Възможностите на ChatGPT за генериране на тестове

Андриан Минчев, Надежда Ангелова,
Габриела Кирякова, Ваня Стойкова, Галя Шивачева

The capabilities of ChatGPT to generate quizzes

Andrian Minchev, Vanya Stoykova,
Gabriela Kiryakova, Nadezhda Angelova, Galya Shivacheva

Abstract:

The role of ChatGPT in education has grown over the past year. Teachers consider the possible benefits of integrating such a tool into learning activities - preparing lesson plans, lectures, or exercises, developing learning materials for training activities during classes, and generating exam tasks and tests. Despite the significant possibilities of generative artificial intelligence, the teacher's role is crucial in verifying the correctness and appropriateness of the generated educational resources.

The paper aims to present the results of a study for evaluating quiz questions generated by ChatGPT on chosen topics from the learning content of various disciplines studied by students at Trakia University.

Keywords: Large language model, ChatGPT, Test generation, Learning management systems

For contacts: Assist. Prof. Andrian Minchev, Trakia University, andrian.minchev@trakia-uni.bg

ВЪВЕДЕНИЕ

Бързото развитие на техниката и технологиите и повишените изисквания към кадрите във всички области на човешката дейност, поставя редица предизвикателства пред образователната система. Нелека е задачата на преподавателите да осигуряват висококачествено обучение, което да бъде адаптирано с темпа на технологичния напредък и повишаващите се изисквания на пазара на труда. Непрекъснатата необходимост от актуализиране на учебното съдържание „в крак“ с иновациите в различни области налага да бъдат проучени и използвани нови съвременни методи за подпомагане на преподавателите, вкл. с помощта на изкуственият интелект (ИИ). Той може да бъде от полза при редица учебни дейности – от създаването на учебен план/учебна програма по дисциплина, последващо разработване на планове на уроци и генериране на учебно съдържание (вкл. актуализирането му) в подходяща форма, до използването му при контрола на знанията на студентите. Оценяването на знанията и уменията изисква разнообразни подходи и средства – чрез тестове с различни видове въпроси, задания, изискващи работа по реални казуси или проучване по определени теми, изпълнение и решаване на конкретни задачи и други. Всеки един от тези методи може да бъде приложен за оценка според нивата от таксономията на Блум, т.е. за запаметяване, разбиране, приложение, анализ, синтез и оценка. Това изисква време и усилия от страна на преподавателя да създаде подходящи инструменти за оценка, които най-адекватно и коректно да оценят знанията и уменията на обучаемите. Процесът по създаването на тестове, които днес масово се използват в университетските системи за е-обучение (СЕО)

при текущата и окончателна оценка на знанията на студентите, също е трудоемък и отнема много от времето на преподавателите във ВУ. От друга страна разширяването на „банките“ с въпроси е от съществено значение за провеждането на изпити и обективна оценка на знанията. В тази насока са проучванията на възможностите за автоматизиране на процеса на създаване на нови изпитни въпроси в Moodle (Pehlivanova, T., Kanchev, K., 2018) (Yuan, Z. et al, 2024). Интегрирането на GhatGPT в CEO би било още една полезна стъпка като ще се автоматизират изцяло и дейностите по импортирането на генерираните въпроси в университетските платформи за е-обучение. Редица преподаватели и изследователи са търсили и предлагали решения по въпросите, свързани с приложението на ИИ в обучението, вкл. при контрола и оценката на знанията на студентите.

След появата на ChatGPT и други езикови модели като Bard, все по-често се дискутира и изследва тяхната употреба, както от страна на преподавателите, така и от страна на обучаемите. Приложението им за генериране на тестови въпроси е обект на проучване от много автори и основният изследователски въпрос е свързан с това доколко коректни са създадените по този начин въпроси и дали е възможно да се разчита изцяло на тези езикови модели. Наличните проучвания за генериране на въпроси с множествен избор в сферата на хуманната и денталната медицина показват, че едно от основните предимства е спестяването на време. В експеримент на (Cheung et al, 2023) се сравняват 50 въпроса генерирани с ChatGPT и други 50, създадени от двама професори по медицина като са използвани едни и същи източници на информация. При оценката им от независими международни оценители на тестове се доказва, че генерирането на въпроси с множествен избор от ChatGPT отнема много по-малко време спрямо това на университетските преподаватели. Статистическият анализ показва, че няма значителна разлика в качеството на въпросите между тези, изготвени от ChatGPT и преподавателите.

Положително отношение и липса на статистическа разлика между генерираните с ИИ тестови въпроси в сравнение с тези, създадени от преподаватели, са отчетени и в изследванията на (Rivera et al, 2024) и (Coskun, O., Kıyak, Y. S., Budakoğlu, I. 2024).

Не се наблюдава и значителна разлика при въпроси, генерирани с двата най-популярни езикови модела ChatGPT и Bard, както и по отношение на качеството на въпросите. Въпреки това, авторите препоръчват на преподавателите да преглеждат внимателно въпросите и да ги адаптират, за да са в съответствие с поставените учебни цели (Ahmed et al, 2024). Този проблем е отразен и от (Lu, K., 2023), според който ChatGPT може да създаде логически разумни въпроси, но не винаги съответстват на учебните цели и в повечето случаи такива въпроси са лесни за отговор, доста често започват по един и същ начин и са с еднотипни конструкции. Областите, в които той използва ChatGPT са Линейна алгебра и Астрономия и от получените резултати са направени изводи за това как могат да се прилагат и усъвършенстват инструменти с ИИ като ChatGPT в подкрепа на образованието.

От друга гледна точка, студентите си служат с ChatGPT за търсене на верните отговори, особено при електронни тестове. Това изисква въпросите и отговорите да бъдат така зададени, че да затруднят студентите (Newton, P., 2023). Препоръчва се използването на изображения, математически изчисления, въпроси, които изискват по-високо ниво за решаване на проблеми (Gonsalves, C., 2023).

ChatGPT, като популярен представител на генеративния изкуствен интелект, би могъл да бъде едно полезно решение, с което да се подпомогне дейността на преподавателя в това отношение. Статията има за цел да представи и анализира резултатите от изследване на възможността за генериране на тестови въпроси по зададени теми от учебното съдържание по различни дисциплини, чрез големи езикови модели (LLM) и конкретно ChatGPT-4. Изследването е проведено със съдействието на преподаватели и студенти в Тракийски университет и обхваща ключови аспекти като качеството на генерираните въпроси, тяхната релевантност към учебното съдържание и ефективността им като инструмент за оценяване на знанията. Търси се отговор на въпроса дали GPT-4 може да бъде използван като иновативно средство за подпомагане на преподавателите и подобряване на качеството на образователния процес.

2. ИЗЛОЖЕНИЕ

2.1. Методика на изследване

Формулировка на задачата

Методиката на изследването включва създаването с помощта на ChatGPT на банка с по 50 въпроса по 2 дисциплини като предварително са зададени раздели от учебното съдържание, които да се съдържат в генерираните въпроси. Генерираните тестове са анализирани от преподаватели като се оценява съдържанието на въпросите и се акцентира върху коректността, съвпадението с предложената тематика, яснота, приложимост. Един от тестовете е апробиран сред студенти и е зададен като тест за самоподготовка за изпит по дисциплината „Компютърни мрежи“, едновременно с тест създаден от водещия преподавател. Направено е проучване на студентското мнение, включващо и въпроси за разпознаваемост и качества на теста генериран от ChatGPT.

Генериране на тестови въпроси с ИИ и импортиране в СЕО

По дисциплината „Компютърни мрежи“ бяха зададени 10 раздела от учебната програма, към които се изискваше от ChatGPT да бъдат генерирани по 5 въпроса от тип множествен избор (с 4 възможни отговора) във формат за импортиране gift. Символите **~%50%** и **~%-100%** определят частично верни и напълно грешни отговори, което позволява на Moodle да начислява или отнема точки според избраните отговори. От генерираните 50 въпроса 40 бяха успешно импортирани в Тракийски електронен университет, базиран на Moodle. Девет въпроса се импортираха неуспешно (слети с предходния въпрос) поради техническа грешка (липса на празен ред между въпросите) при ръчно обобщаване на 10 файла с въпросите от 10-те раздела. Поради синтактично несъответствие с формата gift, използване на кавички, които не са били избегнати със специален символ, не беше импортиран 1 въпрос. Тези въпроси не бяха включени в теста. Възникнаха

проблеми и с въпроси с по 3 верни отговора, при които общата сума на процентите надвишаваше 100, което е недопустимо при Moodle. Тези въпроси също бяха коригирани след намеса на преподавателя.

2.2. Анализ на резултатите от приложението на ИИ при генериране на тестове

При генерирането на въпроси с помощта на GPT-4 по дисциплината „Компютърни мрежи“ се отчитат следните грешки: неточност при формулиране на въпросите, което предполага неразбиране и грешни отговори; неправилен словоред на български език при някои въпроси; неподходяща терминология, нетипична за дисциплината, напр. „имунитет към електромагнитни смущения“, „проводни мрежи“ или „ниска задръжка“; използване на непопулярен за терминологията на дисциплината превод на имена на устройства от английски на български език, вкл. различен превод в различни въпроси (напр. „repeater“ е преведено като „ретранслатор“ и като „повторител“ в два различни въпроса на теста); повторения на един въпрос в различните групи въпрос по раздели

Установено бе, че по дисциплината „Релационни бази данни“ от 50 предложени тестови въпроса в 47 от тях правилният отговор е с видимо и значително по-дълъг текст в сравнение с останалите посочени възможни отговори. Това бързо се забелязва и от студентите и те започват да търсят и следват тази зависимост. Не е възможно директно предоставяне на теста на студентите, т.к. ChatGPT поставя правилния отговор като първи в списъка с отговори и това налага да се включи опцията за разбъркване на отговорите при импортиране в Moodle.

Друг проблем е свързан с коректността на отговорите и връзката им с въпроса. Срещат се въпроси, при които верният отговор не винаги е изчерпателен или точен. Наблюдават се въпроси, на които част от отговорите нямат отношение към самия въпрос и веднага могат да бъдат изключени като верен отговор. Има какво да се желае и по отношение на терминологията, която не винаги е точна (на български език).

2.3. Апробиране на тестовете в студентски групи

В проведеното проучване чрез използване на тестове за самоподготовка по дисциплината „Компютърни мрежи“, генерирани от ИИ и съставени от преподаватели, участваха 19 студенти от специалност „Автоматика и компютърни системи“. Близо 80% (78,9%) дадоха правилен отговор на въпроса и разпознаха кой е теста генериран автоматично с използване на ИИ. Над 80% (84,2%) от респондентите са отговорили, че ги е затруднил повече теста генериран от GPT-4 и предпочитат да бъдат изпитвани с традиционния тест, съставен от водещия преподавател по дисциплината. Като признаци, по които са разпознали теста генериран с ИИ, студентите основно посочват неразбираеми и липса на яснота и точност при някои въпроси. Преобладаващ е отговора, че не могат да посочат точна причина, но считат че това е теста, които не е генериран от преподавателя.

3. Обобщение на получените резултати и изводи

Основните предимства при автоматизиране на дейността за създаване на тестове с помощта на GPT-4 могат да бъдат обобщени по следния начин:

ефективност по отношение на бързина и спестяване на време; разнообразие по отношение на различни варианти на въпросите; обновяване, автоматизация.

За да се прецени доколко подходящо е използването на GPT-4 за генериране на тестови въпроси, е необходимо да се отбележат и следните проблеми и недостатъци:

- Качество на въпросите: Не всички генерирани въпроси може да са с високо качество или да са напълно коректни, с използване на подходяща за дисциплината терминология. Необходим е допълнителен преглед от експерти;
- Релевантност: Въпросите не винаги може да съответстват на конкретния учебен план или учебни цели;
- Риск от генериране на неподходящи въпроси;
- Липса на човешка интуиция: LLM не притежават интуицията и опита на преподавателите, което може да доведе до генериране на въпроси, които не са съобразени с възможностите на обучаемите.

Използването на LLM за генериране на изпитни въпроси може да бъде много полезно, ако се комбинира с експертен преглед и оценка. Това ще гарантира, че въпросите са подходящи и точни за оценка на знанията на обучаемите.

Изследването доказва че LLM спестява време и повишава ефективността в процеса на създаване и оценка на знанията чрез изпитни тестове. Използва се подход за итеративно подобряване на резултата, чрез изпращане на обратна връзка до ChatGPT.

Предвижда се проучването да продължи в посока анализ на ползите от приложението на този метод на създаване на въпроси от различен тип, по други учебни дисциплини и направления, по които се обучават студенти във ВУ.

ЗАКЛЮЧЕНИЕ

Направеното изследване би подпомогнало процеса на генериране на тестове с помощта на ИИ и тяхното импортиране в СЕО, базирано на Moodle като са посочени възможни грешки и проблеми при поставяне на заданието и последващите етапи при създаването на теста в е-курса по дисциплината.

Като резултат от следващо детайлно изследване на различни аспекти от приложението на GPT-4 за автоматично генериране на въпроси, е възможно разработване на детайлни препоръки за оптимизация на процеса на създаване, одобрение и използване на изпитни въпроси, генерирани от GPT-4.

Изследването доказва потенциала на GPT-4 за иновация в образователния процес и възможностите му по отношение на оценяване на знанията на обучаемите. Безспорна е необходимостта от активна намеса на преподавателя или групи от водещи преподаватели, които да одобрят и/или стандартизират банки от въпроси за изпитни тестове. ИИ може да подпомогне част от трудоемките, рутинните дейности по създаване и актуализиране на изпитни тестове, които отнемат голям процент от времето на университетските преподаватели.

Налице са вече разработки за интегриране на ИИ в Moodle и това ще доведе до премахване на етапа на импортиране на въпросите, което изисква намеса на специалист.

Въпреки значителните възможности на генеративния изкуствен интелект, ролята на преподавателя е решаваща при проверката на правилността и целесъобразността на генерираните образователни ресурси.

Докладът е финансиран по Националната научна програма „Интелигентно животновъдство“, финансирана от МОН, съгласно подписаното споразумение № Д01-62/18.03.2021/ Регистър на ТрУ Н003-2021/18.03.2021г.

ЛИТЕРАТУРА

1. Ahmed, W., Azhari, A., Alfaraj, A., Alhamadani, A., Zhang, M., & Lu, C.T. (2024). The quality of dental caries-related multiple-choice questions and answers generated by ChatGPT and Bard language models. *Heliyon, 10, e28198*. doi: 10.1016/j.heliyon.2024.e28198.
2. Cheung, B.H.H., Lau, G.K.K., Wong, G.T.C., Lee, E.Y.P., Kulkarni, D., Seow, C.S., et al. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS ONE, 18(8): e0290691*. doi: 10.1371/journal.pone.0290691
3. Coskun, O., Kıyak, Y.S., & Budakoğlu, I. (2024). ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Medical Teacher, 1-7.* doi: 10.1080/0142159X.2024.2327477
4. Gonsalves, C. (2023). On ChatGPT: what promise remains for multiple choice assessment? *Journal of Learning Development in Higher Education, (27)*. doi: 10.47408/jldhe.vi27.1009
5. Lu, K. (2023). Can ChatGPT Help College Instructors Generate High-Quality Quiz Questions? *AHFE Open Access*, doi: 10.54941/ahfe1002957
6. Newton, P. (2023). ChatGPT performance on MCQ-based exams. EdArXiv. doi: 10.35542/osf.io/sytu3
7. Pehlivanova, T., & Kunchev, K. (2018). A Moodle Plugin for Creating a New Type of Questions. In: Proceedings of ICTTE 2018. *International Conference on Technics, Technologies and Education ICTTE 2018, October 18-19, 2018, Yambol, Bulgaria.* ISSN: 1314-9474, doi: 10.15547/ictte.2018.04.010
8. Rivera Rosas, C., Calleja López, J., Ruibal-Tavares, E., Flores-Felix, C., & Trujillo López, S. (2024). Exploring the potential of ChatGPT to create multiple-choice question exams. *Educación Médica, 24*. doi: 10.1016/j.edumed.2024.100930
9. Yuan, Z., Liu, M., Ding, S., Wang, K., Chen, Y., Peng, X., Lou, Y. (2024), No More Manual Tests? Evaluating and Improving ChatGPT for Unit Test Generation. *Proceedings of the ACM on Software Engineering*, 1(FSE), Article 76. doi: 10.1145/3660783.