

# Предложение

## за

възлагане на преддипломен курсов проект  
към катедра “Компютърна информатика”,  
ФМИ, СУ “Св. Климент Охридски”

**От дипломант:** Симеон Емилов Христов, факултетен № 6MI3400191, специалност „Информатика“, магистърска програма „Изкуствен интелект“

### 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“

**Тема на преддипломния проект:** Създаване на многомодален и многоезичен набор от данни за автоматично отговаряне на въпроси, давани на ученици от различни нива на обучение

### Анотация:

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 да 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.

### Цел на преддипломния проект:

Създаване на многомодален и многоезичен набор от данни за автоматично отговаряне на въпроси, давани на ученици от различни нива на обучение, и прилагане на основни езикови модели за получаване на отправна точка за бъдещи сравнения.

**Задачи, произтичащи от целта:**

- Обзор на съществуващи подходи и формиране на множества от подобни типове множества от данни.
- 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.
- Използване на инструмент за аотиране 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.).
- Създаване на програмен код, който проверява и валидира формата (т.е. структурата) на всяка анотация.
- Използване на Bard за получаване на отправна точка за бъдещи сравнения.

**Ограничаващи/облекчаващи условия:****Срок за изпълнение:**

Дата:

Заявител:

/студент/

/научен р-л/