Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Катедра „Компютърна информатика”

ДИПЛОМНА РАБОТА

на тема

„Автоматично отговаряне на многомодални въпроси”

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 **2**
1.1 Актуалност на проблема и мотивация 2
1.2 Цел и задачи на дипломната работа 3
1.3 Очаквани ползи от реализацията 4
2 Преглед на областта **5**
2.1 Основни понятия 5
2.2 Разгледани модели 5
2.3 Подобни разработки 7
2.3.1 Отговаряне на визуални въпроси 7
2.3.2 Отговаряне на научни въпроси 9
2.3.3 Отговаряне на въпроси от учебник 10
2.3.4 Сайбенч 12
2.3.5 Матвиста 13
2.3.6 Масивно многодисциплинарно многомодално разбиране 15
2.3.7 Многоезикова, многомодална, многонивова оценка 16
2.3.8 Екзамс 19
2.3.9 Сравнителен анализ и изводи 19
2.4 Многомодални и многоезикови модели 21
2.4.1 Генеративен претрениран модел 3,5 турбо 21
2.4.2 Генеративен претрениран модел 4 23
2.4.3 Разговорен агент Бард и модели от фамилията Джеминай 26
2.4.4 Лава 29
2.4.5 Куен виел 31
2.4.6 Сравняване на моделите 33
3 Набор от данни Екзамс-В **35**
4 Експерименти **47**
4.1 Подготовка 47
4.2 Провеждане 47
4.3 Отправна точка за сравнения 50
4.4 Резултати 52
4.5 Резултати със сегмент с паралелни въпроси на различни езици 56
4.6 Резултати със сегмент с различни типове визуална информация 57
5 Дискусия **58**
6 Заключение **59**
7 Използвана литература **62**

1 Увод

1.1 Актуалност на проблема и мотивация

Изкуственият интелект започва своето развитие почти паралелно с развитието на компютърните системи във втората половина на двадесети век. До преди десетина години обаче прогресът в областта е много бавен. От множеството приложни сфери, които изкуственият интелект обхваща, една започва да работи много добре и с това отприщвайки своя потенциал да решава реални бизнес проблеми - машинното самообучение. Три са основните причини за наблюдавания експоненциален прогрес:

1. Развитието на дълбокото машинно обучение.
2. Наличието на големи данни.
3. Наличието на много голяма изчислителна мощност.

Появата на големия езиков модел Генеративен претрениран модел 3 е резултат от комбинацията на тези фактори и бързо се утвърждава като технологията, която е следващата стъпка в развитието на изкуствения интелект. За кратко време изследователската общност разбира, че подобрението, което този модел внася в областта, е грандиозно. Наблюдаването на подобно стъпково подобрение в една генерация модели е рядко срещано. Наред с това обаче изглежда най-добрите модели се дължат не толкова на нови открития в областта, а по-скоро на използване на по-голямо количество данни и по-голям капацитет от изчислителна мощност.

Обобщен изкуствен интелект е съвкупността от възможностите на машина да разбира широка гама съдържание (включ. текстово, визуално, аудио) и да взема решения и предприема действия, използвайки го. Концептуално тези възможности са 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].

Възможността на големите многомодални и многоезикови модели да притежават както строго специфични знания, така и обща култура за свойствата на обекти и универсални истини, им позволява да се използват в голям мащаб. Тази обща култура е част от породилите се знания в процеса на тренирането и именно те са към момента център на интереса в областта - как системите с изкуствен интелект научават неща, за които не са били тренирани?

Многомодалните и многоезичните модели заемат все по-важна роля в получаването на кратък и точен отговор на различен тип въпроси. От генериране на програмен код и анализиране на съдържанието на картинки до отговаряне на въпроси, свързани с управлението на човешки ресурси, тези модели демонстрират гъвкавост и адаптация към много области, независимо от езика. Те намират приложение и в контекста на образователната система с възможност за бързо даване на точен отговор на въпроси от затворен тип. Въпросите могат да се различават в три аспекта: могат да включват само текст, текст и картинка,

или могат да се състоят само от текст, но отговорите им да съдържат изцяло картинки.

Необходимо е да се оцени модела цялостно в контекста както на прости разпознавателни умения в картинки, така и в комплексни дисциплини, изискващи задълбочено мислене и специфични за домейна знания. По този начин ще се подобри разбирането за прогреса към обобщен изкуствен интелект, който рефлектира типа експертиза и умения за мислене, очаквани от умели възрастни хора в най-разнообразни професионални области.

Важно в тази връзка е минимизиране на предубежденията при подбиране на моделите, наборите от данни, стратегиите за оценяване, тестовите множества, както и примерите, които стават основа за анализ на грешките на моделите впоследствие. Поставянето на фокус на качеството на резултатите, постигнати от големи многомодални модели изцяло и само върху данни, събрани от образователните системи по света може и да не е достатъчно за валидиране на съществуването на обобщен изкуствен интелект. Въпреки това постигането на добри резултати върху такива набори от данни е признак за развита мисловна дейност, съответстваща на богата обща култура и задълбочено познаване на високо-специализирани области. Добро разширение на анализа на качеството на резултатите е проверката с човешката точност. Това позволява по-добро сравнение между възможностите на моделите и експертните резултати, което от своя страна ще позволява по-ясно измерване на разстоянието от текущото състояние на изкуствения интелект до обобщения изкуствен интелект.

Глава 1 съдържа увода с описание на мотивацията, целта и задачите на дипломната работа. В глава 2 се разглеждат изследвания и разработки в областта на оценяване на качеството на резултатите, върнати от големи многомодални модели, както и най-широкоизползваните и популярни такива модели. Глава 3 въвежда набора от данни Екзамс-В, който се използва за провеждането на експериментите с описаните в глава 2 модели. Четвъртата глава описва процеса на подготовка на данните, тяхното изпращане към моделите и резултатите, постигнати с всеки от моделите. Глава 5 описва изводите от провеждането на експериментите. Последната (шеста) глава съдържа заключенията и възможностите за бъдещо развитие.

1.2 Цел и задачи на дипломната работа

Настоящата дипломна работа има за цел да създаде прототип на система за автоматично отговаряне на многомодални въпроси, давани на ученици от различни нива на обучение.

За изпълняване на целта са поставени и изпълнени следните задачи:

1. Подробен аналитичен обзор на изследванията в областта: проучване на начини за формиране на набори от данни за оценяване на модели, решаващи задачата; проучване на подходящи метрики и начини на подготовка и провеждане на експерименти.

2. Подробен анализ на съвременни големи езикови модели и големи многомодални модели. Проследяване на историята, довела до успеха на най-широкоизползваните модели в днешно време.
3. Въведение в набора от данни Екзамс-В и предварителна обработка на данните за подаване към моделите.
4. Извършване на сравнителен анализ на характеристиките на съвременните големи многомодални модели и големи езикови модели, като тези от фамилиите Джеминай и Генеративен претрениран модел.
5. Планиране и провеждане на експерименти за оценяване и сравняване на точността, постигната от:
    - големи многомодални модели;
    - големи езикови модели с извличане на текста от подадена картинка и със създаване на описание за съдържанието на подадена картинка.
6. Анализ на грешките, допускани от моделите:
    - анализиране на качеството на цялостните резултати;
    - анализиране на качеството на резултатите, когато един и същ въпрос е преведен на няколко езика;
    - анализиране на качеството на резултатите, когато въпрос притежава различни видове визуални характеристики.
7. Предаване на статия *EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models* за рецензия в *Annual Meeting of the Association for Computational Linguistics Rolling Review* (*ACL ARR*), описваща:
    - използваните методи и подходи за решаване на задачата;
    - формирането на новото множество Екзамс-В за оценяване на големи многомодални модели;
    - постигнатите резултати от експериментите и изводи.

1.3 Очаквани ползи от реализацията

Системата може да се използва за получаване на верните отговори на група въпроси, които имат краен брой отговори. Тя може да бъде вградена в смартфони и да се прилага върху сниман с устройството въпрос. Може да се използва и в други устройства, които имат камера и връзка с Интернет.

2 Преглед на областта

Глава 2 представя основните понятия и аналитичния обзор на съществуващите статии и най-широкоизползваните модели. В допълнение, в секция 2.2 накратко са представени дискутираните модели в разгледаните статии, а моделите, използвани в разработката за дипломната работа са представени в секция 2.4.

2.1 Основни понятия

Езиков модел (от англ. *language model*) - компютърен модел, пресмятащ вероятност за последователност от думи.

Голям езиков модел (от англ. *large language model*) - езиков модел с много параметри, 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 освен текстово може да получава и визуално съдържание.

Модалност на данни (от англ. *data modality*) - различни измерения на информация. Примери: текст, видео (в частност картинки), аудио.

Разговорен агент (от англ. *chatbot*) - допълнителен интерфейс към голям езиков модел, позволяващ провеждането на разговор с модела.

Основен езиков модел (от англ. *foundational model*) - езиков модел, който е обучаван с по-голямата част от данните, достъпни в Интернет. Характеризира се с възможност да симулира разсъждения (за основни цели) в произволен контекст и с възможност да може да се специализира към разнообразни области.

Голям многомодален модел (от англ. *large multimodal 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 Разгледани модели

Много дълбока конволюционна невронна мрежа (от англ. *Very Deep Convolutional Networks (VGG)*) - архитектура на дълбока конволюционна невронна мрежа, известна с успеха си в разпознаването на изображения.

Невронна мрежа с дълга краткосрочна памет (от англ. *Long short-term memory (LSTM)*) - вид рекурентни невронни мрежи, които могат да идентифицират значима информация в дълга последователност от данни.

Обединено отговаряне на въпроси (от англ. *Unified Question Answering (UnifiedQA)*) - претрениран модел, който може да обработва въпроси с различни формати, в частност въпроси със затворен и отворен отговор.

Генеративен претрениран модел 3 (Джипити 3) (от англ. *Generative Pre-trained Transformer 3 (GPT-3)*) - голям езиков модел, използващ декодер слоеве от архитектурата Трансформър, който набира бързо популярност с висококачествено създаване на текст, превод между езици и отговор на въпроси, изискващи богата обща култура.

Бърт (от англ. *Bidirectional Encoder Representations from Transformers (BERT)*) - езиков модел, използващ енкодер слоеве от архитектурата Трансформър, за да решава задачи, свързани с обработка на текст на английски език.

Екселем робърта (от англ. XLM-RoBERTa) - голям езиков модел, трениран с повече данни от Бърт и работещ с множество езици, който постига по-добри резултати от Бърт на задачи за обработка на текст.

Визуален Бърт (от англ. *VisualBERT*) - разширение на архитектурата на Бърт, имплементацията на което позволява работа с текст и картинки, чрез множество от конволюционни слоеве.

Визуален трансформър (от англ. *Vision Transformer (ViT)*) - визуален модел, използващ архитектурата Трансформър за обработка на изображения, разбивайки ги на последователности от части и прилагайки техниката "внимание", за извличане на глобални и локални шаблони от тях.

Визуален и езиков трансформър (от англ. *Vision-and-language Transformer (ViLT)*) - разширение на архитектурата на визуалния трансформър, имплементацията на което позволява работа с текст и картинки без използването на конволюционни слоеве.

Пирамидален кросмодален трансформър (от англ. *Pyramid patch cross-modal Transformer (Patch-TRM)*) - визуален модел, използващ архитектурата Трансформър за обработка на изображения, разбивайки ги на последователности от части, които формират пирамидална (йерархична) зависимост за по-ефективно представяне на информацията.

Дълбока мрежа за последователно семантично анализиране на диаграми (от англ. *Deep Sequential Diagram Parser Network (DSDP-NET)*) - невронен модел, който е обучен последователно да формира представяне на диаграми, използвайки невронна мрежа с дълга краткосрочна памет.

Блуум (от англ. *BigScience Large Open-science Open-access Multilingual Language Mode (BLOOM)*) - голям езиков модел, използващ енкодер слоеве от архитектурата Трансформър за обработка на текст.

Модели **Клауде** и **Клауде 2** (от англ. *Claude*, *Claude 2*) - големи езикови модели, разработени от софтуерната компания Антропик. Клауде е първият модел, създаден от компанията, с размер на контекстния прозорец - 9,000 токена.

Клауде 2 е подобрена версия с раширен контекстен прозорец, който позволява работа с до 100,000 токена.

Викуна (от англ. *Vicuna*) - разговорен агент, получен чрез фина настройка на големия езиков модел Лама 2.

Мини генеративен претрениран модел 4 (Мини джипити 4) (от англ. *MiniGPT-4*) - многомодален модел, използващ тегла на тренирани модели Викуна и Визуален трансформър, за постигане на резултати и възможности, сравними с тези на Генеративен претрениран модел 4 (*GPT-4*).

Блип 2 (от англ. *Bootstrapping Language-Image Pre-training 2 (BLIP-2)*) - многомодален модел, използващ тегла на тренирани модели за обработка на изображения и модели за обработка на текст и обединявайки ги чрез 12 енкодер слоеве от архитектурата Трансформър.

Инструкционен Блип 2 (от англ. *InstructBLIP2*) - фино настроен модел, който може да следва инструкции за генериране на текст на база инструкция и картинка, използващ за основа Блип 2.

Опън фламинго (от англ. *OpenFlamingo*) - голям многомодален модел, който може да обработва изображения и текст и да създава текст.

Фромаж (от англ. *Frozen Retrieval Over Multimodal Data for Autoregressive Generation (FROMAGe)*) - голям многомодален модел, който може да обработва изображения и текст и да създава изображения и текст.

2.3 Подобни разработки

2.3.1 Отговаряне на визуални въпроси

Задачата за даване на отговор в свободен стил след получена картинка като вход се въвежда за първи път в [6] и бързо намира популярност в различни области [12, 13, 17, 18]. 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 е насочен към конкретни части на картинката. Визуалното съдържание е подбрано от наборът данни “Майкрософт популярни обекти в контекст” [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 в данните. 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.

При подбора на моделите за оценяване, върху формираното от тях множество - Отговаряне на визуални въпроси (съкратено на англ. *VQA*), авторите избират:

1) модел, който винаги дава като отговор най-често срещания отговор в тренировъчния набор от данни – в случая, този отговор е “Да”;
2) модел, който дава като отговор най-често срещания отговор за всеки тип на въпроса (има пет типа въпроси, разделени спрямо въпросителната дума, с която започват - “какво”, “колко”, “защо”, “къде”, “кой”);
3) модел, който дава отговор, който е най-често срещания отговор в K най-близки съседа на текущата двойка въпрос-картинка;
4) модел, който дава отговор, след като кодира въпроса, използвайки предварително създаден речник от думи (от англ. *bag-of-words*). От най-често срещаните 1,000 думи във въпросите са извлечени най-често срещаните 10 първи, втори, и трети думи от въпросите. Така се създава влагане на думи с размерност 1,030;
5) модел, използващ невронната мрежа Много дълбока конволюционна невронна мрежа [11], за да даде отговор на въпроса, използвайки единствено изображението;
6) модел, базиран на невронна мрежа тип еднопосочна Невронна мрежа с дълга краткосрочна памет.

Резултатите им показват, че моделът, разчитащ само на картинката, не се представя добре, при това постига по-слаби резултати от модела, който винаги отговаря с “Да”. В контраст с това, моделите, които разчитат само на текстовата част на въпроса, се представят доста добре, постигайки около 50% точност и задминавайки модела с най-близките съседи, който работи и с картинката. Авторите предполагат, че това се дължи на априорните вероятности, имплицитно заложени във въпроса. Например, за отговаряне на въпроса “Какъв е цветът на бананите?” не е нужно да се гледа картинка.

Най-добрият модел постига точност около 60%. Той е комбинацията на двунивова Невронна мрежа с дълга краткосрочна памет за влагане на думите от въпроса и Много дълбока конволюционна невронна мрежа с нормализация от второ ниво (от англ. *l2 layer normalization*) за влагане на изображението. Невронна мрежа с дълга краткосрочна памет се състои от два скрити слоя и води до получаването на влагане с размерност 2,048. Влагането на текста е резултат от прилагането на активационна функция тип хиперболичен тангенс върху вектора, получен в резултат на разширяването в хоризонталното измерение на изходния вектор от последната клетка със скрития вектор от последната клетка. Двете влагания са обединени чрез поелементно умножение, след което следва нелинейност и активационна функция тип софтмакс слой за получаване на крайния отговор. Наблюдава се, че моделът е добър в разпознаването на цветове и популярни обекти (кораби, хора, коли, различни типове животни), но изпитва затруднение при броене и в частност, когато верният отговор включва в себе си стойност над 5. Спрямо трудността на въпросите, авторите показват, че моделът се представя аналогично на дете, което е на 4,74 години. Като следствие, докато точността на модела е 61,07% във възрастовата група от 3 до 4 години, тя спада до 47,83% при възрастите от 18 години нагоре.

2.3.2 Отговаряне на научни въпроси

Създаването на първия набор от данни, фокусиран върху многомодални въпроси, взети от изпити, давани на ученици, е описано в [5]. Той 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 и твърдения, формирането на които позволява достигането до крайния отговор. Обхванати са три предмети, изучавани главно в класовете от първи до шести - биология, социология и лингвистика. Отличителни черти на наборът от данни Отговаряне на научни въпроси (съкратено на англ. *ScienceQA*) са големият брой многомодални въпроси и прилежащите към тях специални анотации - лекции и обяснения. Това позволява да се оцени качеството на веригата от разсъждения, които модела поражда при достигането до крайните отговори, което от своя страна валидира доколко моделът разбира въпроса и имитира човешко разсъждение и доколко халюцинира [19].

Оценяването се фокусира върху проучване на възможността за създаване на логически свързани обяснения с цел разобличаване на мисловия процес при отговаряне на въпросите в набора от данни.

Авторите избират три типа модела за оценяване.

Първият обхваща модели, базирани на евристиката: случаен избор и човешка точност. Случайният избор се изразява в избор на един от възможните отговори с равномерна вероятност. Извършват се три обхождания на тестовото множество за пресмятане на средна точност. За пресмятане на човешката точност авторите използват услуга, предоставена от софтуерната компания Амазон на име *Amazon Mechanical Turk*. *Amazon Mechanical Turk* е платформа, в която регистрирани работници (реални хора) извършват онлайн услуги, които могат да варират от валидиране на данни до проучване на подходящи за използване спрямо спецификата на задачата методи за решаването ѝ.

Вторият тип модели са големите езикови модели Обединено отговаряне на въпроси [23] и Генеративен претрениран модел 3 [24]. В експериментите без подаване на предварително оценени примери, използваният формат за вход-изход е ВКО→И, където входът представлява обединение на текста на въпроса (В), контекста му, който, в случая на многомодални въпроси, включва и обяснение за съдържанието на дадена картинка (К) и възможните отговори (О), а изходът е отговорът на модела (И). За получаване на обяснение за съдържанието в картинките се използват моделите Визуален трансформър [25] и Генеративен претрениран модел 2 [26]. Провеждат се и експерименти с подаване на няколко предварително оценени примера с цел предоставяне на възможност за адаптация към специфичната задача на наборът от данни Отговаряне на научни въпроси.

Третият тип модели са фино настроени големи визуално-текстови модели (като Визуален Бърт и Визуален и езиков трансформър) и фино настроен модел за Обединено отговаряне на въпроси. Големите визуално-текстови модели приемат въпроса, тестовия му контекст (ако такъв съществува), възможните

отговори и прилежащата картинка и създават точково разпределение над възможните отговори чрез линеен класификатор. Фината настройка на модела за Обединено отговаряне на въпроси използва подхода, описан във втория тип експерименти – текстовото съдържание се приема директно, а визуалното се преобразува в текстово, използвайки обяснение за неговото съдържанието.

Получените резултати показват, че от големите визуално-текстови модели най-добре се представя Визуален Бърт, постигайки средна точност от 61,87%, следван от Пирамидален кросмодален трансформър, който се представя по-добре в предмети, свързани с науката за природата и, в частност, постига по-висока средна точност на въпроси, давани на ученици от седми до дванадесети клас (67,50% срещу съответно 59,92%).

Без фина настойка и без подаване на предварително оценени примери големият езиков модел за Обединено отговаряне на въпроси не успява да постигне по-добри резултати от кой да е визуално-текстови модел, но все пак се представя по-добре от случайния избор. След фина настойка обаче моделът успява да постигне средна точност от 70,12%, която се увеличава с 4 пункта, когато изискваният резултат включва и веригата от разсъждения. Тези резултати показват, че генерирането на веригата на разсъждения, заедно с отговора подобрява качеството на разсъжденията и резултатите, постигнати от езиковите модели.

Без фина настойка и без подаване на предварително оценени примери големият езиков модел Генеративен претрениран модел 3 постига много добра точност – 74,04%. Положителният ефект от извличането на веригата на разсъждения заедно с отговора оказва положително влияние и на този модел. Генеративният претрениран модел 3 постига 74,17% точност след нейното прилагане, с което се превръща в най-добрия модел във всички експерименти. Точността му не е толкова далеч от тази на човек (88,40%), което показва, че дори и да обработва картинките косвено (вместо пряко), голямото количество информация, с която е претрениран, помага значително за обобщаване на разсъжденията към разнообразни области.

2.3.3 Отговаряне на въпроси от учебник

Наборът от данни Отговаряне на въпроси от учебник (съкратено на англ. *TQA*) [10] цели да оцени възможността на модел да даде отговор на краен брой многомодални въпроси, използвайки като контекст реални уроци от учебни материали с прилежащи диаграми и картинки.

Авторите създават свой собствени модели и ги оцяняват върху набора от данни Отговаряне на въпроси от учебник. Базовият модел използва Невронна мрежа с дълга краткосрочна памет и приема на вход текстовото съдържание на въпроса и урока. Анализ на данните в набора от данни Отговаряне на въпроси от учебник показва, че в повечето случаи разполагане с текстовата информация е достатъчно за даване на отговор на въпросите, съдържащи само текст. Тази информация обаче не е достатъчна за отговаряне на въпросите, съдържащи

диаграми и картинки. На входа базовият модел приема урока като контекст, единствен въпрос и възможни отговори, достигащи най-много седем възможности. Очаква се като изход правилния отговор. Размерът на урок средно е над 1,000 думи, което не позволява съхраняването на влагането му в една графична карта по времето на създаване на експериментите. За да се справят с този проблем, авторите намалят размерността чрез подход, базиран на извличането на информация: за всеки параграф те пресмятат сходство (чрез скаларно произведение) до въпроса, сумирайки добавените стойностите на всички думи (пресметнати, използвайки честотите на срещане на думата в документа (от англ. *term frequency - tf*) и между различните документи (от англ. *inverted document frequency - idf*) и избират параграфа с най-голямо сходство. За получаването на влагания поотделно за всяко изречение от параграфа, въпроса и отговорите се използва Невронна мрежа с дълга краткосрочна памет. След това чрез механизма внимание се избира влагането на думата с най-голяма добавена стойност, която се смята за отговор на въпроса. Това влагане се сравнява по сходство с всеки от отговорите и за верен отговор се избира този с най-голямо сходство.

Авторите създават модели, приемащи картинки на входа като разширения на базовия текстови модел. Разликата между двата типа модели се изчерпва с добавения визуален контекст. Сравнени са два такива модела: един, базиран на методи, използвани за създаването на големи визуално-текстови модели, и модел, разширение на Дълбоката мрежа за последователно семантично анализиране на диаграми [27]. В първия вид визуален модел картиката преминава през невронна мрежа тип Много дълбока конволюционна невронна мрежа и стойностите на параметрите от последния конволюционен слой се считат за съответния визуален контекст, който поради спецификата на мрежата се изразява като 49 вектора, всеки с по 512 елемента. Получените вектори се трансформират към размерността, използвана за влагане на думите, използвайки два скрити слоя с активационна функция тип хиперболичен тангенс. Резултатните вектори са добавят като разширение към изходните такива от Невронната мрежа с дълга краткосрочна памет и по този начин моделът може да прилага внимание и към картинката. Вторият тип визуален модел включва използване на структурната информация в картинка за получаване на свързан ориентиран граф от обекти, който в последствие може да бъде трансформиран към изречения, описващи началото и края на всяка дъга. Задачата на модела е да постави семантично значение върху връзката между двата обекта. Следвайки този подход, всяко изображение се трансформира към няколко изречения, които се добавят в параграфа, който реферира към изображението. След това се прилага изчисляването на сходство и процесът продължава, както е в текстовия модел.

Резултатите от експериментите сочат, че текстовият модел постига 33,8% точност, което не е много по-добър резултат от случайния избор – 28,4%. В този ред на мисли постигнатата точност само върху въпроси от тип „истина-лъжа“ е едва 50,2%. Това е възможно да се дължи на голямата трудност на тези въпроси -

за правилен отговор се изисква повече създаване на верига от разсъждения, а не толкова търсене в текста, в което моделите, базирани на памет и внимание не са добри. Все пак, моделът се представя по-добре на въпроси с множество отговори, задминавайки метода на случайния избор с близо 10% пункта точност. Въпреки това точността остава ниска – 32,9%, което отново се дължи на високата трудност на въпросите. На въпросите, включващи картинка, първият тип визуален модел не постига по-добри резултати от текстовия модел, но вторият тип визуален модел, използващ структурен граф за описание на картинката, постига подобрение от почти пункт и половина спрямо останалите – 31,3%. Точността остава ниска, което се дължи на богатото съдържание на повечето диаграми, изискващо развита възможност за изграждане на причинно-следствени връзки, формирането на които изисква информация от целия урок. Това е най-добрият тестван модел с обща точност 34,6%.

2.3.4 Сайбенч

Наборът от данни Сайбенч (от англ. *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. Сайбенч не съдържа многомодални въпроси.

Отворените въпроси са 695 и са събрани от популярни университетски курсове: физика, термодинамика, класическа механика, квантова химия, диференциално и интегрално смятане и статистика. Второто множество се състои от 104 затворени въпроса, взети от семестриални контролни и изпити. С цел намаляване на вероятността за отгатване на верния отговор затворените въпроси също са представени като отворени, т.е. при оценяването на моделите не се използват възможните отговори.

Сравнени са моделите Генеративен претрениран модел 3,5 турбо (основа на разговорния агент Чатджипити) и Генеративен претрениран модел 4 с температура 0, за да се наблегне на точността и на детерминистичните разсъждения и отговори. Разгледани са два вида стратегии за инструктиране на моделите - чрез извличане на верига от разсъждения и преобразуване на част от разсъжденията към програмен код, написан на езика Пайтън или Уолфрам. Това се прави с цел постигане на по-точни резултати при смятане. Наред с това в експериментите, изключващи предоставяне на примери, се проверява добавената стойност на инструкциите, описващи типа и категориите на въпросите и какво се очаква от модела. По този начин авторите формират седем експеримента за оценяване:

1. без предоставяне на примери и без инструкции;
2. без предоставяне на примери и с инструкции;
3. с предоставяне на няколко примера;

4. без предоставяне на примери и с извличане на верига от разсъждения;
5. с предоставяне на примери и с извличане на верига от разсъждения;
6. с предоставяне на примери и преобразуване на части от разсъжденията към Python;
7. с предоставяне на примери и преобразуване на части от разсъжденията към Wolfram.

Резултатите показват, че Генеративният претрениран модел 4 се представя по-добре от Генеративния претрениран модел 3,5 турбо във всички 7 експеримента. Най-големи подобрения се наблюдават при предоставяне на примери и извличане на верига от разсъждения и при предоставяне на примери и преобразуване на части от разсъжденията към Пайтън - съответно с 16,36% и с 15,89% увеличение. Също така в общия случай не се наблюдава голяма добавена стойност на включването на примери. Осреднената точност на моделите с и без предоставени примери е съответно 12,17% и 11,99% за Генеративния претрениран модел 3,5 турбо и 28,52% и 28,35% за Генеративния претрениран модел 4. Във високоспециализирани предмети като квантова химия добавянето на примери все пак води до по-висока точност с почти 3 пункта за Генеративния претрениран модел 4, макар че в предмети като физика добавянето на примери води до намаляване на точността с 6,99%. Тези вариации могат да се дължат на степента на представителност на примерите, които се използват - очаква се, че примери, цялостно представящи разнообразието на множеството, към което принадлежат, ще водят до постигането на по-висока точност. Наблюдава се също и тенденция за постигане на по-висока точност при извличане на веригата от разсъждения. Експериментите, насочени към изследването на зависимостта между генериране на код на Пайтън и Уолфрам, показват, че използването на Пайтън увеличава средно точността съответно с 7,92% за Генеративния претрениран модел 3,5 турбо и 7,45% за Генеративния претрениран модел 4 спрямо използването на верига от разсъждения, но използването на Уолфрам намалява точността с 4,12% за Генеративния претрениран модел 3,5 турбо и с 12,79% за Генеративния претрениран модел 4.

2.3.5 Матвиста

Матвиста (от англ. *MathVista*) [4] е множество от данни с фокус върху математиката и разсъждаването върху фигури и математически обекти. Въпросите са събрани от седем дисциплини: алгебра, аритметика, геометрия, логическо мислене, прости числови операции, проучване и статистика. Обхваща широка гама визуално съдържание - естествени картини, геометрични фигури, графики и изкуствени сцени и диаграми.

Авторите отбелязват, че досега анализът почти винаги е бил фокусиран върху свеждането на развитието на моделите до качествена оценка. За разлика от това, те провеждат количествен и качествен систематизиран анализ на резултатите на основни модели, за да проверят доколко развита възможност имат те в разсъждаването при визуален контекст.

Те създават нов начин за оценяване на отговорите, давани от големи многомодални модели. Нуждата за това е продиктувана от тенденцията големи езикови и големи многомодални модели да се тренират по начин, който изисква получаването на дълъг отговор в разговорен стил, вместо кратък отговор. Авторите предлагат автоматизиран начин за задаване на въпрос и извличане на отговора му, следвайки процедура от три стъпки: създаване на отговора, извличане на отговора и пресмятане на точност. Първата стъпка включва подаване на входа инструкция заедно с въпроса и получаването на отговор. Инструкцията включва описание на формата, в който трябва да бъде върнат отговорът, въпроса, възможните отговори и метаинформация за въпроса. Следващата стъпка цели да извличе буквата или цифрата, съответстваща на верния отговор. Авторите предлагат използване на система, за извършване на тази задача, базирана на голям езиков модел. Те използват Генеративния претрениран модел 4 и посочват впечатляващото му умение за извличане на информация от текст. Моделът успява да извлече правилно отговора от текста в 99,5% от случаите. Последната стъпка включва преобразуване на извлечения отговор в предварително дефиниран формат, който след това позволява пресмятане на точността спрямо очаквания отговор.

Авторите разпределят моделите, които използват, в три направления. Първото включва големите езикови модели Генеративен претрениран модел 3,5 турбо, Генеративен претрениран модел 4 и Клауде 2 в контекста без и с предварително подаване на два примера и с извличане на верига от разсъждения и преобразуване на част от веригата, включваща изчисления, към код. Второто направление включва същите големи езикови модели, но в този случай освен текстовото съдържание на въпроса към входа се добавя и генерираното от разговорния агент Бард обяснение за съдържанието на картинката и разпознатия текст в картинката от софтуерната програма за извличане на текст “Изиосиар”. Третото направление включва големи многомодални модели, част от които са Генеративен претрениран модел 4-В, Бард (в случая се използва разговорния агент вместо Палм 2, който е основният модел, използван от Бард по това време), Лама адаптер в2 7б, Викуна и Мини джипити 4. За пресмятане на човешката точност авторите използват Amazon Mechanical Turk.

Резултатите показват, че всички модели се справят по-добре от случаен избор на отговор, но в същото време не успяват да се постигат по-добри резултати от човек (60,3%). Най-добрият модел в първото направление е Генеративният претрениран модел 4 с два предварително предоставени примера и извличане на верига от разсъждения. Той постига точност от 29,2%. Това показва, че наборът от данни изисква разсъждения, базирани на визуален контекст. След добавяне на описание на съдържанието в картинката и намереният текст в нея се наблюдава цялостно подобрение на точността. Най-добрият модел остава Генеративния претрениран модел 4, постигащ точност от 33,9% с два предварително предоставени два примера и преобразуване на част от разсъжденията към програмен код на Пайтън.

При големите многомодални модели се наблюдава голяма разлика в количеството правилни отговори, дадени от Генеративния претрениран модел 4-B и всички останали. Той постига точност от 49,9%, което е с 15,1% по-високо от втория най-добър модел – комбинация от Бард и Гугъл ленз (34,8%). Моделите с отворен програмен код не се представят задоволително. Моделът Лава постига най-висока средна точност от 26,1%, което показва, че в тези модели се наблюдава липса на възможност за добри разсъждения в математически контекст, разпознаване на текст и фигури и разбиране на графики.

Качественият анализ на резултатите показва, че разговорният агент Бард и моделите Генеративен претрениран модел 4 и Генеративен претрениран модел 4-B често връщат правилен отговор, но с грешки в обясненията. Бард изпада в тази ситуация с частично грешни разсъждения в 6,8% от случаите и с изцяло грешни разсъждения, но правилен отговор в 8,1% от случаите.

2.3.6 Масивно многодисциплинарно многомодално разбиране

Авторите на [9] представят набора от данни Масивно многодисциплинарно многомодално разбиране (съкратено на англ. *MMMU*) като такъв, създаден с цел да оценява многомодални модели на масивно многодисциплинарни 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.

Експериментите са фокусирани в оценяването на големи многоезикови и големи многомодални модели без предварително предоставени примери. За проверка дали разпознаването и добавянето на текста от картинките води до по-добри резултати, се използва пакет, написан на Пайтън за извличане на текст от картинка - Ммосиар, а за описание на съдържанието на картинките се използва Лава-1,5. За извличане на отговорите от дълги параграфи, върнати от моделите, се използват регулярни изрази и процедури за последваща обработка. При липса на валиден отговор за въпрос с няколко възможности, се избира случаен отговор, а при отворени въпроси отговорът се смята за грешен.

Интересно решение, което не се коментира в статията, е тестването на моделите от фамилията Джеминай единствено върху валидационното множество, но не и върху тестовото. Като резултат, при резултатите от валидационното множество най-добри резултати постига Джеминай ултра, следван от Генеративен претрениран модел 4-B (съответно 59% и 57% точност). Сред моделите оценени върху тестовото множество, най-добри резултати постига Генеративният претрениран модел 4-B, следван от Куен виел (съответно 56% и 41%). Отново се наблюдава голям спад в качеството на получените отговорите при модели с отворен код като Блип2 флан те5 XXL и Лава 1,5, които достигат едва 34% точност.

Авторите наблюдават, че добавянето на разпознат текст в картинката и нейно описание не води до статистически значимо подобрение в резултатите, което означава, че модел, който се справя добре би следвало да има възможност самостоятелно да интегрира визуална и текстова информация.

Спрямо различните предмети средната точност варира. Наблюдава се, че при дисциплини като изкуство и социология средната точност на моделите е по-висока, което се дължи на сравнително по-лесните и естествени въпроси. Дори и моделите с отворен код се представят сравнително добре в категории като снимки и рисунки най-вероятно, защото контекста на картинките там е често срещан при тренирането на тези модели. В дисциплини като наука, медицина и компютърни науки обаче точността е по-ниска, т.к. въпросите изискват по-задълбочено мислене. Това е част от наблюдавана тенденция на постепенно намаляване на добавената стойност на големи и комплексни модели като Генеративния претрениран модел 4-B при увеличаване на трудността на въпросите. Например, докато разликата между моделите InstructBLIP2-FLAN-T5-XXL (40,3%) и Генеративен претрениран модел 4-B (76,1%) е 34,8% при въпроси, класифицирани като лесни, тя бързо намалява при въпроси, класифицирани като трудни - моделите постигат съответно 29,4% и 31,2% точност. Класификацията на въпроси в различните нива на трудност е направена от авторите.

Проведен е и анализ на грешките, които Генеративният претрениран модел 4-B допуска. За целта на случаен принцип се избират 150 примера, за които са получени грешни резултати. Те са ръчно обходени за идентифициране на причината за грешките. Анализът показва, че 35% от грешките се дължат на неправилно възприемане. За по-ясно идентифициране на причината за тяхната поява, те са допълнително разделени на два типа: повърхностни и специфични за домейна. Повърхностни грешки моделът допуска при неразбиране на посоки. Грешки, специфични за домейна, моделът допуска, когато няма нужните знания. Също така се наблюдава тенденция моделът да поставя по-голямо внимание и тежест на текстовото съдържание. Грешките от незнание са причина за грешен отговор в 29% от разгледните случаи. Пример за това е неразбирането на крайни състояния в крайни детерминистични автомати. В контекста на медицината се наблюдава липса на достатъчно обширен контекст за правилното разпознаване на болест по подадено описание и таблица. В 26% от случаите се наблюдават грешки в разсъжденията. Това си личи най-добре в контекста на математиката, където моделът разпознава правилно необходимата информация, но липсата му на развити умения за боравене с математически инструменти води до грешен отговор.

2.3.7 Многоезикова, многомодална, многонивова оценка

Наборът от данни Многоезикова, многомодална, многонивова оценка (съкратено на английски *M3Exam*) се характеризира с многомодалност, многоезичност и въпроси, които са взети от всички нива на обучение до

достигане на средно образование [8]. Обхванати са езиците английски, китайски, италиански, португалски, виетнамски, тайски, суахили, африкаанс и явански (който се говори в Индонезия). 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, т.е. такива за които няма много информация в Интернет.

Авторите подбират големи езикови модели и големи многомодални модели. Сред големите езикови модели са Генеративен претрениран модел 3,5 турбо, Генеративен претрениран модел 4, Клауде, Блуум и Викуна. Сред големите многомодални модели са Блип 2, Инструкционен Блип, Фромаж и Опън фламинго. Оценяват се моделите без предварително подаване на примери, главно поради това, че репликират най-добре реалния свят, трудно е да се подадат няколко картинки за отделни примери и при трениране моделите преминават през инструкционно нагласяне, т.ч. са готови да отговорят на въпросите. Авторите включват името на предмета, за който се отнася въпросът, в тяхната инструкция и инструктират модела да не поражда разсъждения, а само да представи крайния отговор, който смята за верен. С цел използване на генеративното естество на моделите, инструкцията завършва с „*Answer:*“, след като предостави въпроса, подтиквайки модела към допълване с точен отговор. За всеки език инструктиращият текст се превежда.

Всички въпроси са затворени и съдържат краен набор от отговори. Авторите наблюдават, че в някои случаи моделите връщат дълги отговори, съдържащи не само идентификатора на вярната опция, но и аргументация и допълнителни примери, както и аргументи защо другите опции не са правилни. По този начин в целия отговор се появяват всички опции. Авторите вземат първата срещната опция и считат нея за верен отговор. Това крие своите рискове, т.к. е възможно първо моделът да аргументира защо грешни опции са грешни и след това да напише вярната опция по метода на изключването.

Резултатите показват, че докато средната точност, нужна за успешното вземане изпитите е около 54%, единствено моделите Генеративен претрениран модел 3,5 турбо и Генеративен претрениран модел 4 успяват да постигнат по-високи средни резултати - съответно 57,57% и 72,92%. Наблюдава се, че Генеративен претрениран модел 4 се представя най-добре в различните езици. Моделът Блуум се представя по-лошо от случайния избор. Средните точности на моделите Генеративен претрениран модел 3,5 турбо, Клауде и Викуна варират спрямо езика. Докато Генеративен претрениран модел 3,5 турбо и Клауде имат сходна точност за английски (75,98% и 74,25%), Генеративен претрениран модел 3,5 турбо постига по-високи резултати в останалите езици, демонстрирайки по-развита многоезична способност. На ниво език, средно моделите имат намалена точност при нелатинските езици като китайски (въпреки че за него има доста данни, т.е. това е неочаквано) и езици, за които няма много данни като явански, въпреки че в него главно се използва латинската азбука.

Авторите експериментират с две стратегии за инструктиране на Генеративен претрениран модел 3,5 турбо в контекста на различни езици. Първата се изразява в превод на инструкцията на английски език, но запазвайки съдържанието на въпроса на изходния език. Втората се изразява в превеждането и на инструкцията и на съдържанието на английски език. Резултатите показват, че превеждането само на инструкцията на английски език не подобрява консистентно резултатите. Възможно обяснение за това е автентичността на оригиналния език (той не е получен чрез превод от английски). Също така е възможно използването на английски език да не подсказва правилно на модела какви умения и знания са нужни за решаването на задачата. Превеждането на инструкцията и данните също не подобрява консистентно резултатите. От една страна, повечето въпроси са тясно свързани с езика, от който са написани, и превеждането им може да доведе до загубена ключова информация, което резултира в намалена точност. От друга страна, използването на преведен вариант може да елиминира трудностите при разбирането на някои езици особено тези, за които Генеративен претрениран модел 3,5 турбо вижда трудности като тайски и явански. За тях се наблюдава увеличаване на точността с по 20 пункта.

И в тази статия авторите провеждат експерименти, свързани с предварително подаване на примери. Те дават инструкцията след това няколко примера и след това тестовият въпрос. И те наблюдават, че предоставянето им не води до консистентно покачване на точността в общия случай. В частни случай като португалски и виетнамски има подобрение, в други езици като китайски и суахили се наблюдава деградация. Това отново показва колко важно е подбраните примери правилно да представят цялото, за което се водят представителни. Също така е възможно съществуващите големи езикови модели да са свикнали с формата на въпросите и добавянето на примери да не носи добавена стойност. Допълнителни фактори, които влияят на ефективността на примерите са трудността на езика, знанията на модела, начина на избор и други.

Авторите експериментират и с многомодални модели. По време на създаване на статията обаче не съществуват многомодални модели, които да имат многоезиково разпознаване, затова използваният език е английски. Използва се Флан те5, въпреки че е само текстови модел. Той постига добри резултати. Към него подават само текстовата част на въпроса, а към многомодалните модели - Блип 2 и Инструкционен Блип, за които се използва Флан те5 за влагане, подават първото изображение, ако има повече от едно. Неочаквано е, че многомодалните модели не се представят значително по-добре от Флан те5. Всъщност, единствено Блип 2 постига по-висока точност, но едва с 0,76%. Анализ на грешките показва, че многомодалните модели се затрудняват при разбирането на сложни детайли като детайли за осите в математически въпроси и детайли за картата в географски въпроси. Въпреки че моделите Фромаж и Опън фламинго са специално тренирани, за да могат да работят с множество картинки, те не се представят по-добре и дори се представят по-зле от модели като Блип 2 и Инструкционен Блип, които работят само с една

картинка. Анализът показва, че разбиране дори и на една картинка е трудно за тези модели (Фромаж и Опън фламинго). Това показва, че претрениране с няколко изображения не води непосредствено до по-добро разбиране на многомодален вход и изграждане на причинно-следствени връзки между няколко снимки.

Разпределението на точността на моделите спрямо нивата на учениците е неравномерно. Учудващо, с увеличаване на нивото, авторите не наблюдават ясна тенденция за спад или увеличаване на точността. Най-добрите резултати се наблюдават за второ ниво, а най-ниските - за трето (най-сложното).

2.3.8 Екзамс

Наборът от данни Екзамс (от англ. *Exams*) е създаден от Хардалов и колеги [2, 3]. Фокусира се върху изграждането на набор от данни и оценяване на големи езикови модели върху него. Сравнява се също точността, която постигат те, когато отговарят на еднакви въпроси, преведени на различни езици (т.нар. паралелни въпроси). Експериментите включват фина настройка на многоезичния Бърт модел и модела Екселем робърта. Най-добрият модел е Екселем робърта, който е фино настроен чрез наборите от данни Рейс, Ейай 2 научни набори от данни на английски и на тренировъчното множество на Екзамс, постигайки средна точност от 42%. Авторите наблюдават, че въпросите от предмета „наука за природата“ затрудняват най-много моделите. Това се дължи главно на предметите химия и физика. Предметът информатика също затруднява моделите, т.к. изисква разбиране и създаване на програмен код, както и познаването и работенето с различни бройни системи.

2.3.9 Сравнителен анализ и изводи

Таблица 1 показва сравнителен анализ на експериментите, в разгледаните статии, и получените от тях резултати.

Име	Публи-кован (Година /Месец)	Мно-гомо-дал-ност	Най-добър модел	Втори най-добър модел	При-мери пома-гат	Текст + Кар-тинка по-добр о от текст	OCR & Описание на картинка помагат	Верига на разсъж-денията помага
VQA	2016/10	да	deeper LSTM Q + l2 norm VGGNet	LSTM Q + VGGNet	-	да	-	-
ScienceQA	2022/10	да	GPT-3	UnifiedQA	до 2 приме-ра	да	-	да
TQA	2017/07	да	Text + DPG	Text + Image | Text Only	-	да	-	-
SciBench	2023/07	не	GPT-4	GPT-3.5	не	-	-	да
MathVista	2023/10	да	GPT-4V	Bard + Google Lens	-	да	да	-
MMMU	2023/12	да	GPT-4V	Qwen-VL-PLUS	не	да	не	-
M3Exam	2023/11	да	GPT-3.5-Turbo	BLIP-2	не	не	не	-
Exams	2020/11	не	XLM-R (RACE + SciENs + Exams)	XLM-R (RACE + SciENs)	-	-	-	-

Таблица 1: Сравнителен анализ на подобни разработки – набори от данни и моделите, представили се най-добре върху тяхното тестово множество.

Обзорът на статиите показва, че:

1. Въпроси, успешното отговаряне на които изисква богата обща култура или използването на голям контекст от усвоен материал, са по-трудни от специфични за домейна въпроси.
2. Обединяването на числови представяния на многомодални данни в една структура (например едномерен вектор) допринася за по-голяма трудност, когато добавената им стойност е фокусирана само в една модалност. Внася се шум.
3. За цялостното оценяване на модел е необходимо разнообразие от типове въпроси, които варират в количеството разсъждения, нужни за достигането до правилен отговор, и количеството контекст, необходимо за формирането на тези разсъждения.
4. Тенденцията е модели, тренирани с по-голямо количество данни, да се представят по-добре от модели, тренирани с по-малко количество данни, независимо от контекста и специфичността.
5. Използването на стратегии за инструктиране, които включват извличане на веригата от разсъждения, създадена от модела водят до по-добри резултати.
6. Наблюдава се тенденция моделът Генеративен претрениран модел 4-B да постига по-добри резултати, когато е сравнен с разговорния агент Бард. Това сравнение трябва да се изследва по-подробно, т.к. не е хомогенно (модел се сравнява с разговорен агент).

7. Докато средно точността на големи многомодални модели като Генеративен претрениран модел 4-В не надминава тази на човек, тези модели се справят по-добре от човек в задачи, свързани с детайлно познаване на специфичен домейн.
8. Макар и Генеративен претрениран модел 4-В да постига в общия случай най-добри резултати, все още няма достатъчно изследвания спрямо него и модела Джеминай ултра (основния му конкурент).
9. Повечето статии анализират грешките, допуснати от Генеративен претрениран модел 3,5 турбо, но не и тези, допуснати от разговорния агент Бард.

2.4 Многомодални и многоезикови модели

2.4.1 Генеративен претрениран модел 3,5 турбо

Големият езиков модел е разработен от компанията Опънейай и е в основата на разговорния агент Чатджипити. Доста често в литературата наименованията Генеративен претрениран модел 3,5 турбо и Чатджипити се използват взаимозаменяемо.

Трениран е посредством техниката обучение чрез поощряване, използвайки обратна връзка от човек (от англ. *Reinforcement Learning from Human Feedback (RLHF)*) [29]. Тя позволява ефикасно и ефективно трениране на генеративни модели, които по дефиниция се нуждаят от много голямо количество данни, за да постигнат задоволително представяне. Целта е да се създаде система за контролиране и насочване на изходите от модела, която да го специализира в провеждането на дискусия с човек, без да е необходимо човек ръчно да проверява генерираните изходи. Авторите на Чатджипити прилагат техниката в три стъпки, визуално представени на Фигура 1. Стъпките са:

1) Ръчно създаване на анотирани данни тип инструкции. Това се изразява в случаен избор на инструкция от набор от данни, съдържащ инструкции, симулиращи комуникация между две лица. Човек след това написва какъв би бил идеалният изход от генеративен модел и това се превръща в част от данните, които се използват, за създаване на първонална стратегия, която моделът впоследствие оптимизира (от англ. *supervised policy*).
2) От същия набор от данни с инструкции се избират случайни наблюдения и се подават на модела. От модела се изисква да върне малък брой различни отговори. Те се ранкират от човек спрямо добавената им стойност и уместност. Ранкираните данни се използват за тренирането на нов модел, специализиращ се в предоставянето на качествена оценка множество отговори (от англ. *reward model*).
3) Оптимизиране на стратегията от стъпка 1) с цел максимизиране на оценките, давани от модела, създаден в стъпка 2), използвайки алгоритъма оптимизиране на проксимална политика (от англ.

Proximal Policy Optimization [30]). По този начин човешкото знание, в дестилиран вид, се влага в модела, който оптимизира стратегията.

Фигура 1: Визуално представяне на техниката обучение чрез поощряване, използвайки обратна връзка от човек. Източник: [29].

Фамилията от модели Джипити 3,5 е тренирана върху разнообразни данни, включващи текст и в частност програмен код, събирани до четвъртото тримесечие на 2021 [31]. Моделите в тази фамилия са:

1. code-davinci-002, който се спекулира, че е в основата на разработения от Майкрософт продукт Гитхъб Копайлът [32, 33];
2. text-davinci-002, който е Инструкционен Генеративен претрениран модел [34], базиран на code-davinci-002. Това внася интересна интерпретация на неговите възможности: новите модели, работещи с текст са надградили и са фино настроена версия на модели, специализирани за работа с програмен код;
3. text-davinci-003, който се споменава, че е подобрение над text-davinci-003, но не се посочва в какви аспекти.

Използването на техниката обучение чрез поощряване с обратна връзка от човек при за създаването на разговорния агент, позволява добавянето на настройка, която служи, за да определи контекста, в който да бъде генериран отговорът. Този параметър се нарича „роля“ и се подава заедно с инструкцията. Спрямо него се получават отговори в различен стил – например, ако параметърът „роля“ има стойност “ти си учител по програмиране” и инструкцията е “направи времева и пространствена оценка на следния код”, се очаква моделът да върне искания анализ в академичен стил. Този параметър може да бъде зададен експлицитно при използване на програмния интерфейс за подаване на заявки или може да бъде имплицитно подаден при използване на уеб приложението като се постави като префикс на инструкцията.

Големият потенциал на модела води и до използването му за неправилни и опасни цели. Налични са множество ресурси за заобикаляне на поставените ограничения за безопасността на отговорите [35, 36, 37].

2.4.2 Генеративен претрениран модел 4

Генеративен претрениран модел 4 е голям многомодален модел, който може да приема текст и картинки и да връща текстови отговор [38]. С излизането си бързо нашумява начинът, по който е тестван, и резултатите, които постига. Оценен е върху контролни, давани на ученици и студенти и анализ показва, че Генеративен претрениран модел 4 има възможност да премине успешно почти всички от тях [39]. В частност той се класира сред 10% от най-добрите явяващи се изпит по за адвокат [39].

Малкото количество споделена информация за разработването и характеристиките на модела води до недоволство от общността, фокусирана в развиването на продукти с отворен код [41]. В техническата спецификация [40] се споменава, че поради наличието на много конкуренти и експлоатирането на възможностите на модела за нарушаване на безопасността, не се споделя информация за архитектурата на модела (включ. и големина на модела), използван хардуер, време за трениране, набиране и подготвяне на данни, начин на трениране и подобни. Споделено е, че моделът е трениран да предсказва следващата дума в документ, чрез използване на публично достъпни данни и такива, за които Опънейай има лиценз. Това прави трудно извършването на добър сравнителен анализ и опити за подобряване на вече направеното. В същия момент изглежда е част от тенденция за постепенното преминаване на големи интелектуални центрове от извършване на проучване към продаване на продукти. Една от главните критики на общността е, че направените изводи в следствие на създаването на подобни модели, си остават в кръга на техните създатели и не се споделят. Това контрастира с принципа за независимо репродуциране на резултатите на подобни проучвания [42].

Голяма част от техническата спецификация обръща внимание на притесненията за безопасност на модела и действията, които са били предприети, да поставянето на граници на приемливото, които моделът не може да преминава. Основните действия включват премахване на нецензурираното съдържание от данните за трениране и извършване на допълнително фино настройване на държанието на модела, чрез техниката *RLHF*, за отказване на даване на отговор при подтикващи към неправомерни действия инструкции [39].

Анализ на качеството на резултатите показва, че възможностите на модела идват главно от процеса по претрениране - използването на техниката *RLHF* не подобрява представянето на изпити, давани на ученици, а напротив - без допълнителни усилия, точността намалява [39]. Въпреки това нагласянето на модела да отговоря нормално и приемливо за човек идва от именно в резултат на прилагането на тази техника, т.к. базовият модел изисква нагласяне с помощта на инструкциите дори и само за да разбере, че трябва да отговори на въпросите.

Това означава, че добавена стойност на техниката *RLHF* се изчерпва с превръщането на модела в разговорен агент-асистент. Самите възможности за „разсъждаване“ и „креативност“ идват от тренирането с голямо количество данни. Тези твърдения покават, че няма как моделът да научи нови умения, използвайки единствено обратна връзка от човек. От друга страна обаче по този начин може да се намали количеството токени, отделени за обясняване на контекста и нужни за нагласяне на инструкцията (от англ. *prompt engineering*) при поставяне на задачи.

В техническата документация се прави интересно уточнение – осреднената грешка и количеството ресурси за трениране на Генеративен претрениран модел 4 могат да се предвидят, използвайки вече наличната информация за предишни модели. Фигура 2 показва загубата при предсказване на следващата дума от парче код. За осите се използва логаритмична скала. Зелената точка представлява загубата на Генеративен претрениран модел 4, а първата сива точка от дясно наляво представлява модел, който е изисквал 10,000 пъти по-малко ресурси за трениране. Не са уточнени моделите, които са използвани.

Фигура 2: Предвиждане на загубата на Генеративен претрениран модел 4, чрез напасване на крива върху предишни (известни) резултати. Показва, че загубата може да бъде предвидена изключително точно, следвайки тенденцията на предишни модели. Източник: [40].

На 27.06.2022 г. Опънейай обявява състезание с отворен достъп на име *Inverse Scaling Prize* [43]. Задачата, която се поставя, е предаването на набори от данни, които са така конструирани, че по-големите модели (т.е. моделите с повече параметри) постигат по-ниска точност, отколкото по-малки модели (с по-малко параметри). Едно от множествата, предадени в това състезание, представя задачата за ретроспективното пренебрегване (от англ. *Hindsight*

Neglect) [44]. Събрани са данни в контекста на теорията на игрите. Всяко наблюдение създава статистически контекст на игра, в който:

1) Вероятността за загуба е минимална и очакваната печалба е висока при много опити;
2) Играч избира да играе, но губи;
3) Въпросът, поставен на модела, е дали играчът е постъпил правилно, избирайки да играе.

Пример за подобна ситуация е:

1) Възможно е Джон да играе игра, в която да загуби 900 лева с 5% вероятност и да спечели 5 лева с 95% вероятност.
2) Джон избира да играе играта, но губи.
3) "Изборът на Джон правилен ли е?"

Ако моделът отговори с "Да", то се предполага, че той разбира правилния избор на Джон, от гледна точка на вероятностите и очакваната печалба след повтаряне на опита много пъти. Фигура 3 показва, че с увеличаването на броя параметри моделите постигат все по-ниска точност на тази задача. Генеративен претрениран модел 4 обаче е първият модел, който обръща тази тенденция, постигайки почти 100% точност.

Фигура 3: Зависимост между точност и модел върху задачата за ретроспективно пренебрегване. Генеративен претрениран модел 4 обръща тенденцията за по-лошо представяне, постигайки почти 100% точност. Източник: [40].

Техническата спецификация на Генеративен претрениран модел 4 завършва с уточнението, че моделът е използван и за формулиране на твърденията, форматиране и стилизиране на текста на самата спецификация.

Въпреки че на пръв поглед възможностите на Генеративен претрениран модел 4 са много големи и се представя изключително добре на голямо разнообразие от задачи, някои резултати могат да бъдат подвеждащи. Например, в [45] се твърди, че моделът решава правилно всички наблюдения от тестово множество, съставено само от текстови въпроси, давани на студенти от

Масачузетския технологичен институт (от англ. *Massachusetts Institute of Technology (MIT)*). Използваният набор от данни се състои от общо 4,550 въпроса от дисциплини като математика, компютърни наука и електроинженерство. Авторите оценяват Генеративен претрениран модел 3,5 турбо, кой успява да реши успешно едва една трета от изпита, но Генеративен претрениран модел 4, по данни на авторите, решава правилно всички въпроси. Методът на провеждането на експериментите обаче и начинът на оценяването на получените резултати се приемат с голяма критика дори и от студенти на Масачузетския технологичен институт [46]. Те идентифицират, че 4% от набора данни се състои от нерешими задачи – такива, които в оригиналния вариант включват текст и картинка (нужна за даване на правилен отговор), но картинката е била премахната, въпроси, поставени без контекст (без който не може да се даде адекватен отговор), въпроси, които всъщност изискват предлагането на тема за курсов проект и подобни. В допълнение, оценяването на резултатите е автоматично отново чрез използването на Генеративен претрениран модел 4, което внася предубеденост и шум в качеството на оценката. Към момента статията е официално оттеглена от платформата arXiv.

2.4.3 Разговорен агент Бард и модели от фамилията Джеминай

Бард е многомодален и многоезичен разговорен агент, поддържан от софтуерната компания Гугъл. Той се предоставя за свободно използване през март 2023 и е широко приеман да е пряк конкурент на създадения по-рано разговорен агент Чатджипити. Докато Чатджипити е базиран на архитектурата Генеративен претрениран модел 3,5 турбо, Бард в своето развитие използва различни архитектури - първоначално това е основният модел *Language Models for Dialog Applications* (LaMDA) [47], след това *Pathways Language Model 2* (PaLM 2) [48], а наскоро и най-новия голям многомодален модел на Гугъл – *Джеминай* [49]. Първоначално Бард, 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. Бард се превръща в първия свободно достъпен многоезичен и многомодален модел, който може приема картинки и текст като вход [50]. По това време съществуват и други модели, за които се твърди, че могат да работят с картинки, но те все още не са публични. Един от тези модели е Генеративен претрениран модел 4.

Моделите, които използва Бард, постепенно разширяват неговите възможности - моделът LaMDA се фокусира върху провеждането и поддържането на дълги кореспонденции, по този начин поставяйки солидна основа за функционалността му на разговорен агент. Моделът PaLM 2 надгражда и подобрява възможностите за генериране на текст (и по-специфично, създаването на код и поправянето на грешки в код) и превод между различни езици. Моделите от семейството Джеминай позволяват работа с картинки. В допълнение, разговорният агент има достъп до Интернет и може да

бъде използван, за да резюмира и анализира нова информация. Основният конкурент на Бард – Чатджипити, е ограничен да работи с информация само до определен момент във времето – към февруари 2024 г. това е януари 2023 г.

През декември 2023 дъщерната компания на Гугъл - Диипмайнд, обявява нова фамилия от големи многомодални модели под името Джеминай: Джеминай Нано (най-малкият), Джеминай Про и Джеминай ултра (най-големият). Трите обявени модела се различат по размер и предназначение.

Джеминай Нано е предназначен да се използва в устройства, изискващи ограничени ресурси, например, смартфони. Спрямо наличната памет в устройството, може да използват две вариации на Нано - с размер 1,8 милиарда параметъра или с 3,25 милиарда параметъра. Те са получени чрез прилагане на техниката дестилиране върху двата по-големи модела от фамилията. Моделът с 1,8 милиарда параметъра би използвал около гигабайт памет, което е поносимо за повечето съвременни устройства и показва потенциала на този вид модели. Той е вграден в смартфона Pixel 8 Pro [53].

Моделът Джеминай Про е оптимизиран за балансирана производителност и закъснение при връщане на резултати. Може да обработва различни видове многомодални задачи и да разсъждава над проблеми, изискващи богата обща култура [51].

Моделът Джеминай ултра е описан като подобрена версия на Джеминай Про, постигайки най-добрите резултати в проведените от екипа експерименти. Не се посочват директни сравнения, водещи до идентифицирането на отличаващи характеристики между двата модела. Според авторите хардуерната основа, на която работят те, оказва влияние на тяхното представяне - резултатите, които моделите постигат, са получени, използвайки сървър, работещ с TPU ускорители.

В техническото описание на моделите [49] се споделя силно ограничено описание на архитектурите им. Моделите са базирани на невронна мрежа тип Трансформър, която може да приема текст, картинки, видео и аудио и да връща комбинация от картинки и текст, в частност и програмен код. Въпреки тези възможности, към момента е възможно да се работи само с текст и картинки. Предоставена е и силно диаграма, представяща архитектурата на моделите, но тя е на високо ниво и не представя използваните подходи и по какво се различават те от вече съществуващите. Начинът на трениране е описан и е поставен фокус върху различните проблеми, които екипът е имал и как ги е преодолял. Не се споделя колко време са се тренирали. Моделите могат да работят с размер на контекст до 32,000 токена [54], използвайки SentencePiece токенизатор. Това е ниско в сравнение с размера на контекста от Генеративен претрениран модел 4: 128k [55].

Не се споделят детайли относно набора от данни. Съставен е от уеб документи, книги, и програмен код и съдържа картики, текст и видео (което е разбито до серия от картинки). Авторите отбелязват, че тренирането върху картинки, е довело до по-висока точност при работа с текст.

Компанията обявява моделите посредством клип, който демонстрира способностите на Джеминай ултра. Клипът е приет с критики, т.к. въпреки че моделът не може да работи с видео, клипът показва това. То реално е серия от картинки и на някои места дори се подават и подсказки. Например, има една част с разпознаване на играта морски шах от серия от действия на ръка. Това обаче са три снимки заедно с подсказка, че става въпрос за игра. Също така моделът не връща аудио, а текст, който след това е прочетен от човек.

Диипмайнд не предоставя размерите на Джеминай Про и Джеминай ултра. Това, че размерите на моделите Нано се показват е по-скоро очаквано, т.к. вграждането им в устройства ще позволи тяхното разглеждане.

Сравнението с основния конкурент на серията модели - Генеративен претрениран модел 4, се приема с критики. На фигура 4 се показва, че върху наборът от данни MMLU [52], Джеминай Ултра постига точност от 90% спрямо 86.4%, постигната GPT4, и 89.8% - точността, постигната от човек.

Фигура 4: Точността на човек , сравнена с точността на модела Джеминай Ултра (90%), сравнена с точността на Генеративен претрениран модел 4 (86.4%) върху набора от данни MMLU (89.8%). Източник: [51].

Предоставените резултати обаче са заблуждаващи, т.к. моделите не са сравнени при еднакви условия. Джеминай Ултра постига 90% точност, използвайки техниката извличане на верига на разсъждения (от англ. *chain of thought reasoning*), докато Генеративен претрениран модел 4 е оценен с предварително подаване на 5 примера, техника, за която разгледаната литература сочи, че няма статистически значима добавена стойност. В допълнение, извличането на верига на разсъждения е доста по-сложен и по-интензивен начин за задаване на уточнителни въпроси от предоставянето на примери. Това внася предубеденост, т.к. същината на двете техники за оценяване предполага, че първата ще доведе до по-добри резултати, независимо от модела, който я използва.

На фигура 5 са показани резултатите, които Джеминай ултра постига при предварително подаване на 5 примера и те са по-лоши от тези на Генеративен претрениран модел 4 - 83,7% и 86,4%.

	Gemini Ultra	Gemini Pro	GPT-4
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)

Фигура 5: На първи ред: резултатите на Джеминай ултра (90,04%) и Генеративен претрениран модел 4 (87,29%) при извличане на верига от разсъждения. На втори ред: резултатите на Джеминай ултра (83,7%) и Генеративен претрениран модел 4 (86,4%) при предварително подаване на 5 примера. Източник: [49].

Експериментите с Генеративен претрениран модел 4 с извличане на верига от разсъждения постигат малко по-лоши резултати от тези на Джеминай. Получената точност от 87,29% не е официално потвърдена от Опънейай, а е получена от екип на Диипмайнд, използвайки платения програмен интерфейс на модела. Трудно е да се аргументира кой от двата модела е по-добър – представянето им е сходно и разликите в точностите им са малки.

На 08.02.2024 г. Сиси Хсиао, вицепрезидент и генерален мениджър на продуктите Бард и Гугъл Асистант, обявява, че Бард се ребрандира на Джеминай [56]. Разговорният агент използва модела Джеминай Про.

2.4.4 Лава

Моделът Лава (от англ. *LLaVA*) има две версии - от 17 април 2023 – Лава 1.0 [57] и от 5 октомври 2023 – Лава 1.5 [58]. В настоящата дипломна работа при сравненията на моделите се използва втората версия. Те могат да приемат текст и картинка и връщат текст. Архитектурата им, показана на фигура 6, се състои от две части, трансформиращи отделните модалности към една репрезентация. За влагане на картинката се използва моделът Vision Transformer (ViT-L/336px), а за влагане на потребителската инструкция се прилагат последователно токенизатор и слой за влагане на токените. Получените влагания на картинката и потребителската инструкция се обединяват подават за интерпретация към голям езиков модел - Vicuna v1.5 13B.

Фигура 6: Архитектура на модела LLaVa 1.5. Източник: [58].

Моделът за вграждане на изображението е ViT-L/336px, който е използван и в модела CLIP, създаден от Опънейай [59, 60]. Той разделя изображението на множество от по-малки изображения с размерност 336x336, които след това се вграждат чрез линеен слой. Към всеки пиксел се свързва и стойност, която маркира пространствената информация, за да не бъде тя изгубена при процеса на трансформиране към едномерен вектор. Векторите, получени след добавянето на позиционния контекст, се подават на 24 слоя енкодери от Трансформър архитектурата.

LLaVa 1.0 използва нелинеен прожекционен слой, за да обединяване на пространството на извлечените характеристики от текста и картинката. По този начин той придобива ролята на „преводач“ между двете модалности. Прожекционният слой представлява двунивов персептрон, чиито тегла се настройват по време на трениране.

Наборът от данни, използван за трениране на модела, е изцяло синтетичен, многомодален и поставя на фокус върху използването на наблюдения, които изграждат умението за следване на инструкции. Състои се от 158,000 примера, разпределени в три групи - нормални разговори, детайлни описания на обекти и развиване на комплексни аргументи. Всяко наблюдение има картинка, която е и фокусът на инструкцията. Техниката, използвана при създаването на набора от данни, се основава на симулация на кореспонденция между разговорен агент и човек по подадена картинка и нейно описание. По-точно, авторите подават картинка, пет изречения, описващи я, и инструкция на моделът Генеративен претрениран модел 4-B, на който е поставена задачата

да създаде примерна кореспонденция между разговорен агент и човек, питащ въпроси за картинката. Зададено е явно в инструкцията към Генеративен претрениран модел 4-B, че трябва да се създаде въпроси с различна насоченост – например, към типа обекти, техният брой, какво правят те, къде се намират, какви са им относителните позиции и подобни. Също така се изисква и създаването на въпроси, насочени към физическата структура на обектите и описание на сцените, развиващи се в изображението. С цел развиване на умението за аргументация се създават и примери за стъпки на разсъждаване при излагане на отговорите на въпросите. Избягва се включването на въпроси, отговорите на които не са сигурни, конкретни и точни.

Във версия 1.5 наборът от данни е разширен чрез добавяне на въпроси от VQA. LLaVA е трениран в две фази. В първата фаза теглата на вграждащия визуалната информация модел и големия езиков модел не се променят. Тренира се само прожекционната матрица с цел да се научи да обединява по-начин, позволяващ достигане на минимална загуба, извлечените характеристики от изображението и текста. По същество, това е стъпка за създаване на най-подходящ за големия езиков модел визуален токенизатор.

Във втората стъпка се замразяват само теглата на модела, обработващ картинката. Обединяващата двете модалности матрица продължава да се тренира и заедно с нея се тренира и големия езиков модел.

Използвайки 8 графични карти NVIDIA A100, авторите успяват да създадат модел, постигащ по-добра точност от Куен виел, за 6 часа прекарани в трениране на първа фаза и 20 часа във втора фаза.

2.4.5 Куен виел

Куен виел [28, 64] е серия от големи визуално-текстови модели, създадени от Алибаба Груп. Приемат картинки и текст и връщат текст. Макар и многоезични, те поддържат единствено китайски и английски. Моделите се отличават по това, че са едни от малкото, които могат да работят с повече от една картинка при подаване на инструкция. Основният модел е Куен елем, чрез който след трениране върху многоезичен и многомодален набор от данни авторите създават Куен виел и Куен виел чат. Отделя се специално внимание на възможността на моделите да се представят добре освен в типичните задачи, фокусирани върху работа с картинки - даване на описание и отговаряне на въпроси, но и в задачата за локализиране на най-релевантния обект или регион в картинка на база заявка на естествен език [61].

Моделът Куен виел е с общ размер 9,6 милиарда параметъра като архитектурата му е аналогична на тази на Лава. Състои се от три компонента - визуален енкодер с размер 1,9 милиарда параметъра, голям езиков модел с размер 7,7 милиарда параметъра и адаптер за транслация между двете модалности с размер 80 милиона параметъра. Енкодерът използва архитектурата Vision Transformer, която се инициализира с теглата от модела ViT-bigG, използван в Openclip [62].

Адаптерът компресира последователността от извлечените визуални характеристики до едномерен вектор с 256 елемента. Реализиран е чрез еднослоен модул с крос-атеншън, инициализиран със случайни тегла. Модулът използва група от вграждания като вектори на запитването и характеристиките на изображението като ключове. За запазване на позиционната информация се добавят и двумерни позиционни вграждания. Компресираните характеристики с дължина 256 след това се подават на големия текстови модел.

Процесът по трениране на Куен виел се състои от три фази: две фази за претрениране и финална фаза фино нагласяне чрез инструкции.

В първата фаза се използват 1,4 милиарда двойки от картинки-текст за създаване на умение за описване на изображение. Данните са събрани от Интернет като 77,3% от тях са на английски език, а останалите 22,7% са на китайски език. Поради тази специфика авторите посочват, че качеството на анотациите има голяма вариативност в зависимост от езика. В тази фаза на трениране теглата на езиковия модел са замразени и се тренират само визуалният енкодер и адаптерът. Изображенията се оразмеряват до 224 x 224. Целевата функция е да се намали кросентропията между текстовите токени.

Във втората фаза се добавят данни с високо качество, с висока детайлност на анотациите, с по-висока входна резолюция и със смесени множества от текст и картинки. Чрез тези данни моделът е трениран едновременно върху 7 задачи:

1. поставяне на описание на картинката;
2. отговаряне на въпроси, насочени към картинката;
3. намиране на най-важните участъци в картинката;
4. намиране на най-важните участъци в поставено описание;
5. броене на обекти;
6. разпознаване на текст в картинката;
7. предсказване на следваща дума.

Резолюцията на визуалния енкодер е разширена до 448 x 448, намалявайки количеството загубена информация при down-sampling. Прозорецът на внимание и глобалното внимание също са с по-голям размер в трансформъра. Езиковият модел вече не е замразен и така се тренират всички части на архитектурата.

Последната фаза е фино настройване на модела за задачата на конструиране и поддържане на диалог. В нея се замразяват теглата на визуалния енкодер и се тренират езиковият модел и адаптерът. Резултат от този процес е моделът Куен виел чат. Данните за трениране се разширяват с примери, генерирани от самозапитване на езиковия модел. Авторите отбелязват, че самозапитването в повечето случаи симулира диалог, около едно изображение и обсъжданите въпроси за него нямат голяма вариативност поради трудностите, които моделът среща при разбиране съдържанието на изображението. За да внесат по-голямо разнообразие, те създават допълнително множество от данни, симулиращо диалог върху поне две изображения, чрез ръчна анотация. Общият размер на инструкционните данни е 350 хиляди наблюдения. В допълнение, по

време на трениране многомодалните данни се обединяват с такива от текстови диалози, за да се утвърди универсалността на възможностите за диалог. Така те обединяват уменията за локализация, разбиране и работа с множество картинки в модела Куен виел.

Анализ на резултатите показва, че моделът може успешно да прилага наученото не само към различни типове задачи, от тези, за които е трениран, но и към различни езици.

2.4.6 Сравняване на моделите

Таблица 2 показва сравнителен анализ на най-добрите многомодални модели към текущия момент, които се използват за целите на дипломната работа.

Макар и да няма официално потвърдена информация за размерите на моделите от комапниите Опънейай и Гугъл, съществуват спекулации от техни потребители. Например, една от най-популярните във връзка с модела Генеративен претрениран модел 3,5 турбо е, че неговият размер е 20 милиарда параметъра [63]. Това е почти 10 пъти по-малко параметри от GPT-3 (175 милиарда). За Генеративен претрениран модел 4 се спекулира, че има 1,76 трилиона параметъра и че всъщност се състои от 8 по-малки модела-експерти в различни области, близки до GPT-3, чиито „мнения“ се претеглят и така формират крайния отговор [65, 66, 67].

Модел	**Дата**	**Компа-ния**	**Размер / Брой парамет-ри**	**Месец, до който има данни**	**Размер на използвани данни**	**Мно-го-мо-дал-ност**	**Брой ези-ци**	**Размер на кон-текст-ния прозорец**
gpt-3.5-turbo-1106	ноември 2023	OpenAI	-	септември 2021	-	не	>2	16385
gpt-4-1106-preview	ноември 2023	OpenAI	-	април 2023	-	не	>2	128000
gpt-4-vision-preview	ноември 2023	OpenAI	-	април 2023	-	да	>2	128000
gemini-pro	декември 2023	Google Deepmind	-	началото на 2023	-	не	>2	32000
gemini-pro-vision	декември 2023	Google Deepmind	-	началото на 2023	-	да	>2	12288
LLaVa	декември 2023	-	13 млрд.	-	1,2 млн.	да	1	4096
Qwen-7B	октомври 2023	Alibaba Group	9,6 млрд.	-	1 476,8 млн.	да	2	32000

Таблица 2: Сравнителен анализ на използваните в дипломната работа модели.

Таблица 3 показва сравнение на цените за използване на платените модели. Наблюдава се голяма разлика в предлаганите цени от Гугъл и тези от Опънейай.

Модел	Цена за 1000 токена вход текст (долари)	Цена за 1000 токена вход картинка (долари)	Цена за 1000 токена изход (долари)
gpt-3.5-turbo-1106	0,00100	-	0,00200
gpt-4-1106-preview	0,01000	-	0,03000
gpt-4-vision-preview	0,01000	-	0,03000
gemini-pro	0,00025	-	0,00050
gemini-pro-vision	0,00025	0,00250	0,00050

Таблица 3: Цени за използване на текущите най-добри големи многомодални модели. Източници: [68, 69].

Обзорът на техническите документации на избраните модели показва, че:

1. Има голяма разлика в качеството на връщаните резултати от моделите, разработени от Гугъл и Опънейай, и останалите.
2. Най-добрите модели – Генеративен претрениран модел 4-В и Джеминай ултра, вероятно са много по-големи от всички други.
3. Наблюдава се тенденция за скриване на имплементацията на големите многомодални модели и детайлите за начина, по които те са тренирани. Това е резултат от постепенното преминаване на големите цетрове за проучване към продуктови компании.
4. Макар и подробно оценени в съответните технически документации, все още е трудно да бъдат посочени отличителни характеристики между моделите Генеративен претрениран модел 4-В и Джеминай ултра.
5. Става все по-трудно да се тренират големи многомодални модели без разполагане с голяма инфраструктура и средства. Тенденцията е да се търси спонсор (обикновено голяма продуктова компания), който да финансира процеса.
6. Големите модели успяват да придобият голямо количество умения и знания главно заради количеството и качеството на данните, с които се тренират. Използването на допълнителни техники, като обучение чрез поощряване, използвайки обратна връзка от човек, носи добавена стойност дотолкова, доколкото предразполага модела за общуване с хора.

3 Набор от данни Екзамс-В

Наборът от данни Екзамс-В (на англ. Exams-V) е разширена версия на работата на Хардалов и колеги - Екзамс [2]. Аз участвах пряко в неговото разработване и последващо използване за оценяване на моделите в сътрудничество с колеги от Университета „Мохамед бин Зайед по Изкуствен интелект“ (MBZUAI). Като част от разработването на Екзамс-В, бе предадена за рецензия статията *EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models* в *Annual Meeting of the Association for Computational Linguistics Rolling Review (ACL ARR)*. Наборът от данни ще бъде публично достъпен след публикуването на статията. Подробно разглеждане на набора от данни е направено в написания от мен преддипломен проект.

В Екзамс са събрани тестове, давани на ученици от различни нива на обучение на различни езици. От всички въпроси са оставени само тези, които са текстови и със затворен отговор. Разширението включва добавянето на нова модалност - въпроси с картинки. Общо въпросите са 20932 от 11 езика и 20 учебни предмета. Картинката може да е допълнителна част от въпроса, т.е. той да се отнася за картинката, но може и самите отговори да са (част от) картинки. Примери за такива въпроси са показани на фигури 7 и 8.

Фигура 7: Примери за въпроси, отнасящи се до едно изображение.

Фигура 8: Примери за въпроси, в които отговорите са изображения.

Тези два типа въпроси, макар и подобни, се различават по важната част на броя изображения. В допълнение на това, че моделът получава изображение като вход, то самото може да съдържа в себе си други изображения. Очаква се, че когато въпросът се отнася за едно изображение, модел би го разпознал и работил с него по-лесно. Когато изображенията са повече в общия случай е необходимо да се направи сравнителен анализ на тях - по какво си приличат, по какво се различават, за да се достигне до верен отговор.

Процесът по добавяне на картинки е направен чрез свободни за използвани пакети на Python [20, 21, 22]. Разнообразието от картинки е голямо. В частност, картинките са разделени на четири типа на база съдържанието им: химични символи/формули/структури, таблици, фигури или графики. Такива примери са показани на фигури 9, 10, 11, 12.

Фигура 9: Пример за въпрос, съдържащ химични структури.

8 ما التركيب الجيني للآباء حسب مربع بانيت التالي؟

	---------	---------

A *Bb* و *Bb*

B *Bb* و *bb*

C *BB* و *bb*

D *BB* و *Bb*

Фигура 10: Пример за въпрос, съдържащ таблица.

23. Pozorno promotrite povijesni zemljovid.

Kakav je bio status zemalja označenih na povijesnome zemljovidu smeđom bojom prema Napoleonovu Carstvu?

A. Bile su neutralne prema Napoleonovu Carstvu.
B. Bile su protivnice Napoleonova Carstva.
C. Bile su vazalne zemlje Napoleonova Carstva.
D. Bile su saveznice Napoleonova Carstva.

A. ☐
B. ☐
C. ☐
D. ☐

Фигура 11: Пример за въпрос, изискващ осмисляне на фигура (и по точно, карта).

17. От графиката на променливото напрежение определете неговата ефективна стойност.

А) $\frac{15}{\sqrt{2}}$ V
Б) $15\sqrt{2}$ V
В) 15 V
Г) 0 V

Фигура 12: Пример за въпрос, включващ графика.

Различните многомодални въпроси водят със себе си и различни трудности. Например, въпросът, показан на фигура 9, има комплексни химически структури. Макар и да не съдържа формално картинка разпознаването на тези структури и разбирането на техния смисъл е ключът за решаването на задачата. Възможно е и картинка е богата на съдържание. Това е така във фигура 10, където въпросът съдържа карта. Правилно отговаряне на такива въпроси би изисквало моделът на има от една страна богата обща култура, за да разпознае обектите по картата, в частност да я допълва с необходим контекст (напр. исторически период, който е представен), и от друга страна да има възможност за фина обработка на цялата информация, представена чрез картата. Фигура 11 показва че не винаги за "картинка" може да счита всяка област, състояща се от пиксели с различни от черно и бяло цветове. Таблиците също могат да считат за различен тип многомодалност от текста, т.к. имат определена структура и смисъл, който се движи по колони и редове. Графиките, пример за които е показан на фигура 12, са подобни на таблиците с добавената особеност за пространствено представяне на информацията.

Друга интересна вариация е картинката сама по себе си да не съдържа отговора и да изисква абстрактно мислене и разбиране на импликациите, които се създават от нея. Пример за въпрос, съдържащ такъв тип картинка, е показан на фигура 13. Този тип въпроси не са тривиални, т.к. просто разбиране на съдържанието на картинката без разполагането с общата култура за нейното осмисляне и "прочитане" на нейното послание, няма да е достатъчно.
За правилното отговаряне са нужни мултидисциплинарни знания и изграждане на причинно-следствени връзки.

Форматът на въпросите не винаги е чист. Понякога има припокриване на част от въпроса с друга част, която не допринася към правилното отговаряне на въпроса. Има също и присъствие на части, несвързани с въпроса. Това от една страна е нормално, но от друга страна може да затруднява моделите, т.к. е възможно да се заблудят, анализирайки несъществени части.

38. Pozorno pogledajte slikovni prilog.

Koji je događaj iz europske povijesti u 20. stoljeću prikazan na slikovnome prilogu?

A. gubitak Istočne Pruske nakon Prvoga svjetskog rata
B. završna sovjetska ofenziva na Berlin 1945. godine
C. sovjetska blokada zapadnoga Berlina
D. primjena Marshallova plana

A. ☐
B. ☐
C. ☐
D. ☐

Фигура 13: Въпрос, включващ картинка, провокираща абстрактно мислене.

Статистики за разпределението на въпросите от различните езици спрямо и нива на обучение спрямо броя предмети и модалността на въпросите са представи на таблица 4. От първоначалния набор от езици бяха премахнати испански, словашки, руски и румънски, т.к. за тях разнообразието между въпросите с картинки и тези само с текст не бе достатъчно голямо - например, за руски език всички въпроси бяха без картинки.

В крайното множество данни въпросите с картинка са 23,9%. За някои езици има по-малко въпроси с картинки, което се дължи на подбраните предмети от съответните езици. Предметите представени на таблиците са агрегиран/групиран вариант на имената на предметите, използвани в съответната учебна система. В резултат на това за полски език има един агрегиран предмет - “професионални”, който обединява 55 предмета, свързани с професионално градинарство, аграрни дейности, управление на тежкотоварни машини, инсталация и поддържане на компютърни мрежи и подобни. Това е направено аналогично на [2] и е уместно, т.к. събраните изпити са от профилирани дисциплини, свързани с развиването на професионални знания като механика, градинарство, поддръжка на компютърни устройства и развиване на дейности за помагане на инвалиди. Характерна черта за тези изпити е, че те са дълги (т.е. съдържат много въпроси), но са малко като количество изпити.

Няколко езика съдържат паралелни въпроси - т.е. преведени въпроси на различен език от официалния на страна, в която се изучава предметът. Това са

езиците - хърватски (който съдържа въпроси на италиански, сръбски, словашки и румънски), арабски (който съдържа въпроси на английски). Таблица 5 показва как са разпределени те. Забелязва се, че отношението на въпросите, включващи само с текст, спрямо тези, съдържащи и картинка, следва в повечето случай тенденция да бъде едно към пет. Това е малко над общата тенденция (едно към три), което означава, че разпределението на въпроси с картинки в различните езици не е равномерно, и случайно избран въпрос е около четири пъти по-вероятно да бъде въпрос без картинки, отколкото да е въпрос с картинки. Най-много въпроси (3969) и най-много предмети (13) има от хърватски език. Забелязва се също, че въпросите са подбрани от разнообразни езикови групи, обхващащи германските, славянските и китайско-тибетските езикови семейства. Разнообразието, което добавя арабският език е много важно, т.к. той е единственият включен език, който се изписва и чете от дясно наляво.

Език	Нива на обучение	Брой предмети	Брой въпроси	Въпроси с текст и картинка	Въпроси само с текст
арабски	4-12	6	823	117	706
български	4, 12	4	2132	435	1697
китайски	8-12	6	2635	1991	644
хърватски	12	6	3969	700	3269
английски	11, 12	4	724	181	543
френски	12	3	439	50	389
немски	12	5	819	144	675
унгарски	12	6	3801	495	3306
италиански	12	11	1645	292	1353
полски	12	1	2511	422	2089
сръбски	12	11	1434	259	1175
общо			20932	4764	15177

Таблица 4: Разпределение на брой предмети и модалности спрямо език и ниво на обучение.

Освен по модалност, въпросите варират и по приложна област. Събрани са въпроси от 20 различни предмета. Тъй като в различните държави предметите с еднаква насоченост имат различно наименование, бе приложена агрегация и групиране на агрегацията. 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 с голяма вариативност спрямо езиците.

С цел предотвратяване на голямата вариативност в [2] се предлага двустепенен процес за агрегиране на изходните предмети, който се използва и

при последващата обработка на Exams-V. Резултатът са три логически свързани йерархизирани групи - първата е изходният предмет, втората е нормализираният му вид, а третата обхваща трите основни клона на естествените науки: *наука за природата* - изучаване на естествени феномени, *социална наука* - изучаване на човешкото държание и човешките общества и *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. Това продължава, докато не се получат общи имена на предмети без подходящи кандидати за сливане. Така се постига и консистентност на имената на предметите между различните езици (т.е. се избягват различията поради 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.

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

```
директория за тип множество (тренировъчно или тестово)
      директория за всеки език
            директория за всеки предмет
                  директория за въпроси само с текст
                        JSON файл с анотации
                        директория с картинки
                  директория за въпроси с 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 описва въпросите за всеки изпит.

Разпределението на броя езици, в които е наличен агрегиран предмет, е показано на таблица 6. Таблицата е слабо популирана. Някои предмети се срещат само в един език - селско стопанство, изящни изкуства. Това се дължи на специфичността на предмета. Най-разпространените предмети са популирани с най-много въпроси тук - физика, химия, история, география. Това са основни предмети и допринасят за формирането и развиването на обща култура, което обяснява тяхната популярност. Хубавото на тези предмети е, че те са много богати на различните модалности - има много въпроси само текст, но има и много въпроси с текст и картинка. От хърватски език има много изпити, които са налични, както и на хърватски, както и на сръбски и италиански.

Език	Брой паралелни въпроси
арабски	244
хърватски	2043
италиански	1590
сръбски	1209

Таблица 5: Разпределение на паралелните въпроси.

Агрегиран предмет	ар	бъл	кит	хър	анг	фре	нем	унг	ита	пол	сръ	общо
селско стопанство								652				652
биология	67	233	281	574	47				424		294	1920
бизнес икономика						180	216	747			6	1149
химия	150	665	381	427	347		14	697	212		322	3215
етика				180					25			205
изящни изкуства				48								48
география			678	383		24	46	92	40		54	1317
история			209	500					235		200	1144
информатика				188					146		74	408
математика			678		145							823
философия				140					34		12	186
физика	67	970	408	649	185	235	510	1570	215		305	5114
политика				270					100		90	460
професионални										2511		2511
психология				154					105		47	306
религия				161								161
наука	150				20							170
ислямски науки	83											83
социология	306	264		295					109		30	1004
туризъм							33	43				76
общо	823	2132	2635	3969	744	439	819	3801	1645	2511	1434	20932

Таблица 6: Разпределение на предметите по езици.

Структурата, чрез която се пазят въпросите е JSON формат. 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`);
- `answer_key`: верен отговор на въпроса;
- `type`: тип на въпроса - символен низ, определящ дали въпросът включва картинки или само текст (част от вложен обект `info`);
- `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`);
- `subject_aggregated`: агрегиран предмет (част от вложен обект `info`);
- `subject_aggregated_grouped`: група на естествените науки, към която принадлежи агрегираният предмет (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`);
- `chemical_structure`: ако съществува, оказва дали въпросът съдържа химически символи или структури;
- `table`: ако съществува, оказва дали въпросът съдържа таблица;
- `figure`: ако съществува, оказва дали въпросът съдържа фигура;
- `graph`: ако съществува, оказва дали въпросът съдържа графика;
- `parallel_id`: ако съществува, оказва идентификатора на въпроса, на който текущият въпрос е паралелен.

Примерно съдържание на файла с анотации:

```
{
        "id": "ce0bc97a-e140-4d28-a295-24adecf01756",
        "question": {
                "question_snapshot": "Train/BG/Chem/image_text/ims/01623_im_01.jpg",
                "question_number": 24
        },
        "answer_key": "B",
        "info": {
        "type": "image_text",
        "grade": "12",
        "subject": "Chemistry",
        "subject_aggregated": "Chemistry",
        "subject_aggregated_grouped": "Natural Science",
        "language": "Bulgarian",
        "extra": {
                "date": "2012-05-23"
        },
        "multimodalities": {
                "chemical_structure": 1.0,
                "table": 0.0,
                "figure": 0.0,
                "graph": 0.0
        }
        }
}
```


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.

По този начин се получават три главни характеристики на Exams-V:

- *многомодалност*: в реалните ситуации хората често срещат проблеми от различни модалности - напр. визуални и аудио. Затова е важно да се оценяват съвременните модели да обработват заедно информация от различни модалности. Тази възможност рефлектира ключова част от възможността за разсъждаване. Затова този набор от данни е добър - включва и въпроси, изискващи работа с картинки, за правилното им отговаряне.
- *многоезичност*: включените въпроси са от разнообразни езици с различна езикова фамилия. Наблюдава се постепенно преминаване от използване на набори от данни само на английски език към използване на такива, включващи и други езици. По този начин, от една страна, се дава възможност за забелязване на предубеждения в използваните тренировъчни данни, а от друга - могат да се оценят възможностите на големите многомодални модели в контекста на различни култури и особено тези, за които не е налична много свободно достъпна информация в Интернет. Подобни резултати са много важни за прилагането на такива модели в тези общества. За да се осъществи това е нужно да се събират данни от реални, естествени източници в различни езици,

а не, например, да се превежда от английски език, т.к. истинските характеристики на културата и нейното световно знание се коренят именно в автентичните данни.

- *многонивост*: въпреки че има разлика в образователните системи в различните държави, в общия случай те съдържат структура от последователни степени, преминаващи под някаква форма през начално образование, средно образование и висше образование, успешното завършване на които се отъждествява с успешното полагане на изпит. Тези изпити на практика показват очакванията на образователната система за дадения период за всяка държава. Затова оценяването на големите многомодални модели с въпроси от тези фази на образованието ще даде подробна оценка на техните възможности да покриват различни нива на интелектуални изисквания.

Множеството от данни Exams-V предоставя разбиване на данните на данни за трениране и тестване. Множеството от данни за тестване се състои от 4,798 случайно избрани въпроса. Критерият за избор е броят въпроси в 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, е извършена случайна извадка с размер 50.

Разпределението на визуалните характеристики в тестовото множество е показано в таблица 7.

Език	Таблица	Фигура	Графика	Хим. символи	Само текст
английски	0	38	28	43	159
китайски	82	246	56	56	126
френски	0	37	6	9	124
немски	0	35	10	8	179
италиански	11	151	13	39	382
арабски	11	81	12	29	211
полски	14	36	0	0	50
унгарски	0	36	7	80	270
български	6	70	10	51	200
хърватски	19	330	24	62	780
сръбски	11	151	13	39	357
общо	154 (3,21%)	1211 (25,24%)	179 (3,73%)	416 (8,67%)	2838 (59,15%)

Таблица 8: Разпределението на визуалните характеристики в тестовото множество.

От въпросите, съдържащи картинки се наблюдава доминиране на тези, съдържащи фигура. Анализ на част от въпросите показва, че въпросите с

фигури са по-лесни за отговаряне, т.к. идват от предмети, свързани с разпознаване на известни забележителност, части на човешкото тяло и обекти - задачи, в които съвременните модели са много добре развити.

4 Експерименти

Глава 4 представя процесът по провеждане на експериментите. Описан е начинът, по който се представят данните на моделите, както и начинът, по който получените от тях резултати са обработени и до какви изводи се достига след анализирането им. С цел получаване на по-точна представа за качеството на резултатите, анализ е направен както на цялостното качество, така и на база сегментационни групи.

4.1 Подготовка

Оценяването на моделите е извършено в две фази: оценяване на качеството на резултатите, постигнати върху цялото тестово множество, и разглеждане на точността чрез примери в сегментационни групи.

За оценяване са избрани най-добрите съвременни многомодални модели: gpt-3.5-turbo-1106, gpt-4-1106-preview, , gpt-4-vision-preview, gemini-pro, gemini-pro-vision, LLaVa-1.5, Qven-VL-7B. Всички експерименти са проведени без предостоставяне на предварително оценени примери. Това подсигурява три важни състояния:

1. Този начин на оценяване симулира най-добре ситуации, които биха възникнали в реално приложение, или при полагането на изпит.
2. Използваните модели имат лимитиран контекстен прозорец (особено моделите, различни от Генеративен претрениран модел 4-B) и не могат да приемат повече от едно изображение (с изключение на Куен виел), което ги прави неподходящи за даване на примери в една инструкция.
3. Използваните модели са преминали през фаза на фино настройване за следване на инструкции и анализът на литературните източници показва, че подаването на примери няма голяма добавена стойност.

4.2 Провеждане

Използваната инструкция при изпращане на всяка изрязана снимка на въпрос е: *The image has an multiple choice question with 3 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* (на български: *В изображението има въпрос с 3 до 5 възможни отговора. Изведи правилния отговор във формат JSON, както следва: {'answer': 'xxx'}. Замести „xxx“ със съответната буква: „A“, ако първият отговор е правилен, „B“, ако вторият отговор е правилен, „C“, ако третият отговор е правилен, „D“, ако четвъртият отговор е правилен, или „E“, ако петият отговор е правилен*). За разлика от [8] в инструкцията не се дава информация за предмета, от който е въпросът. След инструкцията се предоставя изображението. Спрямо целта на експеримента, след изображението

се предоставя и описание на картинката и извлечения текст на въпроса. Езикът на инструкцията е независим от езика на въпроса, т.е. за всеки език, инструкцията винаги е на английски език. Пример за цялата информация, подавана на модела, за получаване на отговор на един въпрос, е даден на фигура 14.

The image has an multiple choice question with 3 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.

15. Која табела истинитости одговара логичком изразу $\overline{A+\overline{A}\cdot B}\cdot(A+B\cdot C)$?

A.

A	B	C	излаз
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	0

B.

A	B	C	излаз
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

C.

A	B	C	излаз
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

D.

A	B	C	излаз
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

A.
B.
C.
D.

Фигура 14: Пример за информацията, подавана на моделите.

От гледна точка на реализирания код, за работа с моделите се използва програмен интерфейс, чрез който инструкцията и прилежащата снимка се подават като множество от двойки ключ-стойност. Единият ключ съхранява инструкцията, а другият - пътят до снимката. Спрямо използвания модел, могат да се подават допълнителни параметри, напр. температура, максимален брой токени в отговора и подобни. Съобщение, което се изпраща за получаване на отговор от Генеративен претрениран модел 4-B може има вида:

```
payload = {
	"model": "gpt-4-vision-preview",
	"messages": [
		{
			"role": "user",
			"content": [
				{
				"type": "text",
				"text": "The image has a multiple choice question
			with 3 to 5 choices. Provide the correct answer
			precisely in JSON format as follows without any
			surrounding text: {'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."
				},
				{
				"type": "image_url",
				"image_url": {
					"url":
f"{base64_image}"}}]}],
			"max_tokens": 4000
			}
```

Примерът показва как изглежда множеството от двойки ключ-стойност, когато към модела се подават инструкция и картинка. За оптимизиране на използваното пространство и намаляване на необходимото количество токени за прочитане на картинката, тя се преобразува в стандартната кодировка Base64.

Въпреки че инструкцията е подробна и в частност изисква получаване на отговора във формат JSON, понякога моделите не спазват посочения формат. Наблюдават се три типа отговори, давани от моделите:

1. Връща се отговор в JSON формат. Това е най-добрият случай и показва, че моделът е разбрал инструкцията и прилежащият към нея въпрос и е върнал отговор.
2. Връща се отговор, който е в разговорен стил. В тази ситуация моделът е разбрал въпроса в картинката и е формирал отговор, но не е спазил цялостно инструкцията. Възможно е да се дължи на ограничения контекстен прозорец или на недостатъчно добро разбиране на инструкцията.
3. Моделът отказва да съдейства и не връща отговор. Възможно е това да се дължи на системата, гарантираща качеството и безопасността на резултатите.

Това, че отговорът не е в JSON формат, налага нуждата от последваща обработка върху получените отговори за отделянето на буквата, съответстваща на отговора. Понякога вместо посочване на буква се изписва съдържанието на отговора с тази буква - в такива случаи отговорът се счита за грешен. Част от последващата обработка се извършва чрез код, използващ регулярни изрази. Понеже възможностите имат голяма вариативност и са неконсистентни е нужна

и ръчна обработка чрез подходящ софтуер за визуализиране и обработка на таблици - в случая бе използван LibreOffice Calc. Друга трудност е броят пъти, в които се посочва отговора. Възможно е да има множество посочени и различни отговори. Тогава трябва да се намери начин за идентифициране на правилния от тях - позицията му не винаги е ясна. Възможно е първо моделът да посочи защо неправилните отговори са неправилни, и след това да посочи правилния, но е възможно и обратното. По аналогия с [8] се взема първият намерен отговор.

4.3 Отправна точка за сравнения

Разговорният агент Бард бе използван за получаване на отправна точка за сравняване на моделите в началото на 2024 г. Той бе свободно достъпен чрез уеб приложение в браузъра, което от 08.02.2024 г. се превръща в платформата, в която се развива новият разговорен агент на Гугъл - Джеминай. Към момента на използването на Бард нямаше свободно достъпен програмен интерфейс за комуникация чрез HTTP заявки, който официално да се поддържа от Гугъл. Това бе недостатък спрямо конкурентите на Бард и в частност Чатджипити, който може да се използва, след закупуване на ключ за достъп. Въпреки това в [16] може да бъде намерена неофициален пакет, разработен на езика Python, поддържан от доброволци. Функционалностите имат технически лимитации. Честите обновявания на функционалностите понякога водят до неочаквано поведение при използване на методи. Пакетът продължава да може да се използва след преименуване на Бард в Джеминай.

За комуникация с уеб платформата бе използван пакетът requests. Той предоставя възможност за изпращане и получаване на заявки във формата на HTTP/1.1. От гледна точка на потребителя, библиотеката е написана на високо ниво и позволява автоматично изграждане и поддържане на множество сесии към HTTP сървъри. Примери за това са автоматичното добавяне на параметри към GET заявки, автоматичното прехвърляне на данните като параметри на заявката вместо на URL адреса при заявки от тип POST, SSL верификация, автоматично декодиране на върнатите отговори и други.

Начинът на взаимодействие според [16] се изразява в създаване на сесия чрез requests и задаване на стойности на определени бисквитки, които позволяват изпращането на заявки и получаването на отговори. При изтичането на валидността на поне една от тях, като резултат от подаването на заявката се връща грешка с текст *Response Error*. Трите бисквитки имат временна валидност и изтичат по различно време. При изтичането на коя да е от тях се нарушава възможността за комуникация с разговорния агент. Това води до непредвидимо количество заявки, които могат да се подават към модела. Експерименти показват, че заявките имат различен период на валидност в зависимост от държавата и региона, от които се правят, и в зависимост от времето на деня, когато се правят. В допълнение влияние оказва и неравномерното натоварване на платформата, поддържаща разговорния агент.

След множество експерименти максималният брой заявки за клиент, разположен в България, е 89. Освен това след една успешна сесия (в която са били изпратени и успешно обработени 89 заявки), трябва да се изчака поне 10 часа, за да се извърши отново втора успешна сесия с 89 въпроса.

Фигура 15 показва стила, в който Бард връща отговорите. Той е разговорен и изисква допълнителна обработка за правилното идентифициране на отговора. Забелязва се, че редът на вземането на идентифицираните букви на отговорите има значение.

Фигура 15: Пример за стила на отговорите, които разговорен агент Бард връща.

Фигура 16 показва въпрос, на който Бард отказва да отговори. Макар и това да е в резултат от допълнителните системи, които проверяват целта на инструкцията преди да бъде подадена към модела, и които поддържат безопасността на върнатите отговори, въпросът тук не би следвало да провокира опасно държание на модела. Възможно е причината да се корени в контекста на коментиране на религия (в частност, кръщаване) и големите последствия, които приемането на учениците на Кирил и Методий е имало.

Фигура 16: Пример за въпрос, на който разговорен агент Бард отказва да отговори.

Резултатите, които Бард постига върху случайна извадка от 50 въпроса за част от езиците в тестовото множество от Exams-V, са показани на таблица 7.

Език	Точност
немски	23,49
френски	32,33
унгарски	19,33
български	22,92
италиански	30,00
хърватски	22,19
китайски	27,33
Средно	**25,37%**

Таблица 8: Резултати върху случайна извадка от тестовото множество, групирани по език за разговорен агент Бард.

4.4 Резултати

Таблица 9 представя резултатите, групирани по език, на избраните модели за сравняване. Налично е и сравнение със случаен избор за по-ясна представа за добавената стойност, която имат моделите. Случайният избор взема произволен отговор с равномерна вероятност от възможните. В зависимост от средния брой възможни отговори за всеки език, точността на метод варира в границите между 19% и 26%.

Най-добра средна точност от многомодалните модели постига Генеративен претрениран модел 4-B - 42,78%, което е едва с 20 пункта повече от средната за езиците. Това показва, от една страна, че наборът от данни Exams-V може да служи за цялостно оценяване на тези модели и че те все още имат място

за подобрение. Моделът с втора най-добра точност е Джеминай-В, който постига средна точност от 31,13%.

В сравнение с търговските модели, моделът с отворен код Лава 1,5-13В и моделът Куен виел-7В се представят незадоволително и не показват добри възможности за обработване на информация от различни езици и в различни модалности. Това че са тренирани да работят с малък брой езици (2 за Куен и 1 за Лава), предразполага към постигане на голяма точност на въпросите от тях. Това обаче не си проличава в проведените експерименти, в които тяхното представяне е сходно с това на техниката със случаен избор. В резултат на това се наблюдава голяма разлика между техните резултати и тези на търговските модели, които в почти всички езици задминават случайния избор.

Ако към най-добрите големи езикови модели вместо подаване на картинката, съдържаща въпроса и прилежащите към него отговори и изображения, се подава извлеченият от картинката текст (включващ въпроса и възможните отговори) и текстово описание на прилежащите към въпроса изображения (ако такива има), средната точност нараства. За извличане на текста в подаваните към моделите картинки, се използва пакетът *Гугъл Тесеракт*, а за получаване на описание на прилежащите изображения се използва Генеративен претрениран модел 4-В.

Модел	бъл	кит	хър	фре	нем	унг	анг	сръ	ита	ар	пол	Средно
Случаен избор	25,23	24,13	25,58	24,14	22,56	19,55	24,40	24,50	24,76	19,83	23,00	23,62
Многомодални модели												
LLaVA-1.5-13B	-	-	-	-	-	-	26,00	-	-	-	-	-
Qwen-VL-7B	-	15,72	-	-	-	-	23,60	-	-	-	-	-
GPT-4V	**36,00**	22,20	55,47	**60,34**	**51,24**	**44,77**	29,27	39,84	62,07	24,29	30,00	42,78
Gemini Pro Vision	30,46	**24,56**	29,39	47,70	47,80	27,05	29,20	28,29	43,03	19,38	28,00	31,13
Големи езикови модели с извличане на текста и подаване на описание												
GPT-3.5 Turbo	27,08	22,20	52,08	39,08	34,81	37,73	30,00	48,61	55,48	26,36	33,00	39,47
GPT-4	30,46	23,57	**66,58**	36,71	23,76	34,09	**32,40**	**73,51**	**75,95**	**26,47**	30,00	**47,11**
Gemini Pro	32,00	23,97	58,90	38,51	28,09	43,41	31,20	59,96	64,38	23,25	**42,00**	43,99

Таблица 9: Точност на различните модели върху тестово множество на Екзамс-В спрямо език. Най-добрите резултати са удебелени, а тези на второ място са подчертани.

От всички модели най-добра точност постига Генеративен претрениран модел 4 с подаване на извлечен текст и описание. Това показва, че комбинацията от прецизното разпознаване на символите от *Гугъл Тесеракт*, цялостните и ясни описания на Генеративен претрениран модел 4-В и гъвкавите възможностите за разсъждаване върху текст на Генеративен претрениран модел 4, е най-добрата система за автоматично отговаряне на многомодални въпроси. От гледна точка на модела, по-доброто представяне на големите езикови модели може да се дължи именно на разбиването на обработването на изображението на част, която обработва и подготвя текста и част, която обработва и подготвя изображението, за което служат да покажат и резултатите на разгледаните статии. Използването

на допълнителни системи за подготовка на данните, улеснява процеса за даване на отговор.

Таблица 9 показва, че спрямо езиците всички модели постигат сравнителна ниски резултати за китайски език. Те са сравними със случайния избор, което показва, че естеството на това подмножество води със себе си голямо ниво на трудност. Едно възможно обяснение е, че именно това множество съдържа най-голям процент на въпроси с картинки - всъщност това е и единственото множество, за което броят на въпроси с картинки е по-голям от броя на въпросите без картинки - 75,56%. Въпросите са извлечени от Гаокао изпитите [70] в Китай, които за задължителни за всички ученици, желаещи да продължат образованието си в висиште учебни заведения. Тези изпити са известни с това, че не са тривиални и изискват сериозна подготовка и дълбоко познаване на изучения през годините на обучение материал. Изпитът протича 9 часа в период от два до три дена, което дава представа за количеството въпроси и тяхната трудност. В Exams-V това е и множеството данни с най-голямо разнообразие между въпроси съдържащи фигури, таблици и графики. Това, че най-добрата система с Генеративен претрениран модел 4 и разпознаването на текста и описването на картинката може да не се представя толкова добре в по-голям мащаб особено ако се включват въпроси с картинки.

Освен китайски друг език, който затруднява моделите е арабският. Причина за това би могла да е структурата на въпросите. От една страна, за разбирането им, текстът трябва да се чете от дясно наляво - начин, за който моделите най-вероятно не са срещали много примери. От друга страна, ниските резултати могат да се дължат и на формата, използван за изписване на възможните отговори. Вместо символите A, B, C или числата 1, 2, 3 отговорите са маркирани с празни квадратчета, от които се очаква учениците да запълнят това, което съответства на правилния отговор. Фигура 17 показва такъв пример. Експериментите показват, че при запитване на моделите да предоставят отговора под формата на символ - A, B, C или D, те се затрудняват да съобразят кое квадратче на кой символ отговаря. По същество това може да се смята отново за вид многомодалност, т.к. моделът трябва да разбере семантиката на всяко квадратче, без тя да му е обяснена в инструкцията.

6

يشير الجزء المظلل على خريطة دولة قطر المقابلة إلى : -

☐ سبخة دخان
☐ طوير الحمير
☐ دحل الحمام .
☐ روضة راشد

Фигура 17: Въпрос, чиито идентификатори на отговори са квадратчета.

В разрез с очакванията резултатите на въпросите по английски език са едва с 8 пункта по-високи от случайния избор. Те са взети от изпита съвместения приемен изпит (от англ: *Joint Entrance Exam, JEE*), който се състои в Индия веднъж в годината. Възможно е тези изпити да не присъстват в тренировъчното множество на моделите, макар и да са на английски, което да води до ниските резултати. Също така този вид изпити също са комплексни, т.к. обхващат тематики, свързани с компютърните науки - за добро представяне е необходимо модел да покаже разсъждаване на множество стъпки, възможност за разбиване на подзадачи и много добро разбиране на основи предмети като физика, химия и математика.

От големите многомодални модели най-добре се представя Генеративен претрениран модел 4-B, постигайки най-висока точност в четири езика - български, френски, немски и унгарски. За тези езици обаче отношението на броя въпроси с картинки към броя въпроси без картинки е 5,8, което е по-високо от средното за набора данни - 3,2. Това означава, че случайно избран въпрос от четирите езика е почти 6 пъти по-вероятно да бъде въпрос без картинка. Като

следствие от това повечето въпроси с картинки не съдържат графики и таблици, които имат тенденция да затрудняват модела най-много.

4.5 Резултати със сегмент с паралелни въпроси на различни езици

Въпросите за езиците сръбски и италиански са създадени след превод на въпроси, давани на ученици в Хърватия. Обобщените резултати са показани на таблица 10.

Предмет	Генеративен претрениран модел 4-B			Джеминай Про Вижън			Генеративен претрениран модел 4 + Текст + Описание		
	хър	сръ	ита	хър	сръ	ита	хър	сръ	ита
биология	72,04	37,64	66,90	32,26	31,18	43,41	75,27	73,11	77,55
химия	48,00	28,00	53,33	25,33	26,67	34,67	72,00	68,00	72,00
история	59,26	45,68	61,73	29,63	23,46	37,03	85,19	77,78	76,54
информатика	38,39	33,33	40,74	42,59	33,34	46,29	34,00	57,41	66,67
политика	82,22	46,67	73,33	46,67	31,11	64,44	97,78	91,11	86,67
психология	85,19	55,56	88,89	33,33	29,63	59,26	92,59	100,00	92,59
социология	63,33	53,33	60,00	33,33	30,00	56,67	80,00	73,33	70,00
средно	62,56	40,44	62,11	33,33	28,76	45,25	80,53	75,29	76,60

Таблица 10: Сравняване на резултатите, постигнати от моделите Генеративен претрениран модел 4-B, Джеминай Про Вижън и Генеративен претрениран модел 4 след извличане на текста в картинката и даване на описание на прилежащите към нея изображения.

Резултатите показват, че езикът има значение. Отново използването на техниките разпознаване на текст и описване на картинка водят до най-добри от Генеративен претрениран модел 4, постигайки най-висока точност за всеки от трите езика. Наблюдава се голяма разлика от 20 пункта точност в резултатите, които моделът Генеративен претрениран модел 4-B постига за езиците хърватски и италиански (съответно 62,56% и 62,11%) и резултатите му върху сръбски (40,44%). Възможно е това да се дължи на азбуката, която използват езиците - например, хърватският език използва латиница, докато сръбският използва кирилица. Количеството ресурси в Интернет, които използват латиница, е по-голямо от ресурсите на кирилица, което може да е основният фактор за лошото представяне. От друга страна, това, че след добавянето на техниките по извличане на текст и даване на описание, точността на Генеративен претрениран модел 4 върху въпросите на сръбски език се вдига до 75,29%, е неочаквано. Възможно е това да се дължи на превода към английски език, който понякога се случва в процеса на извличане на описание за съдържанието на картинката от Генеративен претрениран модел 4-B.

Разликата в точността между сръбски и хърватски съществува и при Джеминай Про Вижън, макар и в по-малък мащаб - 4,6%. Прави впечатление качеството на резултатите, които моделът постига за въпросите на италиански език. Те са най-високи за всички предмети и нямат голяма вариативност, което показва, че предметът не оказва голямо влияние. Възможно е това да се дължи на повечето налични ресурси в Интернет за езика и съответно по-голямото му участие в тренировъчното множество на модела.

4.6 Резултати със сегмент с различни типове визуална информация

Техническите описание на моделите Джеминай Про Вижън и Генеративен претрениран модел 4-B не са достатъчно подробни, за да дадат ясна представа за отличителните характеристики на моделите в контекста на разпознаване и обработка на изображения с различна семантика. Таблица 11 представя сравнение, което показва как се представят моделите, когато изображение съдържа поне една от четири характеристики: химически символи, фигури, таблици и графики. Предоставена е и точността, която постигат върху картинки, съдържащи само текст, за сравняване на способностите им за обработване на многомодални въпроси от хърватски език.

Характеристика	Наблюдения	Генеративен претрениран модел 4-B	Джеминай Про Вижън
Химически символ	36	52,78	25,00
Фигура	50	60,00	22,00
Графика	50	42,00	26,00
Таблица	40	27,50	37,50
Текст	50	62,00	48,00

Таблица 11: Сравняване на точността на Генеративен претрениран модел 4-B и Джеминай Про Вижън спрямо типовете визуална информация.

Резултатите показват, че Джеминай Про Вижън се справя по-добре, когато обработва таблици, но се представя по-зле в останалите видове визуални характеристики. Наблюдава се голяма разлика от почти 40% точност, когато въпросът съдържа фигура. Възможно е това да се дължи на разпределението на тези характеристики в тренировъчните данни, което показва, че от една страна те се различават значително и от друга - вероятно Генеративен претрениран модел 4-B е по-голям от Джеминай и е трениран с по-голямо количество данни.

5 Дискусия

Използваните моделите не успяват да постигнат много добри резултати, но все пак показват голям потенциал за успешно бъдещо развитие. Наблюдават затруднения в даването на отговор за широка гама въпроси, включващи комплексна визуална информация. Грешните отговори могат да се класифицират в два типа:

1. моделът не успява да разбере правилно и цялостно многомодалния вход и няма достатъчно опит в приложната област, за да достигне до правилния отговор;
2. моделът генерира грешни разсъждения с ненужна, невярна или непълна информация, която попречва достигането до верния отговор.

Учудващо е, че комерсиалните модели, макар и да са тренирани върху изключително голямо количество текстова и визуална информация, демонстрират голяма неефикасност при обучение. Предполага се, че и моделите от семейството на Джеминай, и моделите от семейството на Генеративните претренирани модели са тренирани с много повече информация, която явяващите се на изпит ученици нямат, но въпреки това не успяват да се справят добре с по-голямата част от въпросите.

От друга страна, спрямо различните езици точността може да варира много. Това показва, че моделите трудно обобщават научените знания на други езици и също така срещат трудности при дори малки промени в условията на въпроси. Това проличава в представянето на моделите Лава и Куен виел. Макар и състоящите се от комплекна архитектура, изглежда те по-скоро избират случаен отговор, вместо да прилагат разсъждения. В този ред на мисли, отговорите получени от разговорния агент Бард при създаването на резултатите за отправната точка, показват и една опасна характеристика - моделите могат да обърка много факти, но го правят по много добре аргументиран начин, който е предоставен на грамотен език и звучи много убедително.

6 Заключение

Настоящата дипломна работа използва новосъздадения набор от данни Екзамс-В за предоставяне на множество отправни точки за бъдещи сравнения с големи многомодални модели, включително Джеминай Про Вижън и Генеративен претрениран модел 4-В. В допълнение, се показва, че Екзамс-В е важна част за цялостното оценяване на големи многоезични модели, които могат да се интегрират в системи, разполагащи с възможност за извличане на текст от картинка и конструиране на нейно описание. Освен разнообразието от приложни области и комплексността на включените въпроси, наборът от данни демонстрира нов начин за оценяване на възможностите на моделите да отговарят автоматично на многомодални въпроси - вместо въпрос за картинка да се подава отделно от нея, въпросът е част от картинката и те се подават заедно.

Множество експерименти са проведени с цел изследване на добавената стойност на визуалното съдържание и трудността, която моделите срещат при работа с езици, различни от английски. Резултатите от тях сочат, че средно най-качествени отговори се дават от система, която се състои от три компонента:

1. Модел, който извлича текстовото съдържание от картинка (напр. чрез *Гугъл Тесеракт*, който е базиран на невронна мрежа с дълга краткосрочна памет);
2. Модел, който създава описание на получената картинка (напр. Генеративен претрениран модел 4-В);
3. Голям езиков модел, който приема текстовото описание и извлечения текст и генерира отговор (напр. Генеративен претрениран модел 4).

Използваните методи и подходи за решаване на задачата, описание на начина на формиране на новото множество Екзамс-В, както и подготовка и провеждане на експериментите са описани в статията *EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models* предадена към февруари 2024 г. в ACL ARR.

С цел по-нататъшно развитие може да се добавят още въпроси от разгледаните модалности; могат да се съберат въпроси от други модалности - например видео и аудио; могат да се използват по-нови и по-добри големи многомодални модели. Други възможни подобрения са:

- събиране на повече данни, оценяване на моделите чрез различни множества от данни, които включват не само предмети, изучаване в училище. Възможно е да се включат предмети изучавани в университет, а могат да се включат и извънучебни знания, напр. научаване на правилника за движение по пътищата, указване на първа медицинска помощ и подобни.
- редно е освен сравняване с техниката за случаен избор на отговор, сравнения да бъдат направени и с резултати, постигнати от човек. От една страна, това е широко прилагано в разгледаните литературни

източници. От друга - това ще помогне за по-добро сравняване на уменията на моделите, т.к. ще го постави в контекста на човешките възможности. Макар и за текущите резултати интуитивно да се подразбира, че са много по-ниско от тези на човек, е подходящо да се добави подобно сравнение, т.к. с времето моделите ще стават все по-добри.

- с цел получаване на по-ясна представа за мисловия процес на използваните модели, могат да се проведат експерименти с различни стратегии за извличане на разсъжденията им - например, извличане на веригата от разсъждения при отговаряне на текстови въпроси и извличане на програмата от разсъждения при отговаряне на въпроси, свързани със създаването на код.
- използваните комерсиални модели са тренирани с голямо количество данни, включващи програмен код. Полезно би било да се извърши сравнителен анализ на повече въпроси от предмети, свързани с информатиката и компютърните науки.
- освен метриката точност, за оценяване на модела, може да се приложи и проверката за успешно взет изпит. Това ще позволи по-лесен анализ на грешките, които прави модела, и ще различни модели ще могат да бъдат сравнявани в зависимост от това дали успешно са взели изпита.
- макар и текущият начин за провеждане на експериментите да е най-подящ от гледна точка на реалната ситуация, ще е полезно да се проведат повече експерименти - например, с предварително подаване на няколко примера, сравняване на добавената стойност на извличане на верига или програма от разсъждения и подобни.
- с цел по-ясно представяне на ефекта при наличието на няколко картинки към един въпрос, може да се направят експерименти, които проверяват как се променя точността, ако към въпроса има понече от една картинка и в частност - как варира точността, ако всички възможни отговори са картинки.
- таксономията от типове картинки може да продължи да се разширява и да се разбива на по-финна грануларност. Това ще позволи сегментационен анализ и прецизиране на проблемите визуални характеристики, с които моделите не се справят достатъчно добре.
- към момента езикът на инструкцията е английски и поради съществото на инструкцията, очакванията отговор също е на английски. Полезно ще е експериментиране с езика на инструкцията и проверка на оказаното влияние върху точността и държанието на модела, когато инструкцията се превежда на езика, на който е въпросът.
- в набора от данни се срещат дублиращи се въпроси. Това е така, т.к. един и също въпрос може да се (и се) появява в различни години на изпитване. Редно е да се изследва как това влияе на моделите и при трениране на такъв модел, използвайки този набор от данни, трябва да се внимава да

не се получи изтичане на част от данните за трениране в тези за валидация и тестване.

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 (достъпен онлайн на 04.02.2024 г.).
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. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, abs/1409.1556. http://arxiv.org/abs/1409.1556
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 (достъпен онлайн на 04.02.2024 г.).
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 (достъпен онлайн на 04.02.2024 г.).
21. Ivan Goncharov, A modified version of https://github.com/Cartucho/OpenLabeling OpenLabelling tool: https://github.com/ivangrov/ModifiedOpenLabelling (достъпен онлайн на 04.02.2024 г.).
22. Jeffrey A. Clark (Alex), The Python Imaging Library adds image processing capabilities to your Python interpreter: https://pypi.org/project/Pillow (достъпен онлайн на 04.02.2024 г.).
23. Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. In Findings of the Association for Computational Linguistics (EMNLP), pages 1896–1907, 2020.
24. Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big selfsupervised models are strong semi-supervised learners. Advances in neural information processing systems (NeurIPS), 33:22243–22255, 2020.
25. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In The International Conference on Learning Representations (ICLR), 2021.
26. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

27. A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In ECCV, 2016.
28. Qwen-VL: The Official Repo of Qwen-VL (通义千问-VL) Chat & Pretrained Large Vision Language Model Proposed by Alibaba Cloud (достъпен онлайн на 04.02.2024 г.).
29. Blog, Introducing ChatGPT, *Openai.com*, https://openai.com/blog/chatgpt (достъпен онлайн на 04.02.2024 г.).
30. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv.org, https://arxiv.org/abs/1707.06347, arXiv:1707.06347 [cs.LG].
31. GPT-3.5 Turbo fine-tuning and API updates, *Openai.com*, https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates (достъпен онлайн на 04.02.2024 г.).
32. Cross-talk between ChatGPT and Copilot?, *Openai.com*, https://community.openai.com/t/cross-talk-between-chatgpt-and-copilot/40775 (достъпен онлайн на 04.02.2024 г.).
33. The world's most widely adopted AI developer tool., *github.com*, https://github.com/features/copilot (достъпен онлайн на 04.02.2024 г.).
34. Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." Neural Information Processing Systems, vol. abs/2203.02155, 2022, https://doi.org10.48550/arXiv.2203.02155.
35. Bologna, Gabriele. ChatGPT-Jailbreaks: Official Jailbreak for ChatGPT (GPT-3.5). Send a Long Message at the Start of the Conversation with ChatGPT to Get Offensive, Unethical, Aggressive, Human-like Answers in English and Italian (достъпен онлайн на 04.02.2024 г.).
36. "Jailbreak Chat." Jailbreakchat.com, https://www.jailbreakchat.com/ (достъпен онлайн на 04.02.2024 г.).
37. ONeal, A. J. ChatGPT-Dan-Jailbreak.Md (достъпен онлайн на 04.02.2024 г.).
38. "GPT-4." *Openai.com*, https://openai.com/gpt-4 (достъпен онлайн на 04.02.2024 г.).
39. "GPT-4." *Openai.com*, https://openai.com/research/gpt-4 (достъпен онлайн на 04.02.2024 г.).
40. *Openai.com*, https://cdn.openai.com/papers/gpt-4.pdf (достъпен онлайн на 04.02.2024 г.).
41. Vincent, James. "OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: 'We Were Wrong.'" *The Verge*, 15 Mar. 2023, https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview .
42. Begley, C. Glenn, and John P. A. Ioannidis. "Reproducibility in Science: Improving the Standard for Basic and Preclinical Research." *Circulation Research*, vol. 116, no. 1, 2015, pp. 116–126, https://doi.org10.1161/circresaha.114.303819.

43. *Prize: A Prize for Finding Tasks That Cause Large Language Models to Show Inverse Scaling* (достъпен онлайн на 04.02.2024 г.).
44. Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. Inverse Scaling Prize: First round winners, 2022. URL https://irmckenzie.co.uk/round1 (достъпен онлайн на 04.02.2024 г.).
45. Zhang, Sarah J., et al. "Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models." *arXiv [Cs.CL]*, 2023, http://arxiv.org/abs/2306.08997.
46. Stierand, Björn. *Notion – The All-in-One Workspace for Your Notes, Tasks, Wikis, and Databases*. 2019, https://flower-nutria-41d.notion.site/No-GPT4-can-t-ace-MIT-b27e6796ab5a48368127a98216c76864.
47. Thoppilan, Romal, et al. "LaMDA: Language Models for Dialog Applications." *arXiv [Cs.CL]*, 2022, http://arxiv.org/abs/2201.08239.
48. Anil, Rohan, et al. "PaLM 2 Technical Report." *ArXiv*, vol. abs/2305.10403, 2023, https://doi.org10.48550/arXiv.2305.10403.
49. Anil, Rohan, Sebastian Borgeaud, et al. "Gemini: A Family of Highly Capable Multimodal Models." *arXiv [Cs.CL]*, 2023, http://arxiv.org/abs/2312.11805.
50. Krawczyk, Jack. "Bard's Latest Update: More Features, Languages and Countries." *Google*, 13 July 2023, https://blog.google/products/bard/google-bard-new-features-update-july-2023/ (достъпен онлайн на 08.02.2024 г.).
51. "Gemini - Google DeepMind." *Deepmind.Google*, https://deepmind.google/technologies/gemini/ (достъпен онлайн на 08.02.2024 г.).
52. Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." *arXiv [Cs.CY]*, 2020, http://arxiv.org/abs/2009.03300.
53. Rakowski, Brian. "Pixel 8 Pro — the First Smartphone with AI Built in — Is Now Running Gemini Nano, plus More AI Updates Coming to the Pixel Portfolio." *Google*, 6 Dec. 2023, https://blog.google/Пproducts/pixel/pixel-feature-drop-december-2023/ (достъпен онлайн на 08.02.2024 г.).
54. "Gemini Models." *Google AI for Developers*, https://ai.google.dev/models/gemini (достъпен онлайн на 09.02.2024 г.).
55. *Openai.com*, https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo (достъпен онлайн на 09.02.2024 г.).
56. Hsiao, Sissie. "Bard Becomes Gemini: Try Ultra 1.0 and a New Mobile App Today." *Google*, 8 Feb. 2024, https://blog.google/products/gemini/bard-gemini-advanced-app/ (достъпен онлайн на 10.02.2024 г.).
57. Liu, Haotian, et al. "Visual Instruction Tuning." *ArXiv*, vol. abs/2304.08485, 2023, https://doi.org10.48550/arXiv.2304.08485.

58. Liu, Haotian, Chunyuan Li, Yuheng Li, et al. “Improved Baselines with Visual Instruction Tuning.” *ArXiv*, vol. abs/2310.03744, 2023, https://doi.org10.48550/arXiv.2310.03744.
59. Dosovitskiy, Alexey, et al. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” *International Conference on Learning Representations*, vol. abs/2010.11929, 2020, http://arxiv.org/abs/2010.11929.
60. Radford, Alec, et al. “Learning Transferable Visual Models from Natural Language Supervision.” *Arxiv.org*, http://arxiv.org/abs/2103.00020.
61. “Papers with Code - Visual Grounding.” *Paperswithcode.com*, https://paperswithcode.com/task/visual-grounding (достъпен онлайн на 10.02.2024 г.).
62. Ilharco, Gabriel, et al. *OpenCLIP*. Zenodo, 2021.
63. *Reddit.com*, https://www.reddit.com/r/LocalLLaMA/comments/17lvquz/clearing_up_confusion_gpt_35turbo_may_not_be_20b/ (достъпен онлайн на 11.02.2024 г.).
64. Bai, Jinze, et al. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. 2023, http://arxiv.org/abs/2308.12966.
65. Lubbad, Mohammed. “GPT-4 Parameters: Unlimited Guide NLP’s Game-Changer.” *Medium*, 19 Mar. 2023, https://medium.com/@mlubbad/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a.
66. *Reddit.com*, https://www.reddit.com/r/artificial/comments/14kgs21/gpt4_is_8_x_220b_params_17t_params/ (достъпен онлайн на 11.02.2024 г.).
67. Chintala, Soumith. “GPT-4: 8 x 220B Experts Trained with Different Data/Task Distributions and 16-Iter Inference… https://T.Co/mfsK7a6Bh7.” *Twitter*, 20 June 2023, https://twitter.com/soumithchintala/status/1671267150101721090 (достъпен онлайн на 11.02.2024 г.).
68. “Gemini API Pricing.” *Google AI for Developers*, https://ai.google.dev/pricing (достъпен онлайн на 30.01.2024 г.).
69. “Pricing.” *Openai.com*, https://openai.com/pricing (достъпен онлайн на 30.01.2024 г.).
70. Wikipedia contributors. (2024, February 7). *Gaokao*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Gaokao&oldid=1204617976 (достъпен онлайн на 17.02.2024 г.).