provide evidence whether this implementation would increase the correctness of final estimation provided by the Evaluator Agent. For each tuple of the dataset, we added the response of the question from ChatGPT, and that response was used as an input of the EA's algorithm for searching matches between student's answer and response from ChatGPT. Example of how these tuples look like is the following (used data from fig. 2):

```
{
  ‘‘test_question’’: ‘‘What are the properties of transactions?’’,
  ‘‘answer’’: ‘‘Transactions should be atomic - it means it’s a
              logical unit of work...’’,
  ‘‘keywords’’: ‘‘atomicity, consistency, isolation, durability’’,
  ‘‘points1’’: 3.1,
  ‘‘points2’’: 3.7,
  ‘‘points_given’’: 4,
  ‘‘max_points’’: 4,
  ‘‘ai_response’’: ‘‘The properties of transactions in the context
                   of databases are...’’,
  ‘‘ai_points’’: 3.3
}
```

To make it clear, `answer` element contains the student's input; `keywords` element contains the lexemes, provided by test creator for EA; `points1` and `points2` are the estimated points by the two algorithms in EA (these algorithms are explained in (Cholakov 2013), but in short they are inspired by the popular ones Soundex[8] and Metaphone[9]); `points_given` is the final estimation from the teacher; `max_points` is the maximum number of points for the answer to this question; `ai_response` contains the response from ChatGPT. For the sake of truth, we must admit that this is simplified demonstration, usually in keywords there are more terms with synonyms, and expressions, other system attributes that contain details regarding estimation methods are also omitted for simplicity. Finally, `ai_points` attribute contains the estimated points using AI response as a base for matching the answer – these points are calculated based on the matching ratio from EA algorithms. Fig. 4 shows approximate comparison between the estimations using keywords and AI response. It presents the points estimated for selected four typical questions with equal maximum points of 4 each. The questions on the graphics are only a representative sample for visualization, a subset of all such used for the experiment.

Elements `points1` and `points2` show the estimation from the two algorithms using keywords as knowledge base, `ai_points` attribute is the estimation from the second (most used) algorithm using AI response as knowledge base (red line), and `points_given` are the finally given points from the teacher. As the lines show, the estimations are pretty close in the chart, but the expected differences are when the answers are not verbose and not correct entirely – the simulation tests are still ongoing and at least at the beginning that's what they reveal. Still there is much work to be done and EA will need a tuning to produce finally expected results.

To complete the estimation comparison, we have to add that when the answer is entirely correct or entirely wrong, all estimations are similar and exact – the limit values of the interval seem to be easily estimated and are not a challenge.

On the other hand, the correctness of the ChatGPT's answer is another topic that needs attention. All aforementioned questions participating in the experiment were tested empirically with ChatGPT, Gemini and Perplexity, and returned answers were manually validated. At first glance, all those did it very well, but this manual approach is applicable when the number of questions is relatively low – as it grows it would put an additional load on the test creators, but initially it was the only reliable approach to test that answers' correctness.

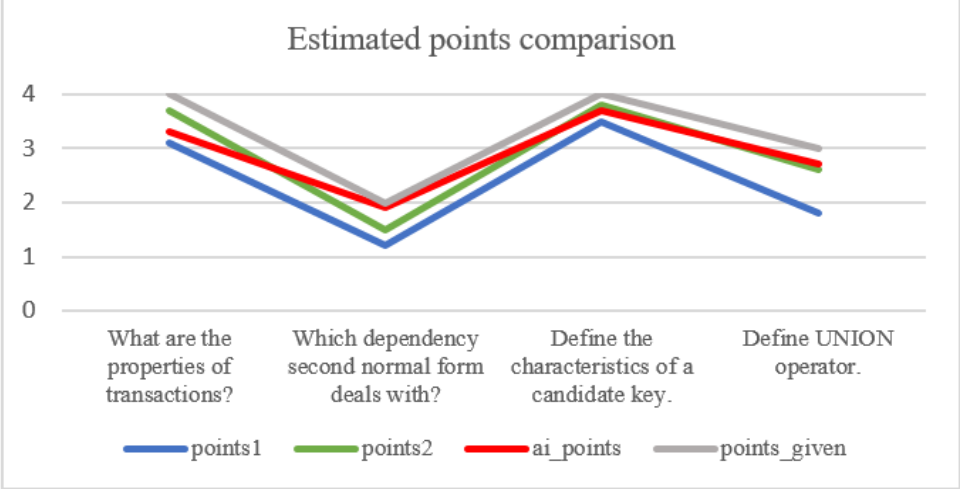

Figure 4. Estimation comparison between algorithms using keywords and AI response for matching

6. Conclusions

Maintaining a high standard of education quality depends on numerous factors that extend beyond the teaching abilities of staff. It also relies on students recognizing the importance of dedication and self-motivation in skill-building. As technology advances, the entire university education process must align with current standards and trends. This isn't solely to capture students' attention but also to create a conducive environment for cultivating professionals. This encompasses various aspects, including the quality

of examinations, a key component of which is ensuring a fair and accurate assessment process – whose automation improvement is the subject of the study. Significantly, enhancing the software components that operate in the background, supporting the overall educational process, constitutes a crucial aspect of building the foundation of modern education.

Improving automated assessment through the usage of LLMs is expected to introduce contemporary functionalities to the responsibilities of the Evaluator Agent. Expanding its own knowledge base poses a considerable challenge, that's why integration with external systems is preferred. This strategic approach allows for a concentration on refining precision in estimation methods rather than dedicating resources to the construction and expansion of own knowledge base.

But there are still many unclear points as the experiment goes that need to be addressed and tested to assess what exactly the benefits of the described implementation will be.

Acknowledgments

This study was made possible thanks to the project European Union – NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BGRRP-2.004-0001-C01.

NOTES

1. Google Classroom, Google for Education. https://edu.google.com/workspace-for-education/classroom [accessed 16 January 2024].
2. Moodle. https://moodle.com/ [accessed 8 February 2024].
3. Blackboard. https://blackboard.com/en-mea/teaching-learning [accessed 5 March 2024].
4. DeLC, Distibuted e-Learning Center. http://delc.fmi.uni-plovdiv.bg/ [accessed 13 January 2024].
5. ChatGPT. https://chat.openai.com [accessed 13 November 2023].
6. Perplexity. https://www.perplexity.ai/ [accessed 13 February 2024].
7. Gemini (chatbot) https://en.wikipedia.org/w/index.php?title=Gemini_(chatbot)&oldid=1206112332 [accessed 11 February 2024].
8. Soundex. https://en.wikipedia.org/wiki/Soundex.
9. Metaphone. https://en.wikipedia.org/wiki/Metaphone.

REFERENCES

ALFEHAID, A., HAMMAMI, M., 2023. Artificial Intelligence in Education: Literature Review on The Role of Conversational Agents in Improving Learning Experience. *International Journal of Membrane*

Science and Technology, vol. 10, pp. 3121 – 3129. DOI 10.15379/ijmst.v10i3.3045.

BENACHOUR, P., EMRAN, M., ALSHAFLUT, A., 2023. *Assistive Technology and Secure Communication for AI-Based E-Learning.* ISBN 978-1-66848-938-3. DOI 10.4018/978-1-6684-8938-3.ch001.

BENKHALFALLAH, F., LAOUAR, M., 2023. Artificial Intelligence-Based Adaptive E-learning Environments. In: *Novel & Intelligent Digital Systems*, Proceedings of the 3rd International Conference (NiDS 2023), pp. 62 – 66. ISBN 978-3-031-44096-0. DOI 10.1007/978-3-031-44097-7_6.

CHOLAKOV, G., 2013. *Hybrid Architecture for Building Distributed Center for e-Learning.* PhD. Thesis, Plovdiv University "Paisii Hilendarski", Plovdiv, Bulgaria.

CHOLAKOV, G., 2020. Approbation of software agent Evaluator in a nonspecific environment for extension of its purpose. In: *2020 International Conference Automatics and Informatics (ICAI)*, pp. 1 – 5. DOI:10.1109/ICAI50593.2020.9311346.

DOYCHEV, E., 2013. *Environment for Provision of eLearning Services.* PhD. Thesis, Plovdiv University "Paisii Hilendarski", Plovdiv, Bulgaria.

FAGBOHUN, O., IDUWE, N.P., ABBDULLAHI, M., IFATUROTI, A., NWANNA, O.M., 2024. Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. *Journal of Artificial Intelligence, Machine Learning and Data Science.* vol. 2, no. 1. DOI:10.51219/JAIMLD/oluwole-fagbohun/19.

FARZANEH, M., VANANI, I.R., SOHRABI, B., 2012. Utilization of Intelligent Software Agent Features for Improving E-Learning Efforts. *International Journal of Virtual and Personal Learning Environments*, vol. 3, no. 1, pp. 55 – 68. DOI 10.4018/jvple.2012010104.

GOYAL, M., KRISHNAMURTHI, R., 2019. Pedagogical Software Agents for Personalized E-Learning Using Soft Computing Techniques. In: *Nature-Inspired Algorithms for Big Data Frameworks.* IGI Global. ISBN 978-1-5225-5852-1. [accessed online 16 February 2024].

HORSEY, J., 2023. Perplexity vs Bard vs ChatGPT. *Geeky Gadgets* https://geeky-gadgets.com/perplexity-vs-bard-vs-chatgpt [accessed 29 February 2024].

JAVA, Words at Work by, 2023. *Battle of the AI's: Chat GPT vs. Perplexity AI vs. Google Bard. Medium* https://medium.com/@jamva/battle-of-the-ais-chat-gpt-vs-perplexity-ai-vs-google-bard-bca76474a30d [accessed 19 February 2024].

KASHI, A., SHASTRI, S., DESHPANDE, A.R., DORESWAMI, J., SRINIVASA. G., 2016. A Score Recommendation System Towards Automat-

ing Assessment In Professional Courses. *IEEE Eighth International Conference on Technology for Education (T4E)*, pp. 140 – 143. DOI 10.1109/T4E.2016.036.

KOSTIC, M., WITSCHEL, H.F., HINKELMANN, K., SPAHIC-BOGDANOVIC, M., 2024. LLMs in Automated Essay Evaluation: A Case Study. *Proceedings of the AAAI 2004 Spring Symposium Series*, vol. 3, no. 1, pp. 143 — 147. DOI 10.1609/aaaiss.v3i1.31193.

MUZURURA, O., MZIKAMWI, T., REBANOVAKO, T.G., MPINI, D., 2023. Application of Artifical Intelligence for Virtual Teaching Assistance (Case study: Introduction to Information Technology). *International Research Journal of Engineering and Technology*, vol. 10, no. 9, pp. 276 – 283. https://issuu.com/irjet/docs/irjet-v10i938

MEISSNER, N., SPETH, S., KIESLINGER, J., BECKER, S. 2024. EvalQuiz – LLM-based Automated Generation of Self-Assessment Quizzes in Software Engineering Education. *Software Engineering im Unterricht der Hochschulen 2024*. Gesellschaft für Informatik, pp. 53 – 64. ISBN 978-3-88579-255-0. DOI 10.18420/seuh2024_04.

RAGHAVENDRACHAR, S., ANAND, V.S., ANUSHREE, H., MANJUNATAN, R., 2023. E-Learning Management System with AI Assistance. *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, pp. 1233 – 1238. DOI 10.22214/ijraset.2023.56730.

RANI, M., VYAS, O., 2015. An Ontology-based Adaptive Personalized E-learning System, Assisted by Software Agents on Cloud Storage. *Knowledge-Based Systems*, vol. 90, pp. 33 – 48. DOI 10.1016/j.knosys.2015.10.002.

SINHA, M., FUKEY, L., SINHA, A., 2021. AI in e-learning. In: *E-learning Methodologies: Fundamentals, technologies and applications*, pp. 126 – 150. ISBN 978-1-83953-120-0. DOI 10.1049/PBPC040E_ch5.

TANJGA, M., 2023. *E-learning and the Use of AI: A Review of Current Practices and Future Directions*. Qeios. DOI 10.32388/AP0208.2.

TAPALOVA, O., ZHIYENBAYEVA, N., GURA, D., 2022. Artificial Intelligence in Education: AIEd for Personalised Learning Pathways. *Electronic Journal of e-Learning*, vol. 20, pp. 639 – 653. DOI 10.34190/ejel.20.5.2597.

UDAYAKUMAR, A., GANESAN, M., GOBHINATH, S., 2022. A Review on Artificial Intelligence Based E-Learning System. In: *Pervasive Computing and Social Networking*, pp. 659 – 671. ISBN 978-981-19283-9-0. DOI 10.1007/978-981-19-2840-6_50.

✉ **Dr. Georgi Cholakov, Assist. Prof.**
ORCID iD: 0000-0001-7971-8434
Faculty of Mathematics and Informatics
University of Plovdiv "Paisii Hilendarski"
Plovdiv, Bulgaria
E-mail: gcholakov@uni-plovdiv.bg

✉ **Dr. Asya Stoyanova-Doycheva, Prof.**
ORCID iD: 0000-0002-0129-5002
Faculty of Mathematics and Informatics
University of Plovdiv "Paisii Hilendarski"
Plovdiv, Bulgaria
E-mail: astoyanova@uni-plovdiv.bg