|  | **Софийски университет „Св. Кл. Охридски”**  Факултет по математика и информатика  *Катедра „Софтуерни технологии”* |  |
| --- | --- | --- |

**ПРЕДДИПЛОМЕН ПРОЕКТ**

на тема

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

Дипломант: **Симеон Емилов Христов**

Специалност: **Софтуерни технологии**

Факултетен номер: **6MI3400191**

Научни ръководители:

проф. д-р Преслав Наков, катедра „Обработка на естествен език“, Университет „Мохамед Бин Зайед по Изкуствен интелект“, Обединените Арабски Емирства

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**](#_heading=h.bxh3qej6es9m)

[**2. Обзор на съществуващи набори от данни 2**](#_heading=h.5bdudfz63zu9)

[2.1. Основни дефиниции 2](#_heading=h.j3ant12jhz03)

[2.2. Набор от данни Science QA 2](#_heading=h.3epl3g5kxmdu)

[2.3. Набор от данни Textbook QA 2](#_heading=h.3b09myx67b2q)

[2.4. Набор от данни Visual QA 2](#_heading=h.hiiov8xw4tw6)

[2.5. Набор от данни SCIBENCH 2](#_heading=h.bj5z9b7cz7p0)

[2.6. Набор от данни MathVista 3](#_heading=h.vz7yg2rwrlar)

[2.7. Набор от данни MMMU 3](#_heading=h.sin0tq8o6o5t)

[2.8. Набор от данни M3Exam 3](#_heading=h.wb2ehw5lyd5s)

[2.9. Набор от данни Exams1 3](#_heading=h.6xm1bycqwoi)

[**3. Наборът от данни Exams2 3**](#_heading=h.o08rd0ex8r9t)

[3.1. Събиране на данните 3](#_heading=h.ktxbf6ipjts3)

[3.2. Предварителна обработка 3](#_heading=h.spnlfjfl2u2d)

[3.3. Анотиране 3](#_heading=h.rl6186g7snao)

[3.4. Валидиране 4](#_heading=h.v4kq47fhcf08)

[3.5. Разбиване на множества за трениране и тестване 4](#_heading=h.8djnlxu7ot29)

[**4. Анализ 4**](#_heading=h.j9l7icyymono)

[1. Езиково разнообразие 4](#_heading=h.fpukny7exdm8)

[2. Обхванати приложни области 4](#_heading=h.8m18lgwpu1qc)

[3. Комплексност на въпросите 4](#_heading=h.vbzjn8dwqc1h)

[**5. Отправна точка за бъдещи сравнения 4**](#_heading=h.4wsmk0efuek8)

[**6. Заключение 4**](#_heading=h.ozibumu7qdzp)

[**7. Използвана литература 4**](#_heading=h.efl9pwdvp34b)

# Увод

Големите езикови модели набират популярност в различни аспекти на човешката дейност. Отличителна тяхна черта е даването на кратък, точен и релевантен отговор на широка гама въпроси. Многомодалните многоезични езикови модели са вид езикови модели, който може да се използва в образователната система за получаване на отговор на въпроси от затворен тип.

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

Целта на преддипломния проект е да се представи многомодален и многоезичен набор от данни, който може да бъде използван за оценяване на големи многомодални модели при отговаряне на въпроси, давани на ученици от различни нива на обучение.

# Обзор на съществуващи набори от данни

## Основни дефиниции

image\_type, text\_type, General purpose visual question answering

## Набор от данни Science QA

## Набор от данни Textbook QA

## Набор от данни Visual QA

## Набор от данни SCIBENCH

## Набор от данни MathVista

## Набор от данни MMMU

## Набор от данни M3Exam

## Набор от данни Exams1

# Наборът от данни Exams2

## Събиране на данните

## Предварителна обработка

Създаване на подходяща файлова структура, в която да са достъпни подбрани изпити (в pdf формат), давани на ученици в различни държави.

Преобразуване на страниците на pdf файловете в картинки

## Анотиране

Използване на инструмент за анотиране за очертаване на границите на въпросите и маркиране на всеки въпрос в един от два класа: само текст или текст и картинка.

Изрязване на всеки въпрос и създаване на анотация (в json формат) за всеки pdf файл. Анотацията включва информация за предмета, по който се провежда изпита, както и данни за всеки въпрос (път до изрязания въпрос, верен отговор, номер на въпроса, и т.н.).

## Валидиране

Създаване на програмен код, който проверява и валидира формата (т.е. структурата) на всяка анотация.

## Разбиване на множества за трениране и тестване

train / test, test-mini

# Анализ

## Езиково разнообразие

## Обхванати приложни области

## Комплексност на въпросите

# Отправна точка за бъдещи сравнения

# 

Използване на Bard за получаване на отправна точка за бъдещи сравнения

# 

# Заключение

Обобщение на изпълнението на началните цели

Насоки за бъдещо развитие и усъвършенстване

# Използвана литература

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.

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.

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.

Antol, Stanislaw, et al. ‘VQA: Visual Question Answering’. CoRR, vol. abs/1505.00468, 2015, <http://arxiv.org/abs/1505.00468>.

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.

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.

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.