Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Катедра „Компютърна информатика”

ПРЕДДИПЛОМЕН ПРОЕКТ

на тема

„Създаване на многомодален и многоезичен набор от данни за автоматично отговаряне на въпроси, давани на ученици от различни нива на обучение”

Дипломант: **Симеон Емилов Христов**

Факултетен номер: **6MI3400191**

Специалност: **Информатика**

Магистърска програма: **Изкуствен интелект**

Научни ръководители:

Проф. Преслав Наков, катедра „Обработка на естествен език“, Университет „Мохамед бин Зайед по Изкуствен интелект“ (MBZUAI)

Проф. д-р Иван Койчев, катедра „Софтуерни технологии“, ФМИ, СУ „Св. Климент Охридски“

Консултант:

докт. Димитър Димитров, катедра „Софтуерни технологии“, ФМИ, СУ „Св. Климент Охридски“

София, 2024 г.

Съдържание

1. Увод

Обобщен изкуствен интелект е възможността на машина да разбира широка гама съдържание (включ. текстово, визуално, аудио) и да взема решения и предприема действия, използвайки го. Концептуално тази възможност е връзката между текущите имплементации, които са силно специфицирани към решаване на конкретна и ясна задача, и системите с обобщен изкуствен интелект, често срещани в научно-популярните филми [1].

Желанието за цялостно и непредубедено измерване на степента на развитие на големи многомодални модели и големи езикови модели е довело до създаването на различни и разнообразни набори от данни. Заедно с това са се развили вече съществуващи такива, често чрез добавянето на нова модалност [6, 17, 18]. Наблюдава се засилено разработване и усъвършенстване на набори от данни в три направления - оценяване на възможността на големи езикови модели да разбират текстово съдържание, оценяване на възможността на големи визуално-текстови модели да създават и модифицират визуално съдържание, следвайки предварително зададено текстово описание, и оценяване на възможността на големи многомодални модели да разбират и създават съдържание, включващо, както визуална, така и текстова част.

Създаването на подобни множества от данни не е тривиална задача [9]. За да се оцени правилно основен модел е необходимо да се използват данни с все по-голяма сложност, която обхваща както всекидневен контекст, така и високо специализирани експертни тематики. Освен по трудност и насоченост, данните трябва да варират и по тип и характеристики. Подсигуряване, че данните са достатъчно като количество и са правилно анотирани, е също важна стъпка в създаването на подобни множества.

Големите езикови модели набират популярност в различни аспекти на човешката дейност. Отличителна тяхна черта е даването на кратък, точен и релевантен отговор на широка гама въпроси. Големите многомодални модели, разбиращи текст и картинки, могат да се използват в образователната система за получаване и/или валидиране на отговори на въпроси от затворен тип. Тези типове въпроси се характеризират с конкретика (те са кратки и ясни) и недвусмисленост (имат краен брой възможни отговори). Те също така могат да имат различна модалност, т.к. освен текстово могат да включват и визуално съдържание.

Настоящият преддипломен проект има за цел да опише създаването на многомодален и многоезичен набор от данни, включващ три типа въпроси, давани на ученици от различни нива на обучение. Въпросите могат да включват изцяло текст, а могат да включват и текст с картинки. Възможно е картинката да бъде част от въпроса, но също е възможно и самите отговори на включват (понякога изцяло) картинки. Разглеждат се само въпроси със затворен отговор, в които има само един верен отговор. Резултатите от този преддипломен проект ще послужат за основа при разработването на дипломна работа, която има за цел да направи сравнителен анализ на свободно достъпни големи многомодални модели и големи езикови модели.

В настоящия преддипломен проект са поставени и изпълнени следните задачи:

- Подробен обзор на съществуващи подходи за създаване на многомодални и многоезикови набори от данни.

- Описание на процеса по създаване на набора от данни Exams2:
    - Създаване на подходяща файлова структура.
    - Предварителна обработка на данните.
    - Анотиране на данните.
    - Валидиране на създадените файлове с анотации.
    - Последваща обработка на данните.
- Използване на чатбота Google Bard за получаване на отправна точка за бъдещи сравнения.

2. Преглед на областта

2.1. Основни дефиниции

Езиков модел (от англ. language model) - компютърна система, съпоставяща вероятност на последователност от думи.

Голям езиков модел (от англ. large language model) - езиков модел, предназначен за използване в големи мащаби и характеризиращ се с възможност за разбиране и генериране на език с общо предназначение.

Голям визуално-текстови модел (от англ. large vision-language model) - голям езиков модел, който може да получава на вход текстово и визуално съдържание.

Модалност на данни (от англ. data modality) - независим канал за протичане на информация при входно/изходни операции с човешко-машинен интерфейс. Най-популярните към момента модалности са текст, видео (в частност картинки) и аудио.

Чатбот (от англ. chatbot) - уеб интерфейс, използващ основен модел, за да имитира човешки разговор чрез обменяне на текстови или аудио съобщения.

Основен модел (от англ. foundation model) - модел, базиран на невронни мрежи, който има много параметри, стойностите на които са се получили в резултат на трениране с голяма част от данни с различни модалности, достъпни в Интернет. Характеризират се с възможност да симулират разсъждения в произволен контекст.

Голям многомодален модел (от англ. large многомодален model) - основен модел, който може да приема като вход данни от различни модалности и да създава резултати от същите или други модалности.

Образователни системи тип K-12 - обхваща годините на получаване на формално или задължително образование от детска градина до завършване на средно образование.

Затворен въпрос (от англ. close ended question) - въпрос, който има краен брой отговори, най-често от два до пет.

Отворен въпрос (от англ. open ended question) - въпрос, който провокира излагане на твърдение и поняка - негова защита или отрицание.

Анотация - множество от етикети, описващи прилежаща метаинформация, към единици данни.

2.2. Обзор на съществуващи набори от данни

2.2.1. Набор от данни Visual QA

Задачата за даване на отговор в свободен стил след получена картинка като вход се въвежда за първи път в [6]. Тя цели да провери доколко големите визуално-текстови модели могат да върнат правилен отговор в текстов вид по подадена картинка и въпрос за нея. Въпросът е отворен и в частност е насочен към конкретни части на картинката. Визуалното съдържание е подбрано от наборът данни MS COCO [14], който предоставя снимки от ежедневни ситуации в реалния свят, които са автентични и без допълнителна обработка. Обхващат широка гама сцени и имат разнообразен контекст от обекти. Поради тази строга специфичност обаче имат много детайли и шум. Авторите констатират, че това внася вид предубеденост и затова добавят втора част към формираното множество, която се състои от абстрактни сцени и обекти. Сцените и обектите могат да се групират по произволен начин и така да се създават нови сцени, които, въпреки че са нереалистични имат изчистен заден план и могат да се използват за създаване на картинки с цел фино оценяване на моделите.

Наборът от данни VQA съдържа въпроси само на английски език. В резултат на това, че се отнасят за конкретни част от изображението, те са лимитирани спрямо контекста и започват обикновено с някоя от четирите въпросителни думи - кой, какво, къде и колко.

Данните във VQA не са фокусирани върху изпити, давани на ученици, но служат за начална точка за създаването на именно такива множества. Така се дава и началото на множество изследвания в областта, явяваща се сечение между компютърното зрение и обработката на естествен език. В нея се появяват по-късно многомодалните модели.

2.2.2. Набор от данни ScienceQA

Създаването на първия набор от данни фокусиран върху многомодални въпроси, взети от изпити, давани на ученици, е описано в [5]. ScienceQA е считан за голяма и важна стъпка към цялостното и систематичното оценяване на големи многомодални модели. Изследва се възможността на модели да създават верига от логически свързани мисли/твърдения, формирането на които позволява достигането до крайния отговор. Обхванати са три предмети, изучавани главно в класовете от първи до шести - Биология, Социология и Науката за езика. Отличителни черти на ScienceQA са големият брой многомодални въпроси и прилежащите към тях специални анотации - лекции и обяснения. Това позволява да се оценят не само крайните отговори, които се получават от модела, но и да се проследят и проверят предоставените обяснения. По този начин може да се валидира доколко моделът разбира въпроса и имитира човешко разсъждение и доколко халюцинира [19].

2.2.3. Набор от данни Textbook QA

Наборът от данни TQA [10] цели да оцени възможността на модел да даде отговор на краен брой многомодални въпроси, използвайки като контекст урок с прилежащи диаграми и картинки.

TQA е създаден, използвайки предмети от шести клас. Съдържа 1,076 уроци от Биология, География и Физика. Уроците са взети от учебници, които се използват в САЩ, но имат и съответни преведени варианти, които се изучават в други държави. Всеки урок съдържа текстово съдържание под формата на параграфи и визуално съдържание под формата на диаграми и реални картинки. Всеки урок съдържа и терминологичен речник, предоставящ основни дефиниции, както и резюме на урока. Резюмето е ограничено до пет изречения и обхваща представените ключови концепции.

Уроците съдържат и уеб линкове към онлайн видеа с инструкции, които обясняват концепциите с повече визуализации. Авторите констатират, че макар съществуващото текстово съдържание да е цялостно и да е самодостатъчно, за разбирането на концепциите в уроците, то не е достатъчно за разбирането на картинките, и предполагат, че учителите ги обясняват по време на час. При оценяване на моделите това води до по-ниска успеваемост на въпросите свързани с картинки, поради по-голямата им трудност. С цел симулиране на обяснението на учителите и съдържанието на онлайн видеата, авторите добавят създадени от тях от три до пет *Инструкционни диаграми* към всеки урок, които имат подробни описания, обясняващи въведените концепции.

След всеки урок има от два до седем въпроса, пряко свързани с представената информация. Броят на въпросите с текст и картинки е значително по-малък от броят на въпросите само с текст. От части това се дължи на по-трудоемката работа по създаването им. За увеличаване на количеството им, представените концепции във всеки урок се използват като ключове за търсене в платформата Google Image Search и първите изведени резултати, които представят съдържание, близко до това, представено в урока, са добавени като нови въпроси.

Наборът от данни е разбит на тренировъчно, валидационно и тестово множество на ниво урок. Има случаи, в които няколко урока имат непразно сечение в концепциите, които те представят. Такива случаи са поставени заедно в съответната разбивка.

2.2.4. Набор от данни SciBench

Наборът от данни SciBench се фокусира върху изследването на представянето на модели върху въпроси от предмети, свързани с Математика [7]. Целта е систематично да се изследват нужните уменията за разсъждаване над и решаване на комплексни математически задачи. Състои се от две множества данни, съдържащи въпроси от трудност в университет: с отворен характер и със затворен характер. Отворените въпроси са 695 и са събрани от популярни университетски курсове, изучаващи Физика, Термодинамика, Класическа механика, Квантова химия, Диференциално и интегрално смятане и Статистика.

За по-голяма близост до реален университетски курс, се включва и втора част с въпроси, която съдържа общо 7 семестриални изпита от три университетски курса в сферата на компютърните науки и математиката. Тези въпроси имат един верен отговор, но са представени като отворени въпроси. По този начин се намалява възможността на модел да отгатва правилния отговор.

Всички въпроси имат един верен отговор и всички отговори са преобразувани към числа с плаваща запетая, закръглени до третата цифра след запетаята. Това позволява недвусмислие при оценяването.

SciBench се отличава по това, че всички въпроси са с отворен отговор и предполагат изразяване на решението чрез логически свързани стъпки от комплексни аритметически операции като диференциране и интегриране. Стъпките на решенията са ръчно преписани от съдържащите ги PDF документи в LaTeX файлове. Това минимизира възможността от тяхното изтичане в тренировъчното множество, използвано при обучаването на многомодалните модели и така се запазва целостта на оценяването.

Характеристиките на това множество са:

- Избраните въпроси изискват разбиране на сложни математически операции и опит в прилагането им. Нужните умения не са ограничени само до прости аритметични операции като събиране и умножение - те засягат диференциране и интегриране и работа с много малки и много големи числа.
- Детайлни решения са включени към избраните проблеми.
- С цел осигуряване на непредубедена оценка, въпросите са подбрани така, че да не могат да се намерят лесно, ако са потърсени онлайн и не могат лесно да се извадят и трансформират към текст. Този процес предотвратява изтичането на въпросите във вече съществуващи тренировъчни множества, които често са използвани при обучаване - например стандартни тестове като SAT.
- За да се предотврати възможността за познаване на верния отговор от крайно множество, всички въпроси са с отворен отговор.

2.2.5. Набор от данни MathVista

Друго подобно множество от данни, което се характеризира с фокус върху Математиката и разсъждаването върху фигури е MathVista [4]. За разлика от останалите разгледани набори от данни, MathVista съдържа 6,141 въпроса, събрани от 28 вече съществуващи многомодални множества и 3 новосъздадени такива (IQTest, FunctionQA и PaperQA). MathVista съдържа въпроси от 7 математически контексти: Алгебра, Аритметика, Геометрия, Логическо мислене, Прости числови операции, Научно мислене и Статистическо мислене. Фокусира се върху 5 типа задачи: отговаряне на въпрос за фигура, решаване на геометрична задача, решаване на математическа задача, отговаряне на въпрос, съдържащ текст, тип пъзел, отговаряне на въпрос, съдържащ текст и картинка. Обхваща широка гама визуално съдържание - естествени картини, геометрични фигури, графики, изкуствени сцени и диаграми. Т.к. това е набор от данни, голяма част от който е формиран от вече съществуващи множества, въпросите варират по трудност, обхват и концепции, към които са обвързани.

2.2.6. Набор от данни MMMU

Авторите на [9] представят MMMU като набор от данни, създаден, за да оценява многомодални модели на масивно мултидисциплинарни задачи,

решаването на които изисква умения, придобити в университетски курсове, и съзнателни разсъждения. Цели да измери успеваемостта на моделите в три основни направления: възприятие, знания и разсъждения.

Обхванати са 30 предмета, включително Изкуство, Бизнес, Здраве и Медицина, Социология и Компютърни науки. Въпросите са събрани в три фази. Първата фаза се изразява във филтриране на курсовете, които ще бъдат обхванати. Критерият за подбор е количеството полезна информация, която се носи от картиките и диаграмите. По този начин предмети като Право и Лингвистика са изключени. Във втората фаза за избраните предмети студенти, изучаващи ги, заемат ролята на анотатори и извличат многомодални въпроси от ресурси достъпни онлайн, както и от учебници, а, където се налага, добавят и нови въпроси, спрямо собствената си експертиза. Данните се събират така, че отговорите да не са непосредствено около въпроси, а да се намират или в отделен документ, или в края на текущия. По този начин се намалява вероятността от замърсяване и теч на данните. Събрани са 13,000 въпроса, следвайки този процес, които след това са подложени на проверка на качеството. Тази проверка се изразява в идентифициране и премахване на дубликати и ръчна проверка на случайна извадка от работата на различни анотатори. В последната стъпка на тази фаза събраните въпроси са групирани спрямо трудност в четири категории: много лесни, лесни, средна трудност и трудни. Въпросите, попаднали в категорията “много лесни” или неотговарящи на целите на набора от данни, са премахнати. Количеството им е около 10%.

MMMU се отличава от останалите набори данни по това, че:

- Обхваща много тематики, които са комплексни. Не са включени въпроси, които са свързани с ежедневен контекст.
- Присъства визуално съдържание от 30 различни типа, включително диаграми, таблици, графики, химични структури, снимки, произведения на изкуството, картини, геометрични фигури и др.
- Обхванатите тематики са покрити в голяма дълбочина - изискват се умения, придобити в разнородни университетски курсове.

2.2.7. Набор от данни M3Exam

Наборът от данни M3Exam се характеризира с многомодалност, многоезичност и въпроси, които са взети от всички нива на обучение до достигане на средно образование [8]. Обхванте са езиците Английски, Китайски, Италиански, Потругалски, Виетнамски, Тайски, Суахили, Африкаанс и Индонезийски език. Изборът е продиктуван главно от желанието да се създаде набор от данни, който има не само лингвистични, но и културни разлики. Така се покриват различни езикови групи и езици с неравномерна популярност, за които няма много информация в Интернет.

След избор на езиците, местни жители, говорещи ги, събират официални изпитни материали, давани на ученици от четвърти, седми, и дванадесети клас. Насоките, които са следват са 1) винаги да се избират изпитите с повече явяващи се ученици, т.е. при наличие на държавни и областни изпити в дадено ниво на обучение, да се избират държавните изпити; и 2) да се събират въпроси от

всички възможни предмети и до 5 изпита от предишни години за всеки предмет, за да се гарантира разнообразие на въпросите.

Като резултат са събрани 435 изпита от 9 държави. Понеже много от събраните изпити са изцяло във формат на картинки и сканирани копия, авторите използват разпознаване на символи (optical character recognition - OCR), за да извлекат съществуващия текст. Получените изпити, извлеченият, редактируем текст и отговорите за всеки от тях са след това подадени към анотатори, които създават затворени въпроси в предварително определена структура. Отворените въпроси биват изключени.

При създаването на въпросите се коригират грешки, допуснати при OCR, и се отделя внимание да се постави максимално много контекстуална информация, нужна за правилното разбиране на въпросите. Всички уравнения са записани посредством LaTeX синтаксис. Като последна стъпка са извършени няколко последователни проверки на качеството на анотациите. В резултат 23% от въпросите имат картинки.

2.2.8. Набор от данни Exams

Наборът от данни е създаден от Хардалов и колеги [2]. Фокусира се върху оценяване на големи езикови модели, в частност, сравнява каква точност постигат, когато отговарят на еднакви въпроси, преведени на различни езици.

Събрани са PDF файлове за всеки изпит, които след това са били преобразувани в текстови файлове, съдържащи въпросите. Създаденият код и събраните данни са свободно достъпни в [3]. Данните представляват въпроси от редовни и поправителни изпити, давани на ученици от различни нива на обучение, главно ученици, завършващи четвърти и дванадесети клас. Изпитите са изготвени от министерствата на образованието на различни държави. Въпросите обхващат широка гама предмети от основни като Биология, Физика, Химия, до високо специализирани като Геология, Иконометрия, Градинарство, както и профилирани предмети. Тези характеристики допринасят за постигането на добро разнообразие на въпросите от гледна точка на приложната област и на комплексността им, т.к. изискват високо-специализирани знания.

В образователните системи на някои от обхванатите държави се провеждат изпити с едно и също съдържание, но преведено на няколко езика. Това води до появата на “паралелни” въпроси. Подобни се наблюдават в изпитите от Хърватия (които се срещат преведени на Сръбски, Италиански и Унгарски), Унгария (които се срещат преведени на Немски, Френски, Испански, Хърватски, Сръбски и Италиански) и Северна Македония (които се срещат преведени на Албански и Турски). Това е отбелязано от авторите и служи като експеримент, който не е извършван в останалите разгледани множества.

Наборът от данни съдържа 24,143 въпроса в 16 езика от 8 езикови групи. Всеки въпрос съдържа от 3 до 5 възможни отговора, от които винаги само един е верен. Анализ на данните е представен на таблица 1.

Език	Езикова група	Брой предмети (агрегирани)	Брой въпроси
Албански	Албански	8	1,505
Арабски	Семитски	5	562

Български	Балто-славянски	6	2,937
Хърватски	Балто-славянски	14	2,879
Френски	Романски	3	318
Немски	Германски	5	577
Унгарски	Угро-фински	10	2,267
Италиански	Романски	12	1,256
Литовски	Балто-славянски	2	593
Македонски	Балто-славянски	8	2,075
Полски	Балто-славянски	1	1,971
Португалски	Романски	4	924
Сръбски	Балто-славянски	14	1,637
Испански	Романски	2	235
Турски	Тюркски	8	1,964
Виетнамски	Австроазиатски	6	2,443

Таблица 1: Статистики за Exams

Настоящият преддипломен проект стъпва върху работата, извършена от Хардалов и колеги. Той използва наготово голяма част от наличните изпити и ги разширява, добавяйки нова модалност. Въпросите се разбиват на въпроси, включващи само текст и въпроси, включващи текст и картинки.

2.3. Сравнителен анализ и изводи

Сравнителна таблица между разгледаните набори от данни и новоизградения Exams2 е представена в таблица 2.

Име	Публикуван (Година/Месец)	Брой въпроси	Брой предмети	Многомодалност	Многоезичност	Обучителен курс	Допълнителен контекст
VQA	2016/10	1,105,904	-	да	не	-	не
ScienceQA	2022/10	21,208	3	да	не	1-12 клас	да
TQA	2017/07	26,260	3	да	не	6 клас	да
SciBench	2023/07	799	10	не	не	Бакалавър	не
MathVista	2023/10	6,141	5	да	не	Бакалавър	не
MMMU	2023/12	11,500	30	да	не	Бакалавър	не
M3Exam	2023/11	12,317	4	да	да	1-12 клас	да
Exams	2020/11	24,143	24	не	да	1-12 клас	не
Exams2	**2024/01**	**17,420**	**19**	**да**	**да**	**1-12 клас**	**не**

Таблица 2: Сравнителен анализ

Създаването на многомодален и многоезичен набор от данни не е тривиална задача. С бързото темпо на развитие на многомодалните модели, правилното и цялостното им оценяване изисква наличие на следните характеристики:

1. **Широк обхват**: Необходимо е да се включат въпроси, свързани както с дисциплини, изискващи математически умения, така и с дисциплини, фокусирани към разбиране на социални взаимодействия, психология и морал. Това гарантира цялостна проверка на развитието и възможността за обобщаване на разсъжденията на големи езикови и големи многомодални модели. В контекста на образователната система, това се изразява чрез включване на въпроси от максимално много предмети, изучавани от ученици и студенти.
2. **Дълбочина**: Докато набори от данни като ScienceQA имат широк обхват и съдържат допълнителна информация към всеки въпрос, в набора от данни липсват високо-специализирани въпроси, отговарянето на които изисква задълбочени знания и наличие практически опит, трупан с години. Това е важно качество, което би гарантирало възможност за създанелно преминаване през множество мисловни стъпки и формиране на сложни аргументи. В контекста на образователната система, това се изразява чрез включване на въпроси от всички нива на обучение на ученици и студенти.
3. **Многоезичност**: Важно е да се провери доколко големите многомодални модели могат да бъдат прилагани към езици, различни от Английски, и езици, за които наличната информация в Интернет е малка.
4. **Многомодалност**: Проблемите, решаването на които изисква работа с няколко модалности (напр. визуална и аудио), са често срещани в реалния свят. Възможността на големи многомодални модели да работят правилно с информация от различни модалности е от ключово значение.

Настоящият преддипломен проект има за цел да развитие работата, създадена от Хардалов и колеги като добави нова модалност на данните в Exams. По този начин Exams2 притежава посочените качества на добър набор от данни за оценяване на многомодални модели.

3. Наборът от данни Exams2

3.1. Същност и цели

Наборът от данни Exams2 е вариант на Exams, в който извлечените въпроси от всеки изпит са многомодални. Те представляват изрязана снимка на въпрос. Моделът трябва да интерпретира текста, който е част на снимката, и да разбере евентуално прилежащи към нея картинки и/или таблици. Целта на създаването Exams2 е да се използва като точка на сравнения за големи многомодални модели. След изграждането на подходящ процес по предварителна обработка на данните (в частност, преминаване от визуално описание на въпроса към текстово такова), Exams2 може да се използва и за оценяване на големи езикови модели.

3.2. Избор на подмножество от данните на Exams

Целта на Exams2 е да представи изрязана снимка на въпрос, който включва от 2 (истина/лъжа) до 5 възможни отговора. Очаква се моделът да може да разпознае идентификатора на реда, който маркира верния отговор. За целта този идентификатор трябва да е или буква от съответна азбука, или цифра. Това налага предварителен избор на подмножество от изпитите, събрани в Exams. Критериите за избор на изпити, които са разширени с многомодални въпроси, са:

1. Изпитът да съдържа отговори, които са букви или цифри. За Арабски език, например, възможните отговори са маркирани с квадратчета, което го прави труден за използване без предварителна обработка.
2. Директорията, в която е поставен изпитът, да съдържа PDF файл. За някои езици налична бе само текстови вариант (т.е. т.е. текстови файл, получен след разпознаване на всички символи - OCR) или файл в тип, различен от PDF (напр. DOCX).

След прилагане на горните критерии бяха премахнати езиците - Албански, Арабски, Северно Македонски, Литвански, Португалски, Турски и Виетнамски. Като следствие, в данните на Exams2 голяма част от изпити са за ученици, завършващи 12 клас и съответно служат за получаване на диплома за висше образование. Това означава, че постигането на добри резултати изисква знания от всички години на обучение.

3.3. Предварителна обработка

Целите на предварителната обработка са:

1. Да се формира финално множество от изпити, които ще бъдат използвани за създаването на набора от данни.
2. Да се преобразуват всички PDF файлове (съответстващи на изпити) в серия от снимки, където всяка снимка представлява един въпрос, и файл с анотации на ниво изпит.

Наборът от данни Exams бе разширен и изменен с цел формиране на финалното множество изпити. Новият вариант включва въпроси от Китайски език и Английски език, които не са част от оригиналното множество. Също така се включват и въпроси с два възможни отговора, тип - “Истина” и “Лъжа”. За Български език, предмет Физика, бяха добавени въпросите от редовните и поправителните матури за 12 клас за годините 2021, 2022 и 2023.

Файловете се преобразуват от PDF формат в снимки, от които лесно да се изрежат въпросите. Технически, най-лесният и бърз начин бе преобразуването на всяка страница към картинка. За това бе използван пакет написан на езика Python - pdf2image [20]. След това бе използван пакетът ModifiedOpenLabelling [21], който позволява създаването на очертания около всеки въпрос във всяка снимка (която е страница с въпроси). Наред с това има възможност и да се поставят различни цветове на очертанията. Те се използват за разграничаване

между двата типа въпроси - само с текст и текст плюс картинка. Резултатът от използването на ModifiedOpenLabelling е директория с текстови файлове. За всяка картинка се създава текстови файл и в него на всеки ред се поставят данните от очертанието на картинката във формата:

```
тип_въпрос точка1 точка2
```

Тук `тип_въпрос` заема стойности `0` или `1`. Ако е `0`, то въпросът е само с текст, иначе е с текст и картинка. Стойността `точка1` маркира координатите на точката, на която се е намирала мишката при първото щракане (т.е. започване на изграждане на очертанието), а `точка2` - координатите при второто щракане (т.е. приключване на очертанието, след което следва запазване на нов ред със съответните стойности в текстовия файл).

Примерно съдържание на текстовия файл:

```
0 0.49570588235294116 0.1645909090909091 0.7382352941176471 0.101
1 0.49423529411764705 0.2815 0.744 0.135
0 0.4792941176470588 0.42095454545454547 0.7225882352941176 0.135
```

Използвайки множеството от текстови файлове, дефиниращи очертанията между въпросите и картинките за всяка страница от всеки изпит, могат да се получат изрязани въпроси. За изрязването на въпросите от всяка страница, бе използвана библиотеката Pillow [22].

3.4. Анотиране

Анотацията представлява файл с метаинформация за въпросите от даден изпит. Избраният тип бе JSON, т.к. представя лесен за създаване, обработка и разбиране йерархизиран вид на нелинейна информация, тип “ключ” - “стойност”. За всеки изпит бе създаден файл `annotations.json`, който представлява списък от JSON обекти, където всеки обект е въпрос. За всеки въпрос се попълва информация за следната структура:

- `id`: уникален идентификатор, генериран от пакета `uuid`;
- `question_snapshot`: път до текущо-описвания изрязан въпрос за съответния изпит (част от вложен обект `question`);
- `question_number`: пореден номер на въпроса в изпита (част от вложен обект `question`);
- `answerKey`: верен отговор на въпроса;
- `grade`: ниво на обучение на ученика, явяващ се на изпита (клас) (част от вложен обект `info`);
- `subject`: предмет, за който се отнася въпроса (част от вложен обект `info`);
- `language`: език, на който е написан въпроса (част от вложен обект `info`);
- `date`: дата на провеждане на изпита (част от вложен обект `extra`).

Примерно съдържание на файла с анотации:


```
{
	"id": "a3fc2e72-2fbd-42f2-9caf-c862d57cf4bb",
	"question": {
	   "question_snapshot": "BG/himiq-2009-sept/page_03_cropped_01.jpg",
	   "question_number": 15
	},
	"answerKey": "B",
	"info": {
	   "grade": 12,
	   "subject": "Chemistry",
	   "language": "Bulgarian",
	   "extra": {
	        "date": "2009-09-01"
	   }
	}
}
```


Попълнената информация е налична при изтеглянето на изпита и в повечето случай е написана и на първите страници от изпита.

За подсигуряване на консистентен формат във всички файловете с анотации, бе създаден програмен код, който проверява дали тази JSON структура, е спазена във всеки файл с анотация.

3.5. Последваща обработка

Всяка образователна система си има своите специфики, което резултира в различия между учебните програми, учебните направления и наименованията на учебните дисциплини. Това води до много голяма вариативност във възможните предмети, което не позволява добър анализ на зависимостта им между различните държави. Изходните предмети са 81, което е голямо количество. Въпреки това, голяма част от тях идват от Полски език - 55. Това се дължи на високо специализираните предмети, част от профилирани специалности.

С цел предотвратяване на голямата вариативност в [1] се предлага двустепенен процес за агрегиране на изходните предмети, който се използва и при последващата обработка на Exams2. Резултатът са три логически свързани йерархизирани групи - първата е изходният предмет, втората е нормализираният му вид, а третата обхваща трите клона на Науката: Наука за природата - изучаването на естествени феномени, Социална наука - изучаване на човешкото държание и общества, Други - Приложна наука, Изкуство, Религия и др. Процесът на агрегация протича в две стъпки: първо всеки предмет се поставя в своя собствена категория и докато има други предмети, подобни на него като име и/или насоченост, те се обединяват, оставяйки по-общото име. Процесът продължава, докато не се получат общи имена на предмети без подходящи кандидати за сливане. Процесът помага и от гледна точка на консистентност на

имената на предметите между различните езици (т.е. преименуване). В частност за Български език, предметите “Човекът и природата” и “Човекът и обществото” се преименуват съответно на “Биология” и “Социология”, което съответства и на тяхната насоченост. След агрегация броят предмети на ниво 2 е 19.

За автоматично оценяване на избраните моделите, първоначалната файлова структура, налична в Exams, бе изменена към формата:

```
директория за всеки език
       директория за всеки предмет
              директория за двата типа въпроси (текст, текст и картинка)
                     JSON файл с анотации
                     директория с картинки
```

Структурата бе образувана след като всички изпити бяха обработени. Новите файлове с анотации са формирани от оригиналните. Като резултат структурата им е сходна, но представляват списък от анотациите за всяка тройка `език-предмет-тип_въпрос` вместо да описват въпросите за всеки изпит.

3.6. Формиране на тестово множество

С цел бъдещи сравнения от данните е формирано тестово можество. То се състои от случайно избрани въпроси. Критерият за избор е броят въпроси на комбинацията `език-предмет-тип_въпрос` да е поне 20. За всички тройки, които изпълняват условието и тяхната бройка е между 20 и 50, са взети всички налични въпроси. За всички тройки, които изпълняват условието и тяхната бройка е над 50, бе извършена случайна извадка с размер 50. В резултат е формирано тестово множество с 3,036 въпроса.

4. Статистики на данните

Наборът от данни Exams2 съдържа 17,420 въпроса в 14 езика от няколко езикови групи. Всеки въпрос съдържа от 2 до 5 възможни отговора, от които винаги само един е верен. Разпределението на въпросите спрямо езика е представено в таблица 3.

Език	Въпроси с картинка и текст	Въпроси с текст	Общо
Български	502	1,630	2,132
Китайски	1,765	870	2,635
Хърватски	713	3,260	3,973
Английски	123	357	480
Френски	66	373	439
Немски	174	645	819
Унгарски	695	3,106	3,801
Италиански	4	40	44

Полски	421	2,090	2,511
Румънски	0	5	5
Руски	0	9	9
Сръбски	67	160	227
Словашки	4	42	46
Испански	89	210	299
Общо	2,858 (24%)	11,927 (76%)	17,420

Таблица 3: Разпределение на въпросите в Exams2.

Първоначалният брой предмети е 81. След агрегация предметите на второ ниво са 19. Таблица 4 показва разпределенията на броят предмети за всеки език.

Език	Брой предмети	Брой агрегирани предмети
Полски	55	1
Хърватски	17	13
Унгарски	8	7
Китайски	6	6
Немски	5	5
Български	4	4
Сръбски	4	4
Английски	3	3
Френски	3	3
Италиански	2	2
Испански	2	2
Румънски	1	1
Руски	1	1
Словашки	1	1

Таблица 4: Разпределение на предметите по език

Таблица 5 показва разпределението на броя въпроси за всеки агрегиран предмет.

Агрегиран Предмет	Брой въпроси	Процент
Физика	4,985	28.62
Професионални	2,511	14.41
Химия	2,437	13.99
География	1,259	7.23

Бизнес икономика	1,149	6.6
Биология	1,088	6.25
Наука	823	4.72
История	709	4.07
Селско стопанство	652	3.74
Социология	559	3.21
Политика	270	1.55
Информатика	188	1.08
Етика	180	1.03
Религия	161	0.92
Психология	154	0.88
Философия	144	0.83
Туризъм	76	0.44
Изкуства	48	0.28
Озеленяване	27	0.15
Общо	17,420	100

Таблица 5: Разпределение на агрегираните предмети

На трето ниво от агрегацията разпределението между трите клона на науката е показано на Таблица 6.

Клон	Брой въпроси	Процент
Наука за природата	9,333	53.58
Социална наука	4,424	25.4
Други	3,663	21.03
Общо	17,420	100

Таблица 6: Разпределение по основните клони на науката

За Китайски език всички въпроси са взети от изпитни банки за 4-ти клас. За Български език има въпроси от 4-ти и от 12-ти клас. За всички други езици въпросите са взети от изпити за 12-ти клас (последният клас, успешното преминаване на който, води до получаване на диплома за завършено средно образование). Поради тази причина трудността на въпросите се очаква да е висока и отговарянето им изисква възможност за комплексни разсъждения.

5. Отправна точка за бъдещи сравнения

През март 2023 софтуерната компания Гугъл предостави за свободно използване многомодален и многоезичен чатбот - Bard. Това е широко приемано да бъде отговор и съответно пряк конкурент на създадения по-рано и постигнал

много успехи чатбот, създаден от компанията OpenAI - ChatGPT. Докато ChatGPT е базиран на архитектурата GPT, Bard използва първоначално LaMDA (Language Models for Dialog Applications), след това PaLM 2 (Scaling Language Modeling with Pathways 2), а наскоро и най-новия голям многоезичен модел на Google - Gemini. Първоначално Bard, използвайки LaMDA, приема инструкции само под формата на текст и извършва разнообразни задачи, свързани с отговор на въпроси, резюмиране на голямо съдържание и създаване на различни видове текстово съдържание [15]. На 13 юли 2023 г. Bard се превръща в първия свободно достъпен многоезичен и многомодален модел, който може приема картинки и текст като вход. По това време съществуват и други модели, за които се твърди, че могат да работят с картинки, но те все още не са публични. Такъв модел е GPT4. Той наследява GPT 3.5 Turbo - големият езиков модел, използван от ChatGPT. Новите модели стъпват върху възможностите на предишните, но ги и надграждат и подобряват - моделът LaMDA се характеризира с възможност за провеждането на дълги разговори, които силно наподобяват реални. Моделът PaLM 2 надгражда и подобрява възможностите за генериране на нов текст и превод между различни езици. Моделът Gemini разширява тези възможности и поставя фокус върху генерирането и обясняването на програмен код, както и работа с нови видове вход - аудио и видео. Като резултат Google Bard може да превключва между различни начини на комуникация и да адаптира своите отговори въз основа на контекста на разговора. В допълнение чатботът има достъп до Интернет и не е ограничен във времето, т.е. може да бъде използван, за да резюмира и анализира нова информация. Основният конкурент на Bard - ChatGPT е ограничен да работи с информация само до определен момент във времето. Към момента това е януари 2023 г.

Bard е свободно достъпен чрез уеб приложение в браузъра. Макар и по-трудно е възможно да бъде достъпен и чрез код и интерфейс [16]. Този интерфейс не е официален и се поддържа от неговия създател, който не е обвързан с Гугъл. В резултат на това функционалностите имат технически лимитации. Понякога се поражда и неочаквано поведение при използването на някои методи, което е следствие от честото обновяване на интерфейса и вътрешната логика на Bard като продукт.

Пример за такава лимитация е горната граница на броя заявки за получаване на отговор, които могат да се изпращат в рамките на определен времеви период. Оказва се, че в зависимост от държавата, от която се правят заявките, и в зависимост от времето на деня, когато се правят заявките, броят на последователните заявки, за които ще може да се върне отговор от Bard, варира значително. От една страна, това е следвствие от неравномерното натоварване, но, от друга, експериментите показват, че в рамките на една сесия могат да се изпратят максимум от 50 до 60 въпроса. Сесията се дефинира от три бисквитки, които се използват в кода при осъществяването на връзката към Google Bard. При изтичането на валидността на поне една от тях, се появява грешка с текст *Response Error*.

Google Bard бе използван за получаване на оценка, служеща за отправна точка при бъдещи сравнения в периода ноември 2023 - декември 2023 до появяване на модела Gemini. Това гарантира, че използваният от чатбота модел през цялото време е PaLM 2. Поради ограничението в броя заявки на ден, Bard бе оценен върху първите 1,400 примера в тестовото можество.

Използваната инструкция при изпращане на всяка изрязана снимка на въпрос е: *The image has an multiple choice question with 2 to 5 choices. Provide the*

correct answer precisely in JSON format as follows: {'answer': 'xxx'}. Replace 'xxx' with the appropriate letter: 'A' if the first choice is correct, 'B' if the second choice is correct, 'C' if the third choice is correct, 'D' if the fourth choice is correct, or 'E' if the fifth choice is correct. Въпреки че инструкцията е подробна и в частност изисква получаване на отговора във формат JSON, получените резултати рядко спазваха посочения формат. В повечето случай бе представен се даден отговор и последващи го обяснения. Това наложи нуждата от последваща обработка върху получените отговори от Bard за изваждане на буквата, съответстваща на отговора.

Постигнатите резултати са добри - 46% точност. Въпреки това те служат, за да покажат, че все още може да се търси подобряване на моделите и че създаденият набор от данни е достатъчно комплексен, за адекватно оценяване на големи многомодални модели. Таблица 7 показва постигнатите резултати на Bard спрямо езика.

Език	Точност
Немски	0.5949
Френски	0.5833
Унгарски	0.5333
Български	0.5292
Италиански	0.5000
Словашки	0.5000
Хърватски	0.4219
Китайски	0.3733
Общо	**0.4579**

Таблица 7: Резултати върху тестовото множество спрямо език

Резултатите спрямо езика са консистентни, с изключение на тези за сръбски и хърватски език. Причините за по-ниските резултати при тях може да варират, но трудността на граматиката и лексиката на езика, както и количеството материали, достъпни в Интернет за съответния език, оказват силно влияние.

Следващите две таблици показват връзката между езика и предмета.

Агрегиран Предмет	Точност
Етика	0.7800
Социология	0.7800
Туризъм	0.6316
Биология	0.5289
Озеленяване	0.4815
География	0.4672
Бизнес икономика	0.4500

История	0.4474
Химия	0.3937
Наука	0.3600
Физика	0.3333
Изкуства	0.2609
Информатика	0.1515

Таблица 8: Резултати върху тестовото множество спрямо предмет

Предметите Информатика и Изкуства се срещат само в един език - Хърватски, което не позволява получаването на цялостна представа за точността на модела. Голяма част от въпросите от предметът Физика съдържат и картинки. Прави впечатление, че Bard не се справя добре на тях. Това показва от една страна, че са с висока трудност, и от друга - разбирането на картинки, свързани с физически феномени, не е достатъчно добро, за да достигне до правилния отговор.

Таблица 9 показва защо е възможно резултатите за сръбски и хърватски език да са по-ниски от тези за останалите езици.

Език	Агрегиран Предмет	Точност
Български	Биология	0.7333
	Химия	0.4600
	Физика	0.3200
	Социология	0.7800
Хърватски	Биология	0.5000
	Химия	0.3500
	Етика	0.7800
	Изкуства	0.2609
	География	0.4200
	История	0.3906
	Информатика	0.1515
Френски	География	0.5833
Немски	География	0.5435
	Туризъм	0.6667
Унгарски	Бизнес икономика	0.4500
	Озеленяване	0.4815
	Туризъм	0.6047
Италиански	География	0.5000
Сръбски	Химия	0.3043
Словашки	Химия	0.5000

Китайски	Биология	0.2800
	Химия	0.3000
	География	0.4200
	История	0.5200
	Физика	0.3600
	Наука	0.3600

Таблица 9: Постигната точност спрямо езика и предмета

Въпросите за Химия при Сръбски език и Изкуства и Информатика при Хърватски език, съдържащи голямо количество химически символи, затрудняват Bard и там не се постигат толкова добри резултати. От друга страна, е възможно комплексността на езика също да оказва влияние.

6. Заключение

Настоящият преддипломен проект представи Exams2 - нов набор от данни, който може да се използва за оценяване на големи многомодални модели. Бяха описани и сравнени съществуващи набори от данни, бяха описани стъпките, следвани за изграждане на Exams2, както и бяха предоставени статистики за въпросите от гледна точка на езика, предмета и типа. С цел бъдещи сравнения, бе използван Google Bard за постигане на отправна точка.

7. Използвана литература

1. Goertzel, Ben. "Artificial general intelligence: concept, state of the art, and future prospects." *Journal of Artificial General Intelligence* 5.1 (2014): 1.
2. Hardalov, Momchil, et al. “Exams: A Multi-subject high school examinations dataset for cross-lingual and multilingual question answering.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5 Nov. 2020, https://doi.org/10.18653/v1/2020.emnlp-main.438.
3. Hardalov, Momchil, Code and Data for Exams: https://github.com/mhardalov/exams-qa
4. Lu, Pan, et al. ‘MathVista: Evaluating Math Reasoning in Visual Contexts with GPT-4V, Bard, and Other Large Multimodal Models’. arXiv [Cs.CV], 2023, http://arxiv.org/abs/2310.02255. arXiv.
5. Lu, Pan, et al. ‘Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering’. arXiv [Cs.CL], 2022, http://arxiv.org/abs/2209.09513. arXiv.
6. Antol, Stanislaw, et al. ‘VQA: Visual Question Answering’. CoRR, vol. abs/1505.00468, 2015, http://arxiv.org/abs/1505.00468.
7. Wang, Xiaoxuan, et al. ‘SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models’. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2307.10635. arXiv.

8. Zhang, Wenxuan, et al. ‘M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models’. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2306.05179. arXiv.
9. Yue, Xiang, et al. ‘MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI’. arXiv [Cs.CL], 2023, http://arxiv.org/abs/2311.16502. arXiv.
10. Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4999–5007, 2017.
11. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In ICCV, 2015
12. J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In EMNLP, 2013
13. Q. Cai and A. Yates. Large-scale semantic parsing via schema matching and lexicon extension. In ACL, 2013
14. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ ar, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, 2014
15. Qin, Haotong, et al. “How Good Is Google Bard’s Visual Understanding? An Empirical Study on Open Challenges.” *Machine Intelligence Research*, vol. 20, no. 5, Oct. 2023, pp. 605–13. *arXiv.org*, https://doi.org/10.1007/s11633-023-1469-x.
16. Bard-API. *Daniel Park*, Minwoo Park 2023. *GitHub*, https://github.com/dsdanielpark/Bard-API.
17. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6904–6913, 2017.
18. Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
19. Huang, Lei, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232, arXiv, 9 Nov. 2023. arXiv.org, http://arxiv.org/abs/2311.05232.
20. Edouard Belval, A wrapper around the pdftoppm and pdftocairo command line tools to convert PDF to a PIL Image list: https://pypi.org/project/pdf2image .
21. Ivan Goncharov, A modified version of https://github.com/Cartucho/OpenLabeling OpenLabelling tool: https://github.com/ivangrov/ModifiedOpenLabelling .
22. Jeffrey A. Clark (Alex), The Python Imaging Library adds image processing capabilities to your Python interpreter: https://pypi.org/project/Pillow .