TRUST AI

Оценка на информация, генерирана от изкуствен интелект

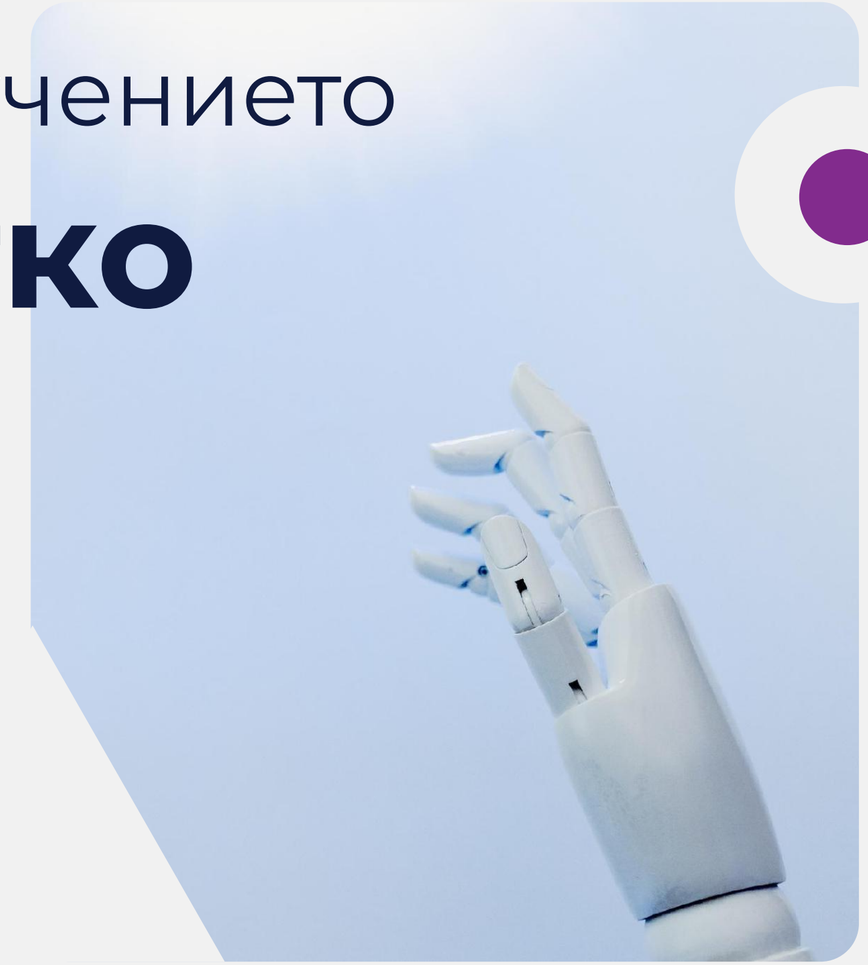
Развиване на умения за критична оценка на точността, надеждността и пристрастността на съдържанието, генерирано от изкуствен интелект

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Цели на обучението

Накратко

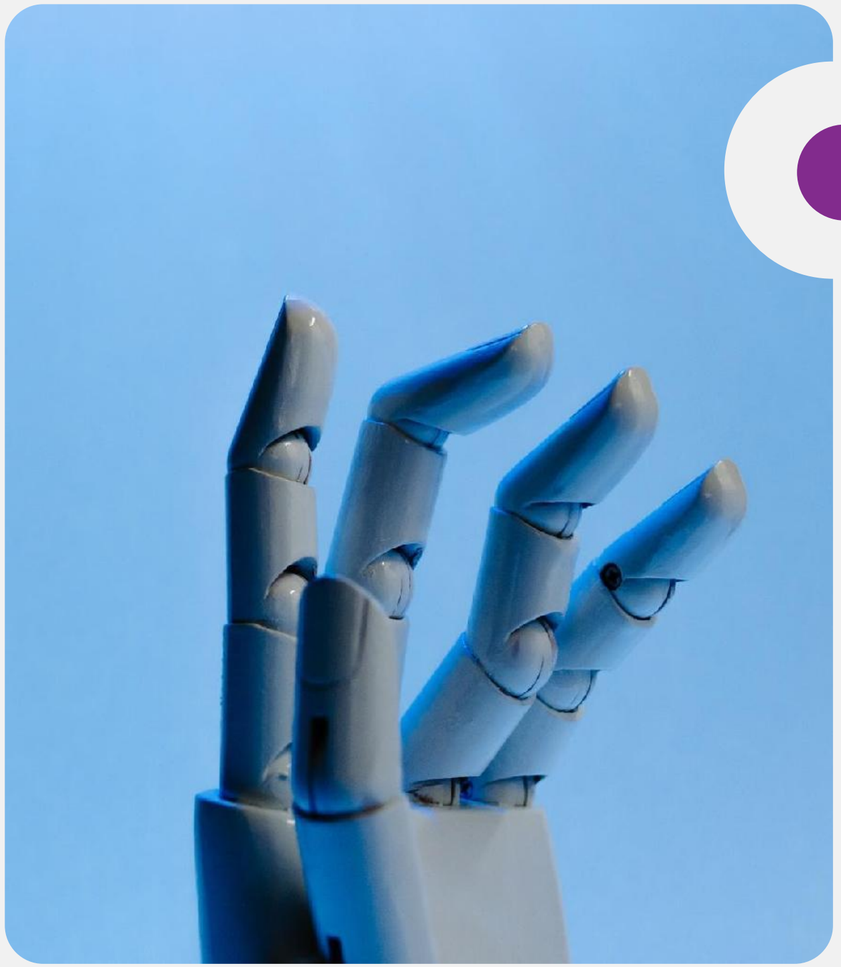


- Идентифициране на трите основни стълба на оценката на съдържанието, създадено от ИИ: точност, надеждност и пристрастност (ARB).
- Прилагайте структурирани рамки за проверка (като CRAAP или ROBOT), специално адаптирани за резултатите от генеративната ИИ.
- Разпознавайте специфични начини на отказ на ИИ, включително халюцинации, измисляне на цитати и културна пристрастност.
- Разработване на практически дейности в класната стая, които превръщат грешките на ИИ в упражнения за критично мислене.
- Въведете протоколи за прозрачност, които изискват от учениците да документират и критикуват използването на инструменти на ИИ.

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Ключови понятия и дефиниции



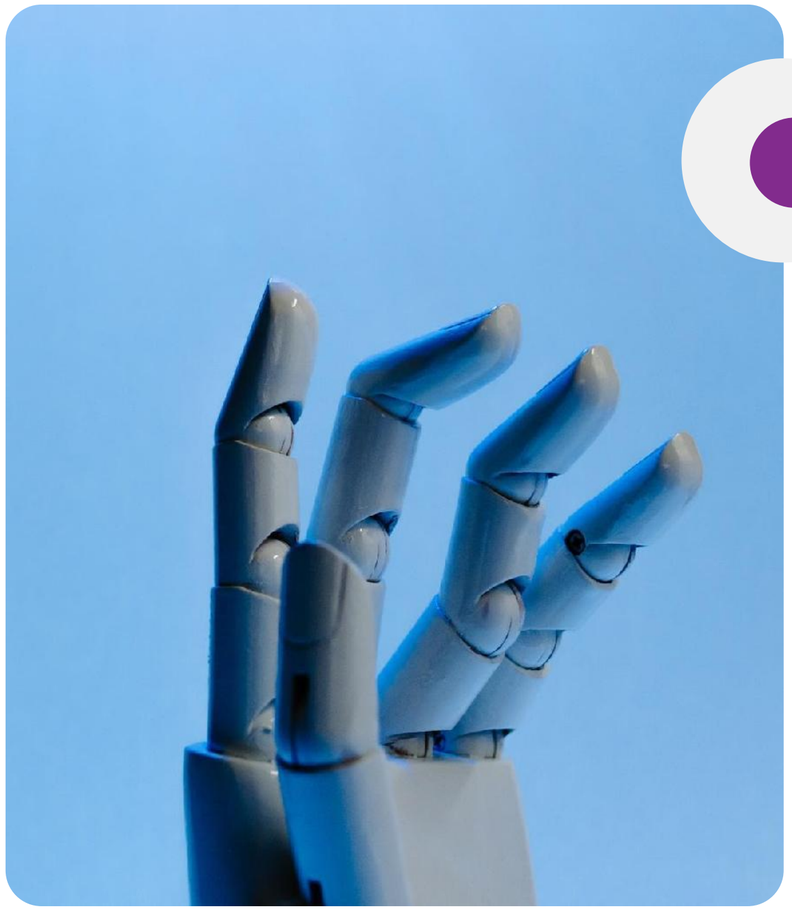
НЕОТЛОЖНОСТТА НА КРИТИЧНАТА ГРАМОТНОСТ В ОБЛАСТТА НА ИИ

- **Промяна в потреблението на информация**: Учениците използват ИИ като основна търсачка и инструмент за написване на текстове, често заобикаляйки традиционната проверка.
- **Капанът на „правдоподобността“**: Големите езикови модели (LLM) са проектирани да звучат убедително, а не непременно да бъдат верни. Те дават приоритет на езиковата вероятност пред фактическата точност.
- **Въздействие върху хуманитарните науки**: ИИ често се затруднява с нюансите, културния контекст и дълбочината на интерпретацията – основни компоненти на хуманитарните науки.
- **Академична интегритет 2.0**: Интегритетът вече не се отнася само до „неизмамата“; той се отнася до проверката на твърденията за истината на инструментите, които използваме.
- **Овластяване**: Оценяването на преподаването пренасочва фокуса от „контрол над студентите“ към „овластяване на учените“ да използват инструментите отговорно.

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Критичната нужда от умения за оценка на ИИ



ПРЕДИЗВИКАТЕЛСТВОТО:

- Инструментите за изкуствен интелект генерират убедителен и плавен текст, който може да прикрие неточности и пристрастност
- Студентите все повече разчитат на ИИ за изследвания, писане и анализ без проверка
- Традиционното откриване на плагиат е недостатъчно за ерата на изкуствения интелект

ВЪЗМОЖНОСТТА:

- Критичната оценка превръща изкуствения интелект от заплаха в мощен педагогически инструмент
- Преподавателите по хуманитарни науки са в уникална позиция да преподават нюансирана оценка
- Развитието на умения за оценяване подготвя студентите за професионален свят, интегриран с изкуствен интелект

НЕОБХОДИМОСТТА:

- Академичния интегритет зависи от проверима и надеждна информация
- Етичното използване на ИИ изисква прозрачност и критично ангажиране

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Съфинансирано от Европейския съюз

Трите стълба: точност, надеждност, безпристрастност

ТОЧНОСТ (Проверка на фактите)

- Могат ли конкретните твърдения да бъдат съпоставени с утвърдени академични източници?
- Съдържа ли съдържанието логически несъответствия?
- Измисля ли ИИ данни, дати или събития (халюцинации)?

НАДЕЖДНОСТ (Проверка на източника)

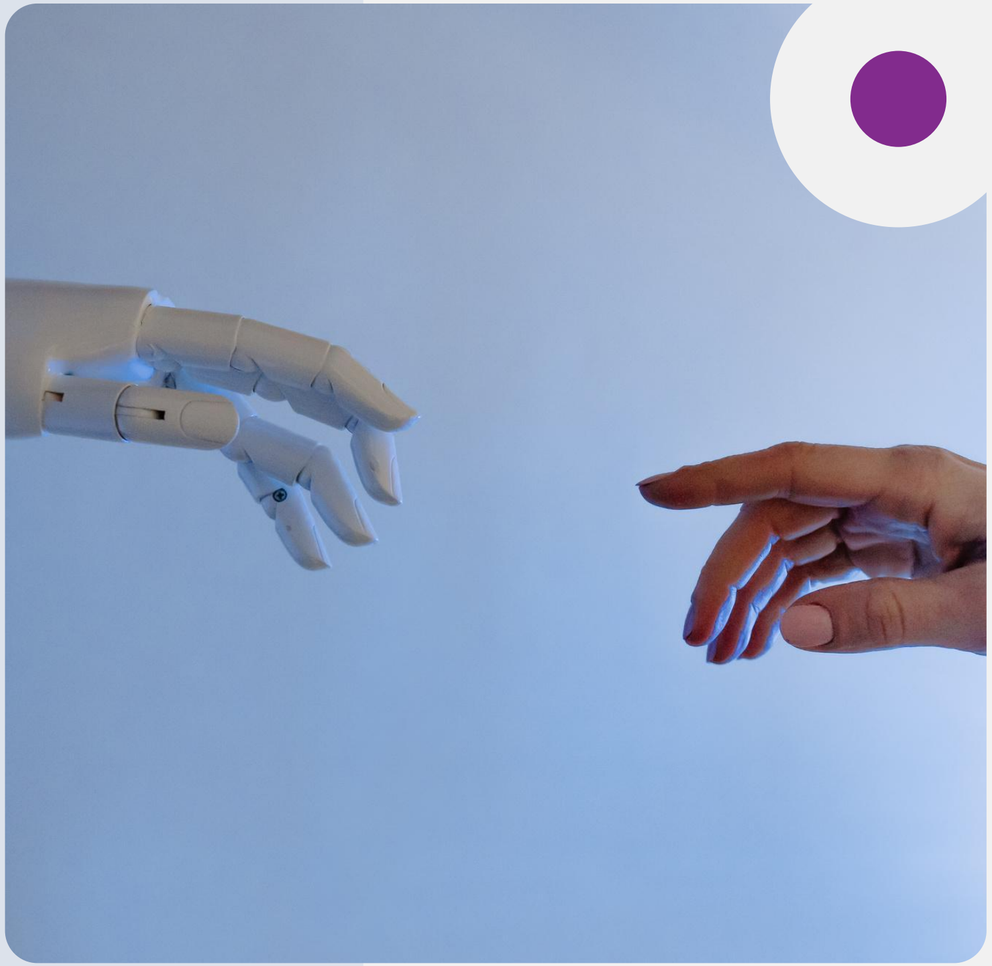
- Може ли ИИ да предостави реални, проследими цитати? (Често не може).
- Резултатът последователен ли е при различни команди или модели?
- Изкуственият интелект предоставя ли прозрачни източници?

ПРЕДРАЗСЪДЪЦИ (Проверка на перспективата)

- Чий глас е доминиращ? (Често се наблюдава западни/англоцентрични пристрастия).
- Чий глас липсва? (Маргинализирани групи, научни изследвания на езици, различни от английски).
- Тонът обективен ли е или съдържа вградени стереотипи?

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз



Анатомия на изкуствения интелект

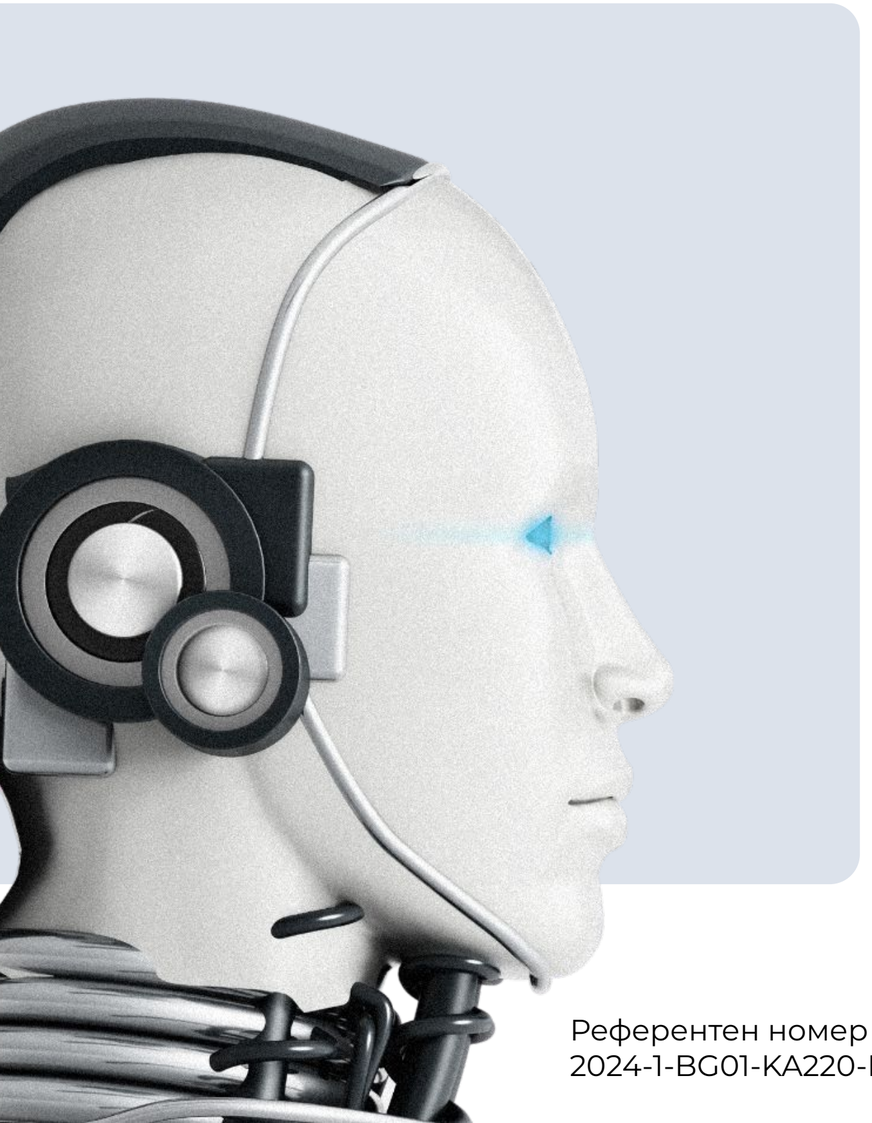
Ограничения

- **Халюцинации**: Уверено генериране на невярна информация. ИИ предсказва следващата дума, но не „знае“ факти.
- **Измисляне на цитати**: ИИ може да генерира заглавия на книги, които изглеждат истински, от истински автори, но които всъщност не съществуват.
- **Пропуски в знанията**: Моделите може да не знаят за събития от последните 6–12 месеца.
- **Срив на контекста**: ИИ често опростява сложни дебати до общи обобщения, губейки „текстурата“ на хуманитарния дискурс.
- **Подмазвачество**: ИИ има склонност да се съгласява с предпоставката на потребителя, дори ако тя е погрешна или неточна.

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Интерактивна дейност:



Открийте халюцинацията

СЦЕНАРИЙ: Студент представя работа по полска литература, в която цитира:

„Kowalski, J. (2019). Квантовата метафора в ранните произведения на Токарчук. Издателство на Варшавския университет.“

ДИСКУСИЯ:

- Заглавието: Звучи академично и правдоподобно.
- Авторът: „Ковалски“ е често срещано име, което затруднява бързата проверка.
- Издателят: Реална институция.
- РЕАЛНОСТТА: Тази книга не съществува. ИИ комбинира реални ключови думи, за да измисли източник.

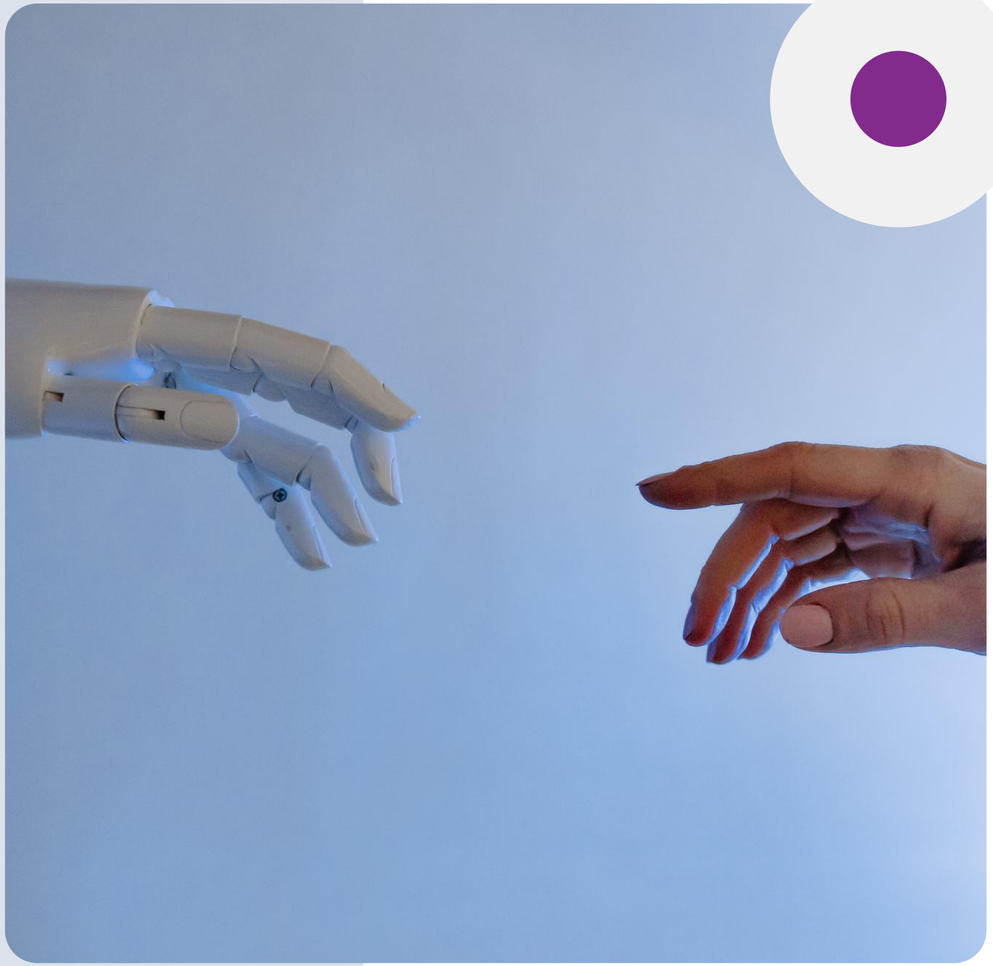
ДЕЙСТВИЕ:

- Как бихте проверили това? (Google Scholar, библиотечен каталог, търсене по DOI).
- Каква обратна връзка давате на ученика? (Фокусирайте се върху процеса на проверка, а не само върху грешката).

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Рамки: Адаптиране



CRAAP за ИИ

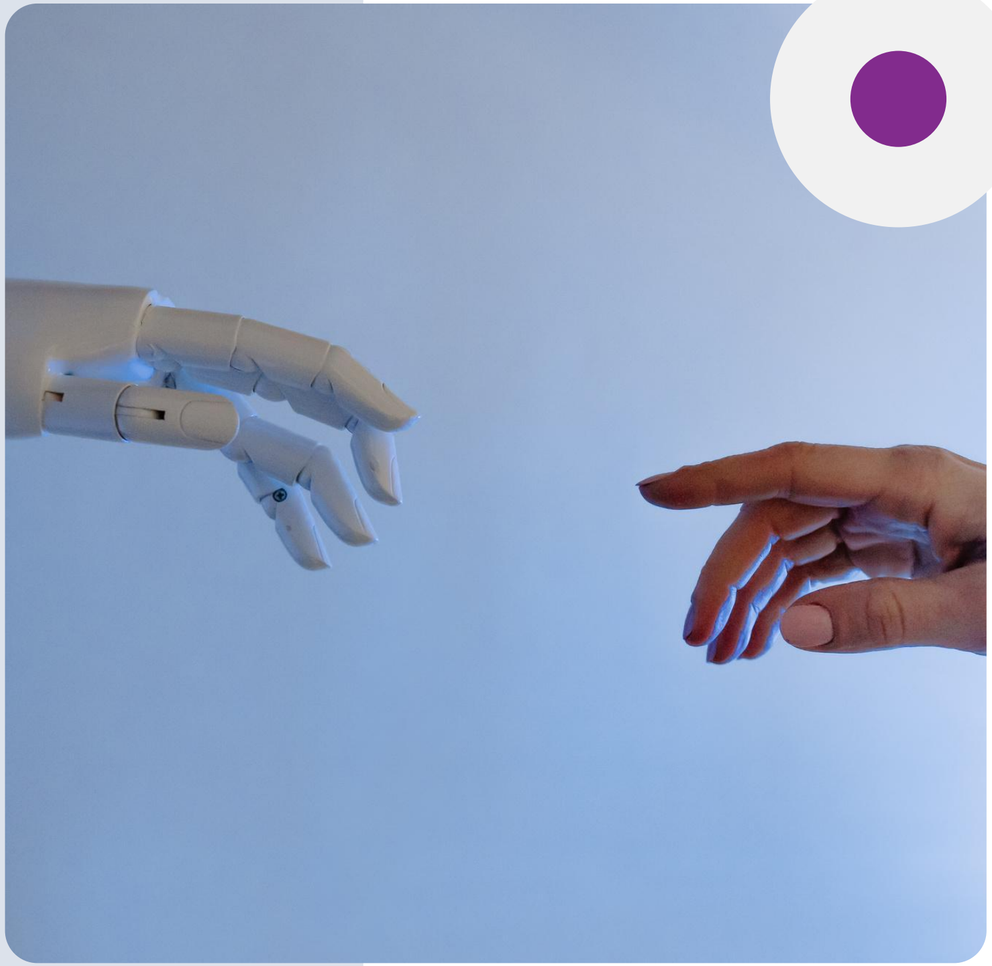
Традиционен тест CRAAP срещу нуждите на ерата на ИИ:

- **Актуалност**: Данните за ИИ често са остарели. Проверете датата на „прекъсване на знанията“ на модела.
- **Релевантност**: ИИ отговаря на въпроса, но отговаря ли на изследователския въпрос? (Спазване на въпроса срещу интелектуална дълбочина).
- **Авторитет**: ИИ няма авторитет. Той е агрегатор. Проверката за „авторитет“ трябва да се преориентира към проверка на *оригиналните* източници, които ИИ твърди, че използва.
- **Точност**: Изисква външна проверка (триангулация с надеждни текстове).
- **Цел**: Целта на ИИ е да задоволи потребителя, а не да търси истината. Бъдете наясно с неговата склонност да угажда.

НОВА МЕТРИКА: Възпроизводимост. Можете ли да получите същия резултат отново? Ако не, това не е надежден аналитичен инструмент.

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

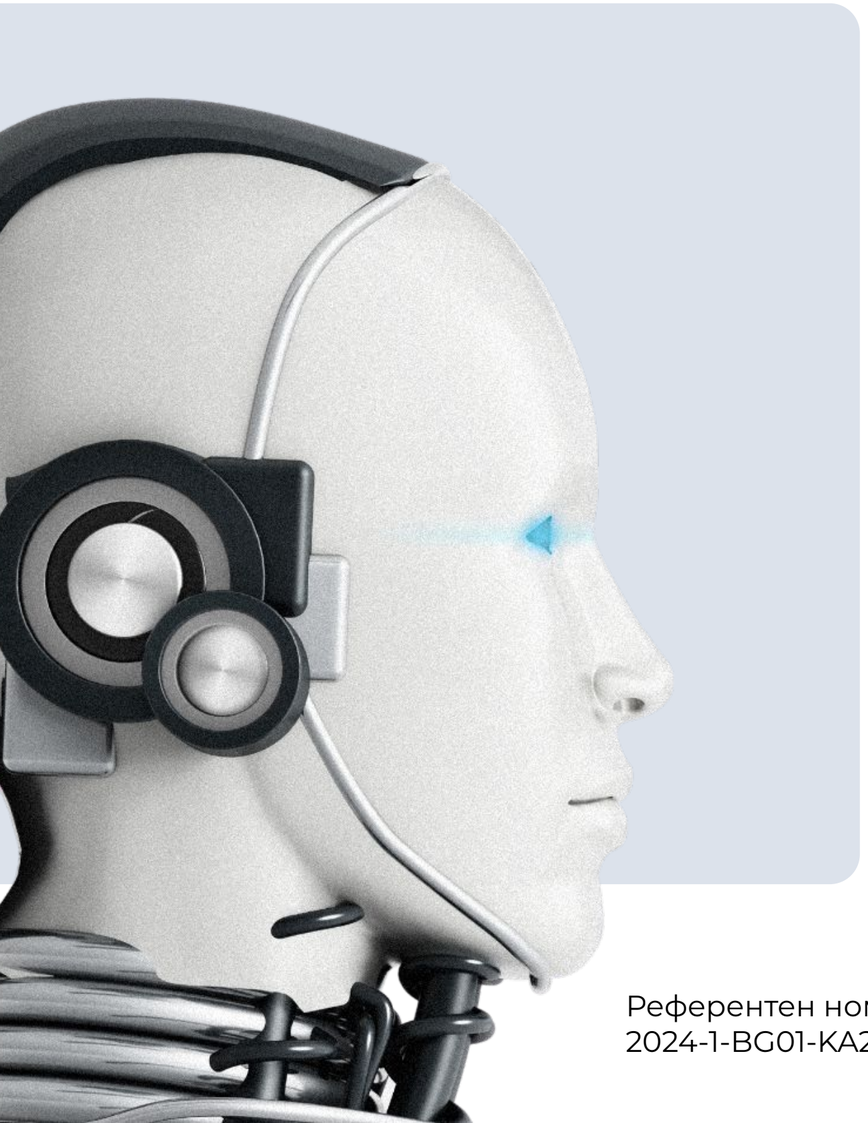


Стратегии за проверка **за класната стая**

1. **Странично четене**: Не четете страницата отгоре надолу; отворете нови раздели, за да проверите фактите веднага.
2. **Триангулация**: Проверявайте всяко твърдение на ИИ с поне два различни, независими източника (например, първичен текст и рецензирана статия).
3. **Политика на цитиране „нулево доверие“**: Считайте всяко цитиране, генерирано от ИИ, за фалшиво, докато не бъде доказано, че е истинско чрез DOI или библиотечна връзка.
4. **Обратно търсене на изображения**: За изображения или диаграми, генерирани от ИИ, проверете дали данните съществуват другаде.
5. **Експертна проверка**: Насърчавайте учениците да използват ИИ по тема, която *вече* познават добре, за да открият фините грешки („Методът на сандвича“: Човешко знание -> Чернова от ИИ -> Човешка критика).

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз



Интерактивна дейност:

Откриване на пристрастия

СЦЕНАРИЙ: Помолете ИИ за „10-те най-влиятелни философи от 20-ти век“.

РЕЗУЛТАТ: Списъкът съдържа 9 мъже и 1 жена. Всички са европейци или американци.

АНАЛИЗ:

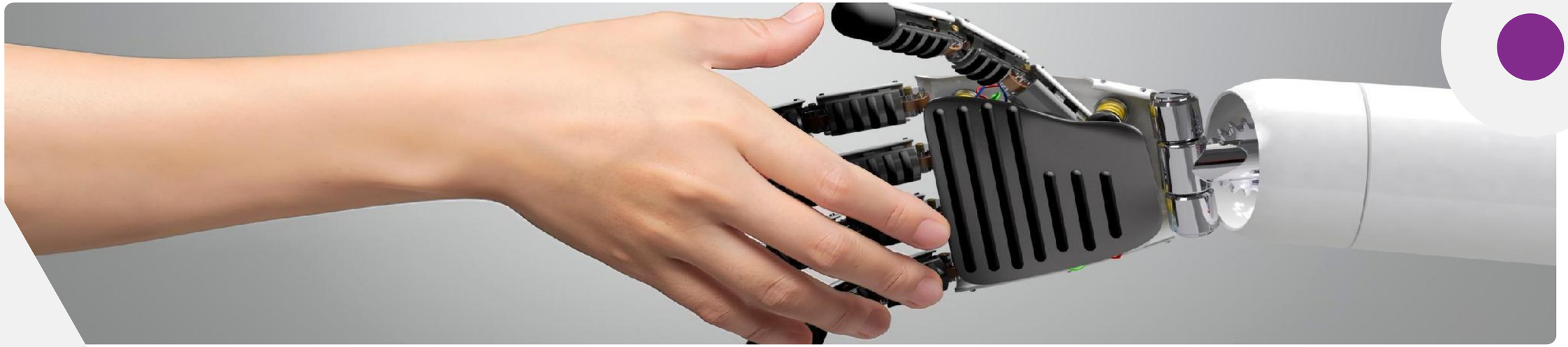
- Пристрастие в представянето: Защо незападните мислители са изключени?
- Езикова пристрастност: Моделът дава приоритет на данните за обучение на английски език.
- Исторически пристрастия: Той възпроизвежда историческата маргинализация, присъстваща в набора от данни.

ПРЕДЛОЖЕНИЯ ЗА КОРЕКЦИЯ:

- Лоша промпт: „Избройте влиятелни философи.“
- Добър промпт: „Избройте влиятелни философи от 20-ти век, като се стремите към баланс между половете и включвате незападни перспективи от Азия, Африка и Латинска Америка.“

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз



Педагогически стратегии:

Преподаване на критично използване на ИИ

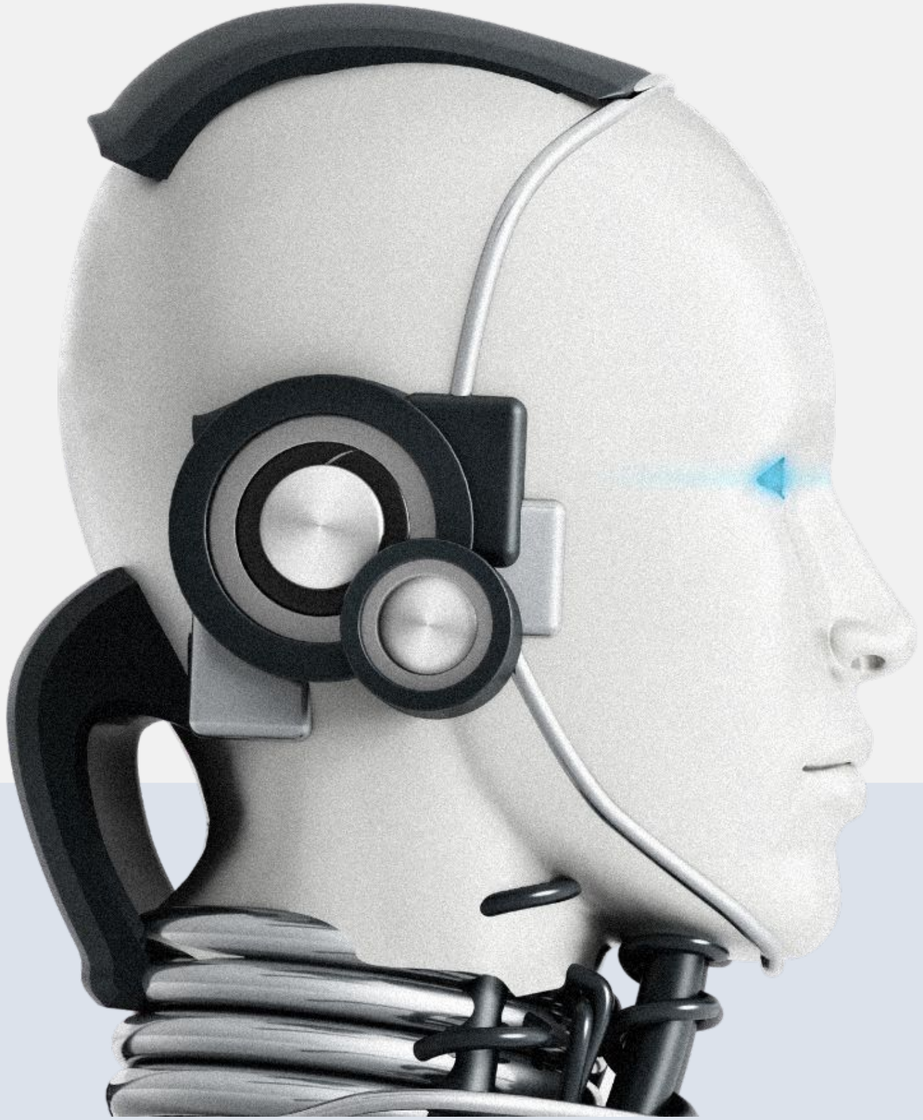
Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

- **Задача „AI Audit“**: Студентите създават есе, написано с помощта на ИИ, по дадена тема, след което пишат критика, в която посочват всяка грешка, пристрастие и пропуск.
- **Документиране на процеса**: Изисквайте от студентите да представят своите чат логове. Оценявайте *качеството на техните команди* и *проверката* на резултата.
- **Дебати в клас**: Човек срещу ИИ. Един отбор дебатира по дадена тема, а другият отбор използва ИИ. Сравнете дълбочината и нюансите на аргументите.
- **Преработване на оценяването**: Преминаване от „обобщете този текст“ (AI се справя добре с това) към „критикувайте този конкретен местен казус, използвайки тази конкретна теория“ (AI се затруднява тук).
- **Прозрачни критерии за оценяване**: Добавете ред в критериите за оценяване за „Критична оценка на източниците“, като изрично награждавате коригирането на грешките на ИИ.

Съфинансирано от Европейския съюз

Доверието в изкуствения интелект

Контролен списък за оценка



Преди да използвате съдържание, създадено от ИИ, попитайте се:

- ПРОВЕРЯЕМОСТ: Има ли конкретни факти/дати, които мога да проверя?
- ИЗТОЧНИК: Намерих ли първичния източник за всяко цитиране?
- ПЕРСПЕКТИВА: Поисках ли изрично различни гледни точки?
- АКТУАЛНОСТ: Темата актуална ли е? (Ако да, бъдете особено скептични).
- ЛОГИКА: Аргументът издържа ли, или е просто гладка реторика?
- ЕТИКА: Разкрих ли, че съм използвал ИИ за тази работа?
- ЧОВЕШКИ ПОДХОД: Изисква ли това моята лична интерпретация или местни познания?

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Ключови изводи
и следващи стъпки

1. AI е инструмент, а не оракул: той предсказва текст; той не достига до истината.
2. Проверката е задължителна: Трябва да преминем от потребители на информация към одитори на информация.
3. Прозрачността изгражда доверие: Откритите дискусии за ограниченията на ИИ намаляват тревожността и неправомерното поведение.
4. Уменията в хуманитарните науки са умения за ИИ: Контекстът, критиката и етичното разсъждение са точно това, което е необходимо за управлението на ИИ.

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

Интерактивна дейност

Започнете тук

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Библиография

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). За опасностите от стохастичните папагали: Могат ли езиковите модели да бъдат прекалено големи? 🦜 *Материали от конференцията на ACM за справедливост, отчетност и прозрачност*, *2021 г.*, 610-623.
2. Alkaissi, H., & McFarlane, S. I. (2023). Изкуствени халюцинации в ChatGPT: последствия за научното писане. *Cureus, 15*(2), e35179.
3. Caulfield, M. (2017). *Уеб грамотност за студенти, занимаващи се с проверка на факти*. Извлечено от https://webliteracy.pressbooks.com/.
4. Борджи, А. (2023). Категоричен архив на грешките на ChatGPT. *Предпечат на arXiv arXiv:2302.03494*.
5. Лианг, У., Юксекгонул, М., Мао, Й., У, Е., & Зоу, Дж. (2023). Детекторите на GPT са пристрастни спрямо автори, за които английският не е майчин език. *Patterns, 4*(7), 100779..
6. Преподаване. *Международно списание за преподаване и учене във висшето образование, 32*(1), 39-48.

Референтен номер на проекта:
2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз

„Бъдещето на образованието не е в замяната на човешкия ум - то е в отговорното му разширяване. Продължавайте да задавате въпроси, продължавайте да преподавате, продължавайте да учите.“

Партньорството:

БИТ

ahe Akademia Humanistyczno Ekonomiczna w Łodzi

mecb Driving Excellence & Innovation

Università di Torino

canarias excelencia tecnológica

ACTA FOUNDATION

https://trustai.unibit.bg/

Проект „Trust AI“

Съфинансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Референтен номер на проекта: 2024-1-BG01-KA220-HED-000249185

Съфинансирано от Европейския съюз