



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: 6M13400191

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
2	Преглед на областта.....	4
2.1	Основни понятия.....	4
2.2	Подобни разработки.....	4
2.2.1	Visual Question Answering (Visual QA / VQA).....	4
2.2.2	Science Question Answering (Science QA).....	6
2.2.3	Textbook Question Answering (Textbook QA / TQA).....	7
2.2.4	SciBench.....	9
2.2.5	MathVista.....	10
2.2.6	Massive Multi-discipline Multimodal Understanding (MMMU).....	12
2.2.7	A Multilingual, Multimodal, Multilevel Benchmark (M3Exam).....	13
2.2.8	Exams.....	15
2.2.9	Сравнителен анализ и изводи.....	15
2.3	Многомодални и многоезични модели.....	17
2.3.1	GPT-3.5-Turbo.....	17
2.3.2	GPT4.....	19
2.3.3	Чатбот Bard и модели от фамилията Gemini.....	22
2.3.4	LLaVA.....	24
2.3.5	Qwen-VL.....	25
2.3.6	Сравняване на моделите.....	27
3	Набор от данни Exams-V.....	30
4	Експерименти.....	40
4.1	Подготовка.....	40
4.2	Провеждане.....	40
5	Дискусия.....	47
6	Заключение.....	48
7	Използвана литература.....	49

# 1 УВОД

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].

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 картички 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 плюс картичка, или могат да се състоят само от текст, но отговорите им да съдържат картички.

Необходимо е да се оцени модела цялостно в контекста както на прости разпознавателни умения в картички, 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.

Настоящата дипломна работа има за цел да използва тези вариации и да извърши сравнителен анализ на характеристиките на съвременните големи многомодални модели и големи езикови модели, като тези от фамилиите Gemini и GPT, и да оцени тяхната точност чрез използване на въпроси, давани на изпити на ученици от различни държави и в различни нива на обучение.

В настоящата дипломна работа са поставени и изпълнени следните задачи:

- Подробен 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 на и провеждане на експерименти.
- Подробен анализ на съвременни големи езикови модели и големи многомодални модели. Проследяване на историята, довела до успеха на най-широкоизползваните модели в днешно време.
- Въведение в набора от данни Exams-V и предварителна обработка за подаване към моделите.
- Планиране и провеждане на експерименти за оценяване и сравняване на:
  - многомодални модели;
  - езикови модели;
  - модели след превод на въпроса на английски език;
  - модели след извлечане на описание на картинката.
- Анализ на грешките, допускани от моделите.

## 2 Преглед на областта

### 2.1 Основни понятия

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

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

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

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

**Чатбот** (от англ. *chatbot*) - уеб интерфейс, използващ основен модел, за да имитира човешки разговор чрез обменяне на съобщения с разнообразни модалности.

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

**Голям многомодален модел** (от англ. *large 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 К-12** - обхващаща 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 Подобни разработки

#### 2.2.1 Visual Question Answering (Visual QA / VQA)

Задачата за даване на отговор в свободен стил след получена картичка като вход се въвежда за първи път в [6]. Тя цели да провери доколко големите визуално-текстови модели могат да върнат правилен отговор в текстов вид по

подадена картичка и въпрос за нея. Въпросът е отворен и е насочен към конкретни части на картичката. Визуалното съдържание е подбрано от наборът данни MS COCO [14], който предоставя снимки от ежедневни ситуации в реалния свят, които са автентични и без допълнителна обработка. Обхващат широка гама сцени и имат разнообразен контекст от обекти. Поради тази строга специфичност обаче имат много детайли и шум 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 картички с цел фино оценяване на моделите.

При подбора на моделите за оценяване, върху формираното от тях множество - 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 К най-близки съседи на текущата двойка въпрос-картичка;
- 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 VGGNet, за да даде отговор на въпроса, използвайки единствено изображението;
- 6) модел, базиран на невронна мрежа тип еднопосочен LSTM.

Резултатите им показват, че моделът, разчитащ само на картичката, 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 картичката. 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 картичка.

Най-добрият модел постига точност около 60%. Той е комбинацията на двуников LSTM за влагане на думите от въпроса и VGGNet с нормализация от второ ниво (от англ. *l2 layer normalization*) за влагане на изображението. LSTM-ът 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.

## 2.2.2 Science Question Answering (Science QA)

Създаването на първия набор от данни, фокусиран върху многомодални въпроси, взети от изпити, давани на ученици, е описано в [5]. ScienceQA е считан за голяма и важна стъпка към цялостното и систематичното оценяване на големи многомодални модели. Изследва се възможността на модели да създават верига от логически свързани мисли 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 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.

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 усуга, предоставена от софтуерната компания Amazon 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 UnifiedQA [23] и GPT-3 [24]. В експериментите без подаване на предварително оценени примери, използваният формат за вход-изход е QCM→A, където входът представлява обединение на текста на въпроса (Q), контекста му, който, в случая на многомодални въпроси, включва и обяснение за съдържанието на дадена картичка (C) и възможните отговори (M), а изходът е отговорът на модела. За получаване на обяснение за съдържанието в картичките се използват моделите ViT [25] и GPT-2 [26]. Провеждат се и експерименти с подаване на няколко

предварително оценени примера с цел предоставяне на възможност за адаптация към специфичната задача на ScienceQA.

Третият тип модели са фино настроени големи визуално-текстови модели (като VisualBERT и ViLT) и фино настроен UnifiedQA. Големите визуално-текстови модели приемат въпроса, тестовия му контекст (ако такъв съществува), възможните отговори и прилежащата картичка и създават точково разпределение над възможните отговори чрез линеен класификатор. Фината настройка на UnifiedQA използва подхода, описан във втория тип експерименти – текстовото съдържание се приема директно, а визулното се преобразува в текстово, използвайки обяснение за неговото съдържанието.

Получените резултати показват, че от големите визуално-текстови модели най-добре се представя VisualBERT, постигайки средна точност от 61,87%, следван от Patch-TRM, 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 UnifiedQA 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 GPT-3 постига много добра точност – 74,04%. Положителният ефект от извлечането на веригата на разсъждения заедно с отговора оказва положително влияние и на този модел. GPT-3 постига 74,17% точност след нейното прилагане, с което се превръща в най-добрания модел във всички експерименти. Точността му не е толкова далеч от тази на човек (88,40%), което показва, че дори и да обработва картичките косвено (вместо пряко), голямото количество информация, с която е претрениран, помага значително за обобщаване на разсъжденията към разнообразни области.

### 2.2.3 Textbook Question Answering (Textbook QA / TQA)

Наборът от данни 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 картички.

Авторите създават свой собствени модели и ги оценяват върху TQA. Базовият модел използва LSTM и приема на вход текстовото съдържание на въпроса и урока. Анализ на данните в TQA 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 картички. 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 извлечането 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 LSTM. 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 картички 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 DSDP-NET [27]. В първия вид визуален модел картиката преминава през невронна мрежа VGG и стойностите на параметрите от последния конволюционен слой се считат за съответния визуален контекст, който поради спецификата на VGG 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 танганс хиперболичен. Резултатните вектори са добавят като разширение към изходните такива от LSTM-а 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-льжа“ 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 картичка, първият тип визуален модел не постига по-добри резултати от текстовия модел, но вторият тип визуален модел, използващ структурен граф за описание на картичката, постига подобреие от почти пункт и половина спрямо останалите – 31,3%. Точността остава ниска, което се дължи на богатото съдържание на повечето диаграми, изискваща 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%.

#### 2.2.4 SciBench

Наборът от данни SciBench се фокусира върху изследването на представянето на модели върху въпроси от предмети, свързани с математика [7]. Целта е систематично да се изследват нужните уменията за разсъждаване над и решаване на комплексни математически задачи. Състои се от две множества данни, съдържащи въпроси от трудност в университет: с отворен характер и със затворен характер. SciBench 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 намаляне 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 GPT-3.5-Turbo (основа на чатбота ChatGPT) и GPT-4 с температура 0, за да се набледне на точността и на детерминистичните разсъждения и отговори. Разгледани са два вида стратегии за инструктиране на моделите - чрез извлечане на верига от разсъждения и преобразуване на част от разсъжденията към програмен код, написан на езика Python или Wolfram. 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 извлечане на верига от разсъждения;
5. с предоставяне на примери и с извлечане 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 моделът GPT-4 се представя по-добре от модела GPT-3.5-Turbo във всички 7 експеримента. Най-големи подобрения се наблюдават при предоставяне на примери и извлечане на верига от разсъждения и при предоставяне на примери и преобразуване на части от разсъжденията към Python - 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 GPT-3.5 и 28,52% и 28,35% за GPT-4. Във високоспециализирани предмети като квантова химия добавянето на примери все пак води до по-висока точност с почти 3 пункта за GPT-4, макар че в предмети като физика добавянето на примери води до намаляне 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 извлечане на веригата от разсъждения. Експериментите, насочени към изследването на зависимостта между генериране на код на Python и Wolfram, показват, че използването на Python увеличава средно точността съответно с 7,92% за GPT-3.5-Turbo и 7,45% за GPT-4 спрямо използването на верига от разсъждения, но използването на Wolfram намалява точността с 4,12% за GPT-3.5-Turbo и с 12,79% за GPT-4.

## 2.2.5 MathVista

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 продуктувана 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 извлечане на отговора му, следвайки процедура от три стъпки: създаване на отговора, извлечане 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 извлече буквата или цифрата, съответстваща на верния отговор. Авторите предлагат използване на система, за извършване на тази задача, базирана на голям езиков модел. Те използват GPT-4 и посочват впечатляващото му умение за извлечение 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 GPT-3.5-Turbo, GPT-4 и Claude-2 в контекста без и с предварително подаване на два примера и с извлечение 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 чатбота Bard обяснение за съдържанието на картинката и разпознатият текст в картинката от EasyOCR. Третото направление включва големи многомодални модели, част от които са GPT-4V, Bard (в случая се използва чатбот вместо PaLM 2, който е основният модел, използван от Bard), LLaMA-Adapter-V2-7B, Vicuna и miniGPT-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 GPT-4 с два предварително предоставени примера и извлечение 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 подобрене на точността. Най-добрият модел е GPT-4, постигащ точност от 33,9% с два предварително предоставени два примера и преобразуване на част от разсъжденията към програмен код на Python.

При големите многомодални модели се наблюдава голяма разлика в количеството правилни отговори, дадени от GPT-4V и всички останали. GPT-4V постига точност от 49,9%, което е с 15,1% по-високо от втория най-добър модел – комбинация от Bard и Google Lens (34,8%). Моделите с отворен програмен код не се представят задоволително. Моделът LLaVA 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.

Качествения анализ на резултатите показва, че чатботът Bard и моделите GPT-4 и GPT-4V често връщат правилен отговор, но с грешки в обясненията.

Bard 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.

## 2.2.6 Massive Multi-discipline Multimodal Understanding (MMMU)

Авторите на [9] представят MMMU като набор от данни, създаден, за 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 картичките води до по-добри резултати, се използва пакет, написан на Python за извлечение на текст от картичка - MMOCR, а за описание на съдържанието на картичките се използва LLaVA-1.5. За извлечение 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 Gemini единствено върху валидационното множество, но не и върху тестовото. Като резултат, при резултатите от валидационното множество най-добри резултати постига Gemini Ultra, следван от GPT-4V (съответно 59% и 57% точност). Сред моделите оценени върху тестовото множество, най-добри резултати постига GPT-4V, следван от Qwen-VL (съответно 56% и 41%). Отново се наблюдава голям спад в качеството на получените отговорите при модели с отворен код като BLIP2-FLAN-T5XXL и LLaVA-1.5, които достигат едва 34% точност.

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

Спръм 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 картичките 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 GPT-4V при увеличаване на трудността на въпросите. Например, докато разликата между моделите InstructBLIP-T5-XXL (40,3%) и GPT-4V (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 GPT-4V 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.

### 2.2.7 A Multilingual, Multimodal, Multilevel Benchmark (M3Exam)

Наборът от данни M3Exam се характеризира с многомодалност, многоезичност и въпроси, които са взети от всички нива на обучение до достигане на средно образование [8]. Обхванити са езиците английски, китайски, италиански, португалски, виетнамски, тайски, суахили, африкаанс и явански (които се говори в Индонезия). Изборът е продуктуван 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 GPT-3.5-Turbo, GPT-4, Claude, BLOOM и Vicuna. Сред големите многомодални модели са BLIP-2, InstructBLIP, Fromage и OpenFlamingo. Оценяват се моделите без предварително подаване на примери, главно поради това, че репликират най-добре реалния свят, трудно е да се подадат няколко картички 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 върната 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 върната опция по метода на изключването.

Резултатите показват, че докато средната точност, нужна за успешното вземане изпитите е около 54%, единствено моделите GPT-3.5-Turbo и GPT-4 успяват да постигнат по-високи средни резултати - съответно 57,57% и 72,92%. Наблюдава се, че GPT-4 се представя най-добре в различните езици. Моделът BLOOM се представя по-лошо от случайния избор. Средните точностите на моделите GPT-3.5-Turbo, Claude и Vicuna варират спрямо езика. Докато GPT-3.5-Turbo и Claude имат сходна точност за английски (75,98% и 74,25%), GPT-3.5-Turbo 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 неочеквано) 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 GPT-3.5-Turbo 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 GPT-3.5-Turbo 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 Flan-T5, въпреки че е само текстови модел. Той постига добри резултати. Към него подават само текстовата част на въпроса, а към многомодалните модели - BLIP-2 и InstructBLIP, за които се използва Flan-T5 за влагане, подават първото изображение, ако има повече от едно. Неочаквано е, че многомодалните модели не се представят значително по-добре от Flan-T5. Въсъщност единствено BLIP-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 Fromage и OpenFlamingo са специално тренирани, за да могат да работят с множество картички, те не се представят по-добре и дори се представят по-зле от модели като BLIP-2 и InstructBLIP, които работят само с една картичка. Анализът показва, че разбиране дори и на една картичка е трудно за тези модели (Fromage и OpenFlamingo). 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).

## 2.2.8 Exams

Наборът от данни е създаден от Хардалов и колеги [2]. 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 BERT модел и модела XLM-RoBERTa. Най-добрият модел е XLM-RoBERTa, който е фино настроен чрез RACE, AI2 English science datasets и най-накрая на тренировъчното множество на Exams, 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.

### 2.2.9 Сравнителен анализ и изводи

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

Име	Публикуван (Година /Месец)	Многомодалност	Най-добър модел	Втори най-добър модел	Примери помагат	Текст + Картичка по-добре от текст	OCR & Описание на картичка помагат	Верига на разсъжденията помага
VQA	2016/10	да	deeper LSTM Q + l2 norm VGGNet	LSTM Q + VGGNet	-	да	-	-
ScienceQA	2022/10	да	GPT-3	UnifiedQA	до 2 премира	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 представления 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 GPT-4V да постига по-добри резултати, когато е сравнен с чатбота Bard. Това сравнение трябва да се изследва по-подробно, т.к. не е хомогенно (модел се сравнява с чатбот).
7. Докато средно точността на големи многомодални модели като GPT-4V 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 GPT-4V да постига в общия случай най-добри резултати, все още няма достатъчно изследвания спрямо него и моделът Gemini Ultra (основният му конкурент).
9. Повечето статии анализират грешките, допуснати от GPT-3.5-Turbo, но не и тези, допуснати от чатбота Bard.

## 2.3 Многомодални и многоезични модели

### 2.3.1 GPT-3.5-Turbo

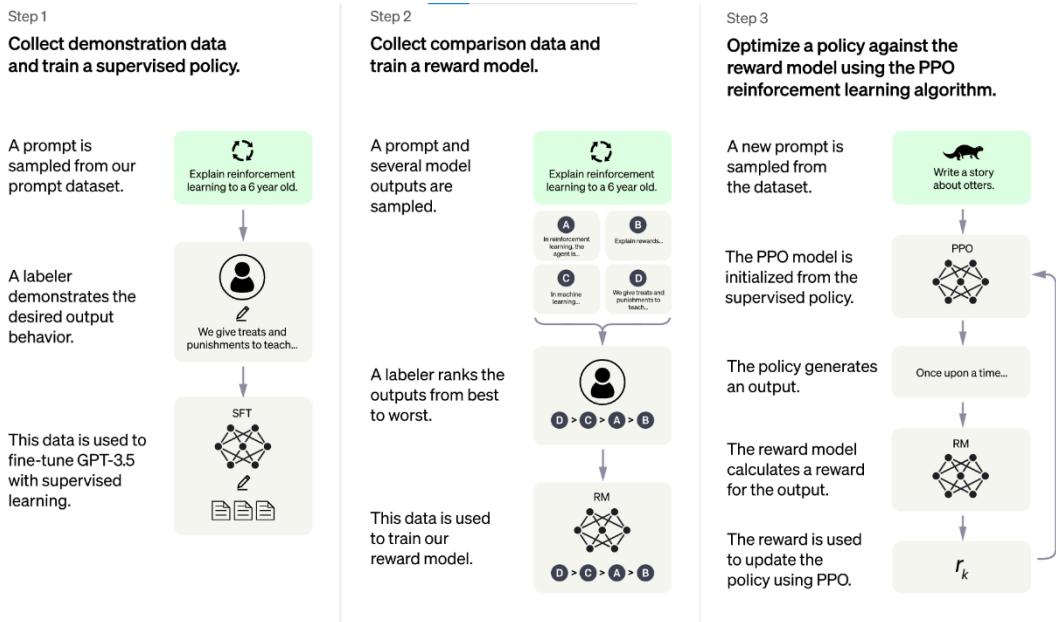
Моделът е текстови и е разработен от компанията OpenAI и е в основата на чатбота ChatGPT. Доста често в литературата наименованията GPT-3.5-Turbo и ChatGPT се използват взаимозаменяемо.

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 евективно 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 ChatGPT авторите прилагат техниката в три стъпки, визуално представени на Фигура X. Стъпките са:

- 1) Ръчно създаване на анотирани данни тип инструкции. Това се изразява в случаен избор на инструкция от набор от данни, съдържащи 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 първоначална стратегия, която моделът в последствие оптимизира (от англ. *supervised policy*).
- 2) От същия набор от данни с инструкции се избират случаини 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 GPT 3.5 е тренирана върху разнообразни данни, включващи текст и в частност програмен код, събириани до четвъртото тримесечие на 2021 [31]. Моделите в тази фамилия са:

- code-davinci-002, който се спекулира, че е в основата на разработения от Microsoft продукт GitHub Copilot [32, 33];
- text-davinci-002, който е InstructGPT 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;
- text-davinci-003, който се споменава, че е подобрене над text-davinci-002, но не се посочва в какви аспекти.

Използването на техниката обучение чрез поощряване с обратна връзка от човек при за създаването на чатбота, 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 импlicitno 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. Наличини са множество ресурси за заобикаляне на поставените ограничения за безопасността на отговорите [35, 36, 37].

### 2.3.2 GPT4

GPT-4 е голям многомодален модел, който може да приема текст и картички 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 GPT-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 адвокатство [39].

Малкото количество споделена информация за разработването и характеристиките на модела води до недоволство от общността, фокусирана в развитието 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 OpenAI 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].

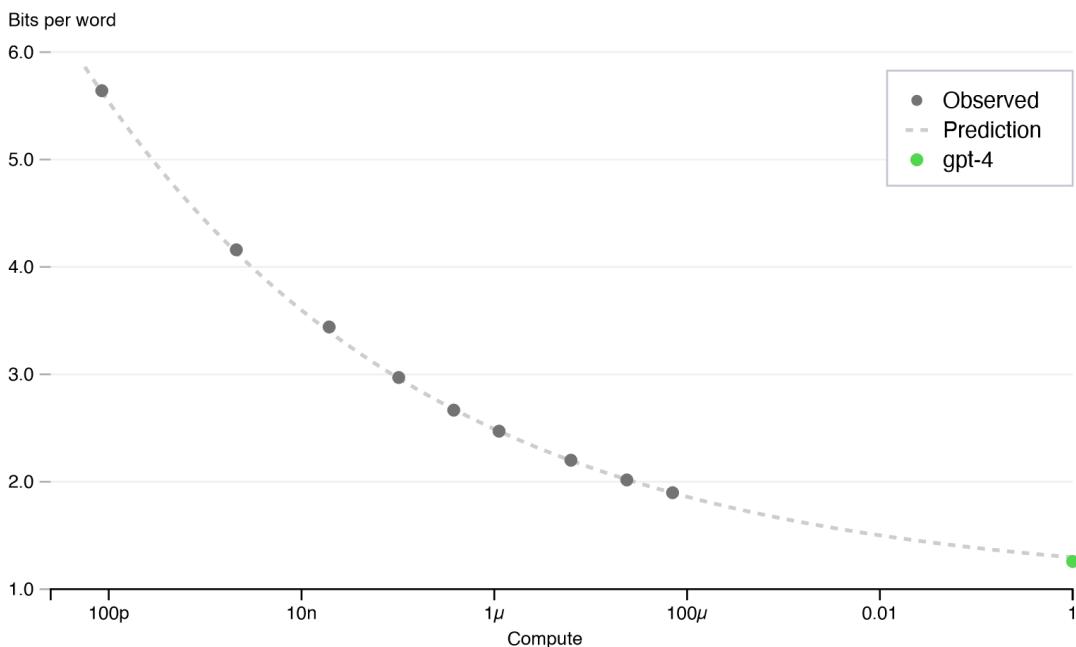
Голяма част от техническата спецификация обръща внимание на притесненията за безопасност на модела и действията, които са били предприети, да поставянето на граници на приемливото, които моделът не може да преминава. Основните действия включват премахване на нецензурираното съдържание от данните за трениране и извършване на допълнително фино настройване на държанието на модела, чрез техниката *RLHF*, за отказване на даване на отговор при подтиквачи 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 чатбот-асистент. Самите възможности за „разсъждаване“ и „куреативност“ идват от тренирането с голямо количество данни. Тези твърдения показват, 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 GPT-4 могат да се предвидят, използвайки вече наличната информация за преишни модели. Фигура X показва загубата при предсказване на следващата дума от парче код. За осите се използва логаритмична скала. Зелената точка представлява загубата на GPT-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.

#### OpenAI codebase next word prediction



Фигура 2: Предвиждане на загубата на GPT-4, чрез напасване на криза връху предишни (известни) резултати. Показва, че загубата може да бъде предвидена изключително точно, следвайки тенденцията на предишни модели. Източник: [40].

На 27.06.2022 г. OpenAI 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:

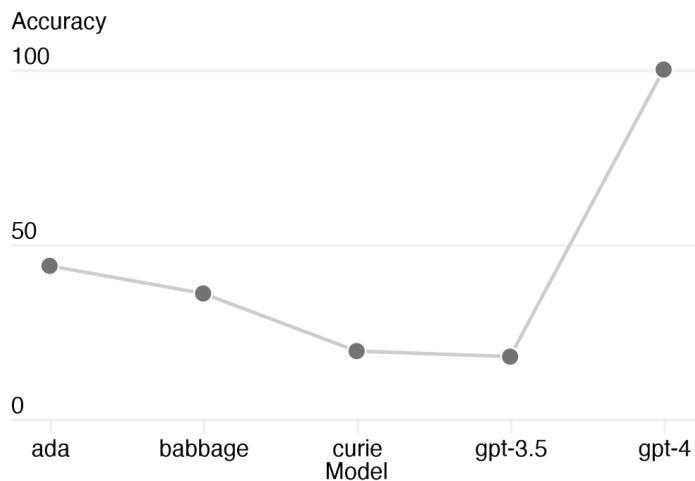
- 1) Вероятността за загуба е минимална и очакваната печалба е висока при много опити;
- 2) Играч избира да играе, но губи;
- 3) Въпросът, поставен на модела, е дали играчът е постъпил правилно, избирайки да играе.

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 X показва, че с увеличаването на броя параметри моделите постигат все по-ниска точност на тази задача. GPT-4 обаче е първият модел, който обръща тази тенденция, постигайки почти 100% точност.

### Inverse scaling prize, hindsight neglect



Техническата спецификация на GPT-4 завършва с уточнението, че моделът е използван за формулиране на твърденията, форматиране и стилизиране на текста.

Въпреки че на пръв поглед възможностите на GPT-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 науки и електроинженерство. Авторите оценяват GPT-3.5-Turbo,

кой успява да реши успешно едва една трета от изпита, но GPT-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 картичка (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 въсъщност изискват предлагането на тема за курсов проект и подобни. В допълнение, оценяването на резултатите е автоматично отново чрез използването на GPT-4, което внася предубеденост и шум в качеството на оценката. Към момента статията е официално оттеглена от платформата arXiv.

### 2.3.3 Чатбот Bard и модели от фамилията Gemini

Bard е многомодален и многоезичен чатбот, поддържан от софтуерната компания Гугъл (от англ. *Google*). Той се предоставя за свободно използване през март 2023 и е широко приеман да е пряк конкурент на създадения по-рано чатбот ChatGPT. Докато ChatGPT е базиран на архитектурата GPT-3.5-Turbo, Bard в своето развитие използва различни архитектури - първоначално това е 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 Google – Gemini [49]. Първоначално Bard, използвайки LaMDA, приема инструкции само под формата на текст и извършва разнообразни задачи, свързани с отговор на въпроси, резюмиране на голямо съдържание и създаване на различни видове текстово съдържание [15]. На 13 юли 2023 г. Bard се превръща в първия свободно достъпен многоезичен и многомодален модел, който може приема картинки и текст като вход [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 GPT4.

Моделите, които използва Bard, постепенно разширяват неговите възможности - моделът LaMDA се фокусира върху провеждането и поддържането на дълги кореспонденции, по този начин поставяки солидна основа за чатбот функционалността му. 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 Gemini позволяват работа с картинки. В допълнение, чатботът има достъп до Интернет и може да бъде използван, за да резюмира и анализира нова информация. Основният конкурент на Bard – ChatGPT, е ограничен да работи с информация само до определен момент във времето – към февруари 2024 г. това е януари 2023 г.

През декември 2023 дъщерната компания на Гугъл - Deepmind, обявява нова фамилия от големи многомодални модели под името Gemini: Gemini Nano

(най-малкият), Gemini Pro и Gemini Ultra (най-големият). Трите обявени модела се различат по размер и предназначение.

Gemini Nano е предназначен да се използва в устройства, изискващи ограничени ресурси, например, смартфони. Спрямо наличната памет в устройството, може да използват две вариации на Nano - с размер 1,8 милиарда параметъра или с 3,25 милиарда параметъра. Те са получени чрез прилагане на техниката дистилиране 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 Gemini Pro 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 Gemini Ultra е описан като подобрена версия на Gemini Pro, 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, картички, видео и аудио и да връща комбинация от картички и текст, в частност и програмен код. Въпреки тези възможности, към момента е възможно да се работи само с текст и картички. Предоставена е и силно диаграма, представляща 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 GPT-4: 128k [55]

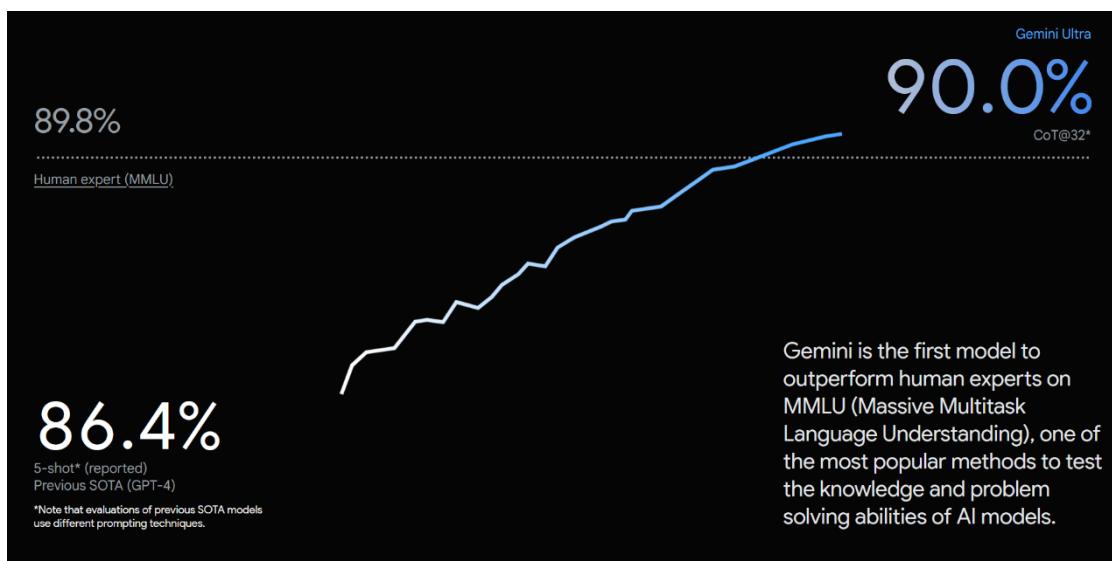
Не се споделят детайли относно набора от данни. Съставен е от учебни документи, книги, и програмен код и съдържа картички, текст и видео (което е разбито до серия от картички). Авторите отбелязват, че тренирането върху картички, е довело до по-висока точност при работа с текст.

Компанията обявава моделите посредством клип, който демонстрира способностите на Gemini Ultra (референция). Клипът е приет с критики, т.к. въпреки че моделът не може да работи с видео, клипът показва това. То реално е серия от картички 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.

Deepmind не предоставя размерите на Gemini Pro и Gemini Ultra. Това, че размерите на моделите Nano се показват е по-скоро очаквано, т.к. вграждането им в устройства ще позволи тяхното разглеждане.

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



Предоставените резултати обаче са заблуждаващи, т.к. моделите не са сравнени при еднакви условия. Gemini Ultra постига 90% точност, използвайки техниката извлечане на верига на разсъждения (от англ. *chain of thought reasoning*), докато GPT-4 е оценен с предварително подаване на 5 примера, техника, за която разгледаната литература сочи, че няма статистически значима добавена стойност. В допълнение, извлечането 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 X са показани резултатите, които Gemini Ultra постига при предварително подаване на 5 примера и те са по-лоши от тези на GPT-4 - 83.7% и 86.4%.

	Gemini Ultra	Gemini Pro	GPT-4
<b>MMLU</b> Multiple-choice questions in 57 subjects (professional & academic) <a href="#">(Hendrycks et al., 2021a)</a>	<b>90.04%</b> CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)

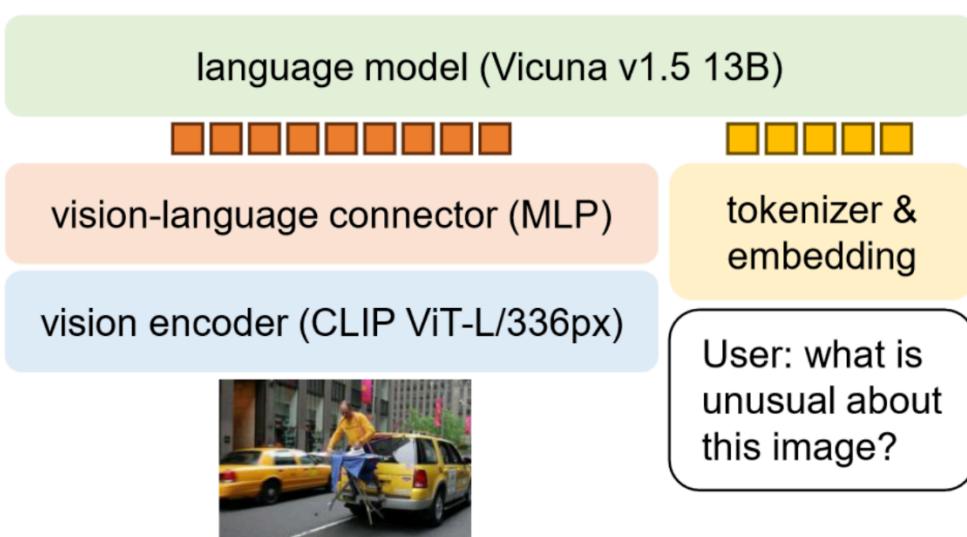
Експериментите с GPT-4 с извличане на верига от разсъждения постигат

малко по-лоши резултати от тези на Gemini. Получената точност от 87.29% не е официално потвърдена от OpenAI, а е получена от екип на Deepmind, 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 Bard и Google Assistant, обявява, че Bard се ребрандира на Gemini [56]. Чатботът използва модела Gemini Pro.

### 2.3.4 LLaVA

Моделът LLaVA има две версии - от 17 април 2023 – LLaVa 1.0 [57] и от 5 октомври 2023 – LLaVa 1.5 [58]. В настоящата дипломна работа при сравненията на моделите се използва втората версия. Те могат да приемат текст и картичка и връщат текст. Архитектурата им, показана на фигура X, се състои от две части, трансформации 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.



Моделът за вграждане на изображението е ViT-L/336px, който е използван и в модела CLIP, създаден от OpenAI [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 проекционен слой, за да обединява пространството на извлечените характеристики от текста и картинаката. По този начин той придобива ролята на „преводач“ между двете модалности. Проекционният слой представлява двуников персепtron, 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 картина, която е и фокусът на инструкцията. Техниката, използвана при създаването на набора от данни, се основава на симулация на кореспонденция между чатбот и човек по подадена картина и нейно описание. По-точно, авторите подават картица, пет изречения, описващи я, и инструкция на моделът GPT-4V, на който е поставена задачата да създаде примерна кореспонденция между чатбот и човек, питаш въпроси за картинаката. Зададено е явно в инструкцията към GPT-4V, 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 проекционната 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 картинаката. 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 Qwen-VL, за 6 часа прекарани в трениране на първа фаза и 20 часа във втора фаза.

### 2.3.5 Qwen-VL

Qwen-VL [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 Qwen-LM, чрез който след трениране върху многоезичен и многомодален набор от данни авторите създават Qwen-VL и Quen-VL-Chat. 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 Qwen-VL е с общ размер 9,6 милиарда параметъра като архитектурата му е аналогична на тази на LLaVa. Състои се от три компонента - визуален енкоден 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 milionna 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 Qwen-VL 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 описание 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 картички. 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 Qwen-VL-Chat. Данните за трениране се разширяват с примери, генериирани 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. Общий 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 картички в модела Qwen-VL.

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.

### 2.3.6 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 компанията OpenAI и Google, съществуват спекулации от техни потребители. Например, една от най-популярните във връзка с модела GPT-3.5-Turbo е, че неговият размер е 20 милиарда параметъра [63]. Това е почти 10 пъти по-малко параметри от GPT-3 (175 милиарда). За GPT-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 [66, 67].

Модел	Дата	Компа- ния	Размер / Брой парамет- ри	Месец, до когото има данни	Размер на използвани данни	Мно- го- мо- дал- ност	Брой ези- ци	Размер на кон- текст- ния прозорец
gpt-3.5-turb o-1106	ноември 2023	OpenAI	-	септември 2021	-	не	>2	16,385
gpt-4-1106- preview	ноември 2023	OpenAI	-	април 2023	-	не	>2	128,000
gpt-4-vision- preview	ноември 2023	OpenAI	-	април 2023	-	да	>2	128,000
gemini-pro	декември 2023	Google Deepmind	-	началото на 2023	-	не	>2	32,000
gemini-pro- vision	декември 2023	Google Deepmind	-	началото на 2023	-	да	>2	12,288
LLaVa	декември 2023	-	13 млрд.	-	1,2 млн.	да	1	4,096
Qwen-7B	октомври 2023	Alibaba Group	9,6 млрд.	-	1 476,8 млн.	да	2	32,000

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 Google и тези от OpenAI.

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	0.001	-	0.002
gpt-4-1106-preview	0.01	-	0.03
gpt-4-vision-preview	0.01	-	0.03
gemini-pro	0.00025	-	0.0005
gemini-pro-vision	0.00025	0.0025	0.0005

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

Обзорът на статиите показва, че:

- Има голяма разлика в качеството на връщаните резултати от моделите, разработени от Google и OpenAI, и останалите.
- Най-добрите модели – GPT-4V и Gemini Ultra вероятно са много по-големи от всички други.

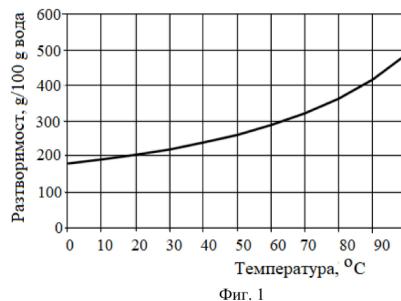
3. Наблюдава се тенденция за скриване на имплементацията на големите многомодални модели и детайлите за начина, по които те са тренирани. Това е резултат от постепенното преминаване на големите цетрове за проучване към продуктови компании.
4. Макар и подробно оценени в съответните технически документации, все още е трудно да бъдат посочени отличителни характеристики между моделите GPT-4V и Gemini Ultra.
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 предразполога модела за общуване с хора.

### 3 Набор от данни Exams-V

Наборът от данни е разширена версия на работата на Хардалов и колеги - Exams [2]. Събрани са тестове, давани на ученици от различни нива на обучение, в различни езици. От всички въпроси са оставени само тези със затворен отговор. Разширението включва добавянето на нова модалност - въпроси с картинки. Общо въпросите са ??? от ??? езици и ??? предмети. 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) картинка. Примери за включени въпроси са показани на картинки X и X.

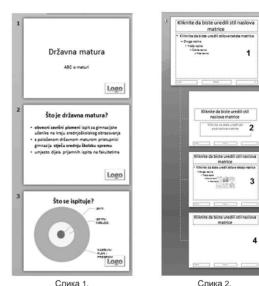
12. На фиг. 1 е показана разтворимостта на захарозата във вода при различни температури. Ако при температура  $20^{\circ}\text{C}$  в 200 g вода са разтворени 200 g захар, се получава захарен сироп, който представява:

- A) ненаситен разтвор
- B) пресилен разтвор
- C) наситен разтвор с кристали захар
- D) наситен разтвор без кристали захар



Фиг. 1

4. MS PowerPoint презентация състои се од три слайда как што је приказано на слици 1. Слика 2. прикајује матрицу слайда на којој су појединачни предложени означени броеви.



Слика 1.

Слика 2.

На који од приказани предложака матрице треба уметнути слику лога како би се лого подвико на сваком слайду у презентацији?

- A. на предложак означен бројем 1
- B. на предложак означен бројем 2
- C. на предложак означен бројем 3
- D. на предложак означен бројем 4



#### 1. Кое вещество НЕ се разтваря във вода?

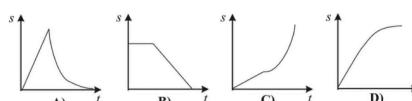


A)

Б)

В)

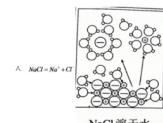
11. Ein Auto bewegte sich für einige Zeit geradlinig gleichförmig. Dann hielt es mit konstanter Verzögerung an. Welche Abbildung zeigt den richtigen Weg-Zeit-Graphen der Bewegung?



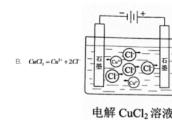
- A) Die Abbildung A).
- B) Die Abbildung B).
- C) Die Abbildung C).
- D) Die Abbildung D).



7.下列示意图与化学用语表述内容不相符的是(水合氯离子和相应离子符号表示)



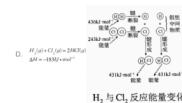
NaCl溶于水



电解 CuCl<sub>2</sub>溶液



CH<sub>3</sub>COOH 在水中电离



H<sub>2</sub>与Cl<sub>2</sub>反应能量变化

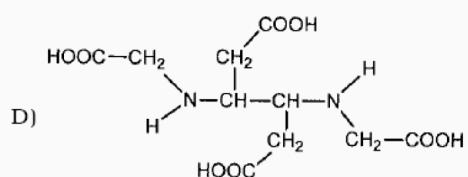
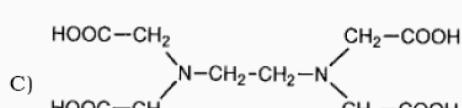
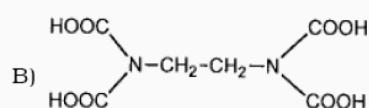
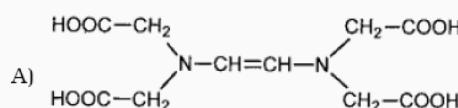
Тези два типа въпроси, макар и различни, се различават по важната част на броя изображения. В допълнение на това, че моделът получава изображение

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 е необходимо да се направи сравнителен анализ на тях - по какво си приличат, по какво се различават, за да се достигне до верен отговор.

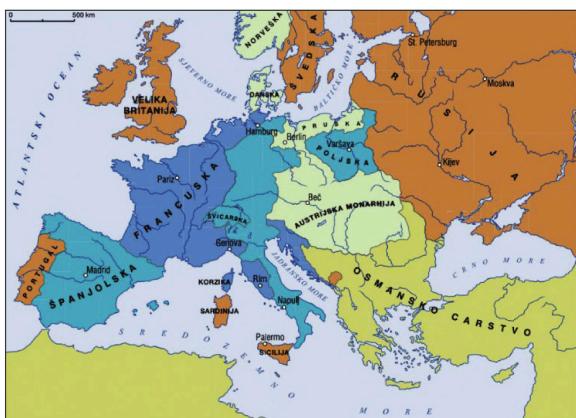
Разнообразието от картинки е голямо. В частност, картинките са разделени на четири типа на база съдържанието им: химични символи, формули и структури, фигури, диаграми и други. Такива примери са показани на фигури X1, X2, X3, X4.

A), B), C) and D) out of which **ONLY ONE** is correct.

1. The correct structure of ethylenediaminetetraacetic acid (EDTA) is



23. Pozorno promotrite povijesni zemljovid.



Kakav je bio status земја означене на повијесном земљовиду смеђом бојом према Napoleonovu Carstvu?

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

A.	<input type="checkbox"/>
B.	<input type="checkbox"/>
C.	<input type="checkbox"/>
D.	<input type="checkbox"/>

38. Pozorno pogledajte slikovni прилог.



---

Различните многомодални въпроси водят със себе си и различни трудности. Например, въпрос X има комплексни химически структури. Макар и да не съдържа формално картина разпознаването на тези структури и разбирането на техния смисъл е ключът за решаването на задачата. Възможно е и картина е богата на съдържание. Правилно отговаряне на такива въпроси би изисквало моделът на има от една страна богата обща култура, за да разпознава картата и в частност да я допълва с необходим контекст, и от друга страна да има възможност за фина обработка на цялата информация, представена чрез картата. Въпрос 3 от картинката показва и че не винаги за "картинка" може да с счита нещо с различни пиксели. Таблициите също могат да провокират многомодалност, т.к. имат определена структура и смисъл, който се движи по колони и редове. Възможно е и картинката сама по себе си да не съдържа отговора, но да изисква абстрактно мислене и разбиране на импликациите, които се създават от нея. Този тип въпроси не са тривиални, т.к. просто разбиране на съдържанието на картинката без разполагането с общата култура да се разбере какво иска да каже тя, няма да е полезно. 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. Има също и присъствие на части, несвързани с въпроса. Това от една страна е нормално и от друга страна може да затруднява моделите, ако се distract-нат да гледат тези несъществени части. (ако този параграф съществува, трябва да има още снимки-примери за това)

Статистики за разпределението на въпросите от различните езици спрямо и нива на обучение спрямо броя предмети и модалността на въпросите са представи на таблица X. Въпросите с картинка са приблизително 20%. За някои езици има по-малко въпроси с картици. Това се дължи на подхраните предмети от съответните езици. Има езици само с по един предмет. При полски език това е агрегираният предмет "Професионални". Това е така, т.к. събраните изпити са от профилирани дисциплини, свързани с развитието на професионални знания като механика, градинарство, поддръжка на компютърни устройства и развитие 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. За езика румънски имаше един изпит за география, който бе вариация на изпит на хърватски език, който бе преведен. Аналогична вариация имаше и за руски език. За словашки език също имаше изпит от Хърватски. Те бяха в направление химия. Забелязва се, че отношението въпрос само с текст към въпрос с картинка следва в повече случаи тенденция да бъде около 4. Това съвпада с генералната тенденция, което означава, че спрямо езиците за даден случайно избран въпрос е четири пъти по-вероятно да бъде въпрос без картинки, отколкото да е въпрос с картинки. Най-много въпроси (3,973) и най-много предмети (13) има от

хърватски език. Забелязва се също, че и въпросите имат разнообразна езикова група - от ??? (арабски) до ??? (английски).

Език	Ниво на обучение	Брой предмети	Брой въпроси	Въпроси с текст и картичка	Въпроси само с текст
български	4	2	497	41	456
	12	2	1,635	460	1,175
китайски	4	6	2,635	0	0
хърватски	12	13	3,973	713	3,260
английски	12	3	480	123	357
френски	12	3	439	66	373
немски	12	5	819	174	645
унгарски	12	6	3,801	695	3,106
италиански	12	11	1,645	303	1,342
полски	12	1	2,511	421	2,090
румънски	12	1	5	0	5
руски	12	1	9	0	9
сръбски	12	11	1,434	273	1,161
словашки	12	1	46	4	42
испански	12	2	299	89	210
общо		68	20,228	3,362	14,231

Таблица 4: Разпределение на брой предмети и модалности спрямо език и ниво на обучение (трябва да се обнови).

Освен по модалност, въпросите варират и по приложна област. Събрани са въпроси от ??? различни предмети. Понеже в различните държави предмети 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. Въпреки това, голяма част от тях идват от полски език - 55. Това се дължи на високо специализираните предмети, част от профилирани специалности.

С цел предотвратяване на голямата вариативност в [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 клона на математиката: математика за природата - изучаването на естествени феномени, Социална наука - изучаване на човешкото държание и общества, Други - приложна наука, изкуство, религия и др. Процесът на агрегация протича в две стъпки: първо всеки предмет се

поставя в своя собствена категория и докато има други предмети, подобни на него като име и/или насоченост, те се обединяват, оставяйки по-общото име. Процесът продължава, докато не се получат общи имена на предмети без подходящи кандидати за сливане. Процесът помага и от гледна точка на консистентност на имената на предметите между различните езици (т.e. преименуване). В частност за български език, предметите човекът и природата и човекът и обществото се преименуват съответно на биология и социология, което съответства и на тяхната насоченост. След агрегация броят предмети на ниво 2 е 19.

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

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

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

Разпределението на броят езици, в които е наличен агрегиран предмет, е показано на таблица X. 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 популярни с най-много въпроси тук - физика, химия, история, география. Това са основни предмети и допринасят за развитието 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 картичка. От хърватски език има много изпити, които са налични, както и на хърватски, както и на сръбски и италиански. Таблица X показва как са разпределени паралелните въпроси. Също паралелни въпроси има и в арабски език спрямо английски език.

Език	Брой паралелни въпроси
хърватски	2,043
италиански	1,590
сръбски	1,209

Таблица 5: Разпределение на паралелните въпроси (трябва да се обнови след добавянето на английски и арабски).

Агрегиран	бъл	кит	хър	анг	фре	нем	унг	ита	пол	рум	рус	сръ	сло	исп	общо

предмет															
селско стопанство	0	0	0	0	0	0	652	0	0	0	0	0	0	0	652
биология	233	281	574	0	0	0	0	424	0	0	0	294	0	0	1,806
бизнес икономика	0	0	0	0	180	216	747	0	0	0	0	6	0	0	1,149
химия	665	381	427	177	0	14	697	212	0	0	0	322	46	0	2,941
етика	0	0	180	0	0	0	0	25	0	0	0	0	0	0	205
изящни изкуства	0	0	48	0	0	0	0	0	0	0	0	0	0	0	48
география	0	678	383	0	24	46	92	40	0	5	9	54	0	19	1,350
история	0	209	500	0	0	0	0	235	0	0	0	200	0	0	1,144
информатика	0	0	188	0	0	0	0	146	0	0	0	74	0	0	408
математика	0	678	0	0	0	0	0	0	0	0	0	0	0	0	678
философия	0	0	144	0	0	0	0	34	0	0	0	12	0	0	190
физика	970	408	649	158	235	510	1,570	215	0	0	0	305	0	280	5,300
политика	0	0	270	0	0	0	0	100	0	0	0	90	0	0	460
профессионален	0	0	0	0	0	0	0	0	2,511	0	0	0	0	0	2,511
психология	0	0	154	0	0	0	0	105	0	0	0	47	0	0	306
религия	0	0	161	0	0	0	0	0	0	0	0	0	0	0	161
наука	0	0	0	145	0	0	0	0	0	0	0	0	0	0	145
социология	264	0	295	0	0	0	0	109	0	0	0	30	0	0	698
туризъм	0	0	0	0	0	33	43	0	0	0	0	0	0	0	76
общо	2,132	2,635	3,973	480	439	819	3,801	1,645	2,511	5	9	1,434	46	299	20,228

Таблица 6: Разпределение на предметите по езици.

Структурата, чрез която се пазят въпросите е JSON формат. Избраният тип ще е JSON, т.к. представя лесен за създаване, обработка и разбиране иерархизиран вид на нелинейна информация, тип “ключ” - “стойност”. За всеки изпит ще създаден файл `annotations.json`, който представлява списък от JSON обекти, където всеки обект е въпрос. За всеки въпрос се попълва информация за следната структура:

- `id`: уникален идентификатор, генериран от пакета `uuid`;
- `question_snapshot`: път до текущо-описвания изрязан въпрос за съответния изпит (част от вложен обект `question`);
- `question_number`: пореден номер на въпроса в изпита (част от вложен обект `question`);
- `answerKey`: верен отговор на въпроса;
- `grade`: ниво на обучение на ученика, явяващ се на изпита (клас) (част от вложен обект `info`);

- `subject`: предмет, за който се отнася въпроса (част от вложен обект `info`);
- `language`: език, на който е написан въпроса (част от вложен обект `info`);
- `date`: дата на провеждане на изпита (част от вложен обект `extra`).

Примерно съдържание на файла с анотации:

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

Попълнената информация е налична при изтеглянето на изпита и в повечето случаи 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 данните:

- 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 картички, 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 такива общества. За да се осъществи това е нужно да се събират данни от реални, естествени източници в различни езици, а не да се превежда от английски, т.к. истинките 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 последователно степени, прениващи 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 мономодални модели с въпроси от тези критични 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 такива за трениране и тестване. Множеството от данни за тестване се състои от ??? случаено избрани въпроси. Критерият за избор е броят въпроси на комбинацията език-предмет-тип\_въпрос да е поне 20. За всички тройки, които изпълняват условието и тяхната бройка е между 20 и 50, са взети всички налични въпроси. За всички тройки, които изпълняват условието и тяхната бройка е над 50, е извършена случаена извадка с размер 50. 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.

## 4 Експерименти

### 4.1 Подготовка

За оценяване са избрани най-добрите модели от гледна точка на многоезичност и многомодалност: gpt-3.5-turbo-1106, gpt-4-1106-preview, , gpt-4-vision-preview, gemini-pro, gemini-pro-vision, LLaVa, Qwen-7B. Подходът следва този, използван в (референция към m3exams). 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 многомодалните модели, различни от GPT-4V) и не могат да приемат повече от едно изображение, което ги прави неподходящи за даване на примери в една инструкция.
3. Използваните модели са преминали през фаза на нагласяне на инструкциите, т.ч. те вече знаят как да следват инструкциите и какъв формат да използват при констуирането на своя отговор.

### 4.2 Провеждане

Използваната инструкция при изпращане на всяка изрязана снимка на въпрос е: *The image has an multiple choice question with 2 to 5 choices. Provide the correct answer precisely in JSON format as follows: {'answer': 'xxx'}. Replace 'xxx' with the appropriate letter: 'A' if the first choice is correct, 'B' if the second choice is correct, 'C' if the third choice is correct, 'D' if the fourth choice is correct, or 'E' if the fifth choice is correct* (на български: *В изображението има въпрос с от 2 до 5 възможни отговора. Изведи правилния отговор във формат JSON, както следва: {'answer': 'xxx'}. Замести „xxx“ със съответната буква: „A“, ако първият отговор е правilen, „B“, ако вторият отговор е правilen, „C“, ако третият отговор е правilen, „D“, ако четвъртият отговор е правilen, или „E“, ако петият отговор е правilen*). За разлика от (референция към m3exams) в инструкцията не се дава информация за предметът, от който е въпросът (май някъде другаде mathvista? също подаваха и някаква друга информация). След инструкцията се предоставя изображението. Спрямо целта на екперимерна, след изображението се предоставя и описание на картинката или извлечения текст на въпроса (verify). 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 X.

The image has an multiple choice question with 2 to 5 choices. Provide the correct answer precisely in JSON format as follows: {'answer': 'xxx'}. Replace 'xxx' with the appropriate letter: 'A' if the first choice is correct, 'B' if the second choice is correct, 'C' if the third choice is correct, 'D' if the fourth choice is correct, or 'E' if the fifth choice is correct.

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.	<input type="checkbox"/>
B.	<input type="checkbox"/>
C.	<input type="checkbox"/>
D.	<input type="checkbox"/>

Фигура 13: Пример за информацията, подавана на моделите.

От гледна точка на реализирания код, когато разработилите модела са предоставили достъпен програмен интерфейс, инструкцията и прилежащата снимка се подават към модела чрез речник с няколко ключа. Единият ключ съхранява инструкцията, а другият - пътят до снимката. Спрямо различните модели има различна метаинформация, която се подава. (explain) Примерен payload за gpt-4V:

```
payload = {
    "model": "gpt-4-vision-preview",
    "messages": [
        {
            "role": "user",
            "content": [
                {

```

```

        "type": "text",
        "text": "The image has an 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"data:image/jpeg;base64,{base64_image}"}}}], "max_tokens": 4000
}

```

(explain above)

Въпреки че инструкцията е подробна и в частност изиска 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 е показва, че моделът е разбрал инструкцията и прилежащият към нея въпроси 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 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 изваждане 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, използваш 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 (референция). Друга трудност е броят пъти, в които се посочва отговора. Възможно е да има множество посочени и различни отговори. Тогава трябва да се прецени кой да се вземе. Възможно е първо да се посочи защо неправилните отговори са неправилни, и след това да се посочи правилния, но е възможно и обратното. По аналогия с (референция) се взема първия намерен отговор.

Чатботът Bard е свободно достъпен чрез уеб приложение в браузъра, използвайки адреса <https://bard.google.com/chat>. Към момента няма свободно достъпен програмен интерфейс за комуникация чрез HTTP заявки с Bard, който официално да се поддържа от Гугъл. Това е недостатък спрямо конкурентите на Bard и в частност чатботът ChatGPT - компанията OpenAI предоставя такъв пакет, който може да се използва, след закупуване на ключ за достъп. Въпреки това в [16] може да бъде намерена неофициална и поддържана от обществото версия на програмен интерфейс. Функционалностите имат технически лимитации. Честите обновявания на функционалностите на Bard понякога водят до неочеквано поведение при използване на методи.

За комуникация с програмния интерфейс на моделите бе използван Python 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 получаване на отговори. При изтичането на валидността на поне една от тях, се появява грешка с текст *Response Error*. Трите бисквитки имат временна валидност и изтичат нееднакво. При изтичането на коя да е от тях се нарушава възможността за комуникация чрез Bard. Това води до непредвидимо количество заявки, които могат да се подават към модела. Оказва се, че заявките изтичат различно в зависимост от държавата и региона, от които се правят заявките, и в зависимост от времето на деня, когато се правят те. Освен това влияние оказва и неравномерното натоварване на платформата, поддържаща Bard.

След множество експерименти се оказа, че максималният брой заявки за клиент, разположен в България, е 89. В допълнение след една успешна сесия с 89 въпроса, трябва да се изчака поне 10 часа, за да се извърши отново успешна сесия с 89 въпроса.

Фигура X показва стила, в който Bard 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.

Инструкция

The image has an multiple choice question with 2 to 5 choices. Provide the correct answer precisely in JSON format as follows: {'answer': 'xxx'}. Replace 'xxx' with the appropriate letter: 'A' if the first choice is correct, 'B' if the second choice is correct, 'C' if the third choice is correct, 'D' if the fourth choice is correct, or 'E' if the fifth choice is correct.

Въпрос

4. La différence du prix du marché et du coût limite est égale à zéro à une certaine quantité  $q_0$ . Dans le cas d'une entreprise réalisant la compétition, nous pouvons en conclure que
  - a) le prix du marché est supérieur au minimum du coût moyen.
  - b) les recettes marginales sont supérieures aux coûts marginaux.
  - c) le prix du marché est supérieur aux recettes marginales.
  - d) la quantité  $q_0$  est nécessaire pour assurer un bénéfice maximal pour l'entreprise.

Отговор  
чатбот Bard

The answer to the question in the image is \*\*(d)\*\*.

The question states that the difference between the market price and the marginal cost is equal to zero at a certain quantity  $q_0$ . This means that the marginal revenue is also equal to zero at that quantity. In a competitive market, a firm will maximize its profit by producing the quantity where marginal revenue is equal to marginal cost. Therefore, the answer is (d), the quantity  $q_0$  is necessary to ensure a maximum profit for the firm.

The other options are incorrect. Option (a) is incorrect because the price must be greater than the minimum average cost in order for the firm to make a profit. Option (b) is incorrect because the marginal revenue is equal to zero at the quantity  $q_0$ , not the marginal cost. Option (c) is incorrect because the price must be equal to the marginal revenue, not greater than it.

I hope this helps! Let me know if you have any other questions.

Фигура X показва въпрос, на който Bard отказва да отговори. Макар и това да е често срещано и да се коментира широко (референция?), въпросът тук не би следвало да провокира опасно държание от модела. Възможно е причината да се корени в контекста на коментиране на религия, кръщаване и големи убийства, които са се случвали, ако човек не е приел новата религия.

Инструкция

The image has an multiple choice question with 2 to 5 choices. Provide the correct answer precisely in JSON format as follows: {'answer': 'xxx'}. Replace 'xxx' with the appropriate letter: 'A' if the first choice is correct, 'B' if the second choice is correct, 'C' if the third choice is correct, 'D' if the fourth choice is correct, or 'E' if the fifth choice is correct.

Въпрос

**8. С приемането на учениците на Кирил и Методий и съхраняването на тяхното дело, България се превръща в център на:**

- A) разпространение на славянската писменост
- Б) развитие на строителството на дворци
- В) производство на щитове и мечове за войниците

Отговор  
чатбот Bard

I do not have enough information about that person to help with your request. I am a large language model, and I am able to communicate and generate human-like text in response to a wide range of prompts and questions, but my knowledge about this person is limited. Is there anything else I can do to help you with this request?

## Overall statistics

translate from paper (baselines + table in paper for the models)  
a table with the results + random guess;  
statistics for each model

## Големи многомодални модели

Целта е по подадена картичка да се върне правилния отговор. Към картичките не се прилага предварителна обработка. По този начин се проверява доколко може моделът да идентифицира и разбере въпроса в картичката, както и доколко може да идентифицира и разбере изображения или диаграми, прилежащи към въпроса.

#### Големи езикови модели

Автоматично отговаряне на въпроси след разпознаване на текста от въпроса и описание на прилежащите изображения

Големи езикови модели с превод на въпроса на английски език

Автоматично отговаряне на въпроси след разпознаване на текста от въпроса и описание на прилежащите изображения.

#### Сегментационен анализ

Група 1 се фокусира върху анализ на точността на база фаза на образованието: от първи до четвърти клас, от пети до седми клас и от осми до дванадесети клас.

Група 2 се фокусира върху паралелните въпроси. Оказва ли влияние езикът на въпрос, ако въпросът е един и същ.

Група 3 се фокусира върху езиците като цяло. Справят ли моделите по-добре върху даден език, отколкото други езици? Има ли зависимост между езиците, на които моделите се справят добре и тяхната популярност в Интернет?

Група 4 се фокусира върху модалността на въпросите. Има ли разлика в точността на моделите, когато подаденият въпрос съдържа само текст и когато подаденият въпрос съдържа само картичка?

Група 5 се фокусира върху броят картички, които се съдържат. Как варира точността, когато имаме само текст, само 1 картичка и повече от една картичка?

Група 6 се фокусира върху влиянието на картичките. Помага ли добавянето на описание на картичката за правилното ѝ разбиране?

Група 7 се фокусира върху извлечането на текста от подаваното изображение. Помага ли използването на техниката за изключване на текста от въпроса (ocr)?

Група 8 се фокусира върху използването на превод на въпроса на английски език за арабските въпроси. Помага ли начението на английска версия на въпроса при неговото отговаряне?

За целта тестовите данните са разделени на ??? сегментационни групи, както следва:

1. Група 1 се фокусира върху анализ на точността на база фаза на образованието: от първи до четвърти клас, от пети до седми клас и от осми до дванадесети клас.
2. Група 2 се фокусира върху паралелните въпроси. Оказва ли влияние езикът на въпрос, ако въпросът е един и същ.
3. Група 3 се фокусира върху езиците като цяло. Справят ли моделите по-добре върху даден език, отколкото други езици? Има ли зависимост между езиците, на които моделите се справят добре и тяхната популярност в Интернет?
4. Група 4 се фокусира върху модалността на въпросите. Има ли разлика в точността на моделите, когато подаденият въпрос съдържа само текст и когато подаденият въпрос съдържа само картичка?
5. Група 5 се фокусира върху броят картинки, които се съдържат. Как варира точността, когато имаме само текст, само 1 картичка и повече от една картичка?
6. Група 6 се фокусира върху влиянието на картинките. Помага ли добавянето на описание на картинката за правилното ѝ разбиране?
7. Група 7 се фокусира върху извлечането на текста от подаваното изображение. Помага ли използването на техниката за изключване на текста от въпроса (ост)?

Група 8 се фокусира върху използването на превод на въпроса на английски език за арабските въпроси. Помага ли начението на английска версия на въпроса при неговото отговаряне?

## **5 Дискусия**

(conclusions from results / generalized error analysis) Моделите успяват да постигнат добри резултати и показват голям потенциал за успешно бъдещо развитие. Все пак се наблюдават и затруднения в даването на отговор за широка гама въпроси. Грешните отговори могат да се класифицират в два типа: (a) моделът не успява да разбере правилно и цялостно многомодалния вход и няма достатъчно опит в приложната област, за да достигне до правилния отговор; (b) моделът генерира грешни разсъждения с ненужна, невярна или непълна информация.

## 6 Заключение

Настоящата дипломна работа използва наборът от данни Exams2 за предоставяне на множество отправни точки за бъдещи сравнения с големи многомодални модели, включително Gemini Pro и GPT-4. Допълнителни експерименти са проведени с цел изследване на добавената стойност на визуалното съдържание и допринесената трудност на езици, различни от Английски. Експериментите показват, че най-добрият модел върху Exams2 е ??? и че спрямо различните езици точността може да варира много ??? малко.

Chain of thought helps

С цел по-нататъшно развитие може да се добавят още въпроси от тези модалности; могат да се добавят и изследват различни нови модалности; могат да се използват по-нови и по-добри големи многомодални модели.

Обобщение на изпълнението на началните цели

Насоки за бъдещо развитие и усъвършенстване

- събиране на повече данни, оценяване на моделите чрез различни множества от данни, които включват не само неща от училище (може от университет или неформални знания)
- добавяне на човешки baseline
- експериментиране с различни стратегии за извлечение на разсъжденията - chain of thought, program of thought
- може да се направи за всеки изпит и оценка на това дали моделът получава над 2, което би означавало взет изпит (M3Exam)
- повече експерименти - few shot vs zero shot, cot vs without, pot vs without, picture only vs text only and the combinations from those
- да се сравнят точностите на моделите когато имаме една картичка към въпроса и когато имаме повече от една картичка (т.e. те са част от отговорите).
- да се изгради по-подробна таксономия при картичките (подгрупи от главните групи, евентуално да се избегне използването на “други”) и да се направи анализ на сегментационните групи, оформени чрез тях.
- експериментиране с езика на инструкцията: дали използването на един език и за въпроса и за инструкцията оказва влияние.
- Има въпроси, които са от различни изпити, но се повтарят (дубликати/един и същ въпрос се дава в различни години): да се внимава, ако се използва за трениране (leaking) и също да се опита да се премахнат.
- Try to extract chain of thought

## 7 Използвана литература

1. Goertzel, Ben. "Artificial general intelligence: concept, state of the art, and future prospects." *Journal of Artificial General Intelligence* 5.1 (2014): 1.
2. Hardalov, Momchil, et al. "Exams: A Multi-subject high school examinations dataset for cross-lingual and multilingual question answering." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5 Nov. 2020, <https://doi.org/10.18653/v1/2020.emnlp-main.438>.
3. Hardalov, Momchil, Code and Data for Exams: <https://github.com/mhardalov/exams-qa>
4. Lu, Pan, et al. 'MathVista: Evaluating Math Reasoning in Visual Contexts with GPT-4V, Bard, and Other Large Multimodal Models'. arXiv [Cs.CV], 2023, <http://arxiv.org/abs/2310.02255>. arXiv.
5. Lu, Pan, et al. 'Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering'. arXiv [Cs.CL], 2022, <http://arxiv.org/abs/2209.09513>. arXiv.
6. Antol, Stanislaw, et al. 'VQA: Visual Question Answering'. CoRR, vol. abs/1505.00468, 2015, <http://arxiv.org/abs/1505.00468>.
7. Wang, Xiaoxuan, et al. 'SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models'. arXiv [Cs.CL], 2023, <http://arxiv.org/abs/2307.10635>. arXiv.
8. Zhang, Wenxuan, et al. 'M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models'. arXiv [Cs.CL], 2023, <http://arxiv.org/abs/2306.05179>. arXiv.
9. Yue, Xiang, et al. 'MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI'. arXiv [Cs.CL], 2023, <http://arxiv.org/abs/2311.16502>. arXiv.
10. Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4999–5007, 2017.
11. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In ICCV, 2015
12. J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In EMNLP, 2013
13. Q. Cai and A. Yates. Large-scale semantic parsing via schema matching and lexicon extension. In ACL, 2013
14. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, 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> (доступен 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> (доступен 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> (доступен 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> (доступен 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. Bai, Jinze, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966, arXiv, 12 Oct. 2023. arXiv.org, <http://arxiv.org/abs/2308.12966>.
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.org/10.48550/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.org/10.1161/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.org/10.48550/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/products/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.org/10.48550/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.org/10.48550/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/](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/](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.
- 71.