



Финансирано от
Европейския съюз

Следващо Поколение ЕС



План за възстановяване и
устойчивост



Република България

ПРОТОКОЛ ЗА РАЗРАБОТВАНЕ НА ЛАБОРАТОРЕН ПРОТОТИП

на дигитален продукт (езиков ресурс)

НАБОР ОТ ДАННИ MMLU-BG ЗА ТЕСТВАНЕ НА ГОЛЕМИ ЕЗИКОВИ МОДЕЛИ,

планиран резултат от

Работен пакет 6. Тестване и оценка на фино настроените големи езикови модели
по проекта „Инфраструктура за фина настройка на предварително обучени големи езикови модели“, договор No ПБУ – 55 от 12.12.2024 г. /BG-RRP-2.017-0030-C01/
от Механизма за възстановяване и устойчивост за изпълнение на инвестиция по C2I2 „Повишаване на иновационния капацитет на Българската академия на науките (БАН) в сферата на зелените и цифровите технологии“ от Плана за възстановяване и устойчивост

Протокол: IFGPT-2025-DATA-002

Версия на протокола: 1.0

Име и версия на продукта: **MMLU-BG version 1.0**

Описание на продукта: **Набор от езикови данни за оценка на големи езикови модели**

Ключови думи: **големи езикови модели, MMLU, оценка, набор от данни**

Keywords: **LLMs, MMLU, evaluation, dataset**

Ръководител: **Светла Коева**

Автори: **Светла Коева, Тинко Тинчев, Ивелина Стоянова, Валентина Стефанова, Светлозара Лесева, Цветана Димитрова, Мария Тодорова, Христина Кукова, Михаела Московска**

Текущо ниво на техническа готовност на продукта: **TRL 4**

Дата: **15.09.2025 г.**

Общо описание на оригиналния набор от данни MMLU

Наборът от данни **MMLU-BG** е версия на набора от данни **MMLU** (Hendrycks et al., 2021) на български език. MMLU (Hendrycks, Dan; Burns, Collin; Basart, Steven; Zou, Andy; Mazeika, Mantas; Song, Dawn; Steinhardt, Jacob (2021). Measuring Massive Multitask Language Understanding. ICLR. <https://arxiv.org/abs/2009.03300>) е един от най-широко използваните набори от данни за оценка на големи езикови модели (LLM benchmarks). MMLU съдържа 15,908 въпроса, които изискват избор на верен отговор измежду няколко отговора. Обхванати са 57 различни тематични

области от хуманитарните науки, социалните науки, природните, точните и технологичните науки (STEM – наука, технологии, инженерство и математика) и други области. Всеки въпрос в MMLU е предназначен да тества модела върху знания от различните нива на образование или експертиза – на ниво начално образование, средно образование, специализирано средно образование (колеж) и професионално ниво. Чрез набора от данни MMLU се проверяват не само фактологични знания, но и способност за разсъждение и прилагане на идеи в различни контексти.

Разработването на набора от данни **MMLU-BG** се основава на оригиналния набор от данни MMLU, като е отчетен опитът от разработването на по-новите версии на набора от данни, както и опитът при разработването на многоезични данни. Във връзка с това в изложената по-долу процедура са въведени процедурни стъпки за подбор на тематичните области и подобряване на качеството на данните.

Организация на данните в MMLU-BG

MMLU-BG следва структурата и организацията на данните в оригиналния набор от данни MMLU.

Всяка от 57-те тематични области в MMLU съдържа три части: **Dev (Разработка)**, **Val (Валидиране)**, **Test (Тестване)**. Тази структура е обоснована от възможността за прилагане на few-shot демонстрации, настройка за голям брой параметри и надеждна оценка на големите езикови модели.

Dev (Разработка) е най-малката част и съдържа точно 5 въпроса за всяка тематична област. Тази част е предназначена за few-shot learning демонстрации, където моделът получава няколко примерни въпроса и отговора преди да бъде тестван.

Val (Валидация) съдържа средно 27 въпроса за тематична област. Тази част се използва основно за избор и настройка на конфигурационните параметри на модела. Валидационното множество от въпроси позволява на изследователите да експериментират с различни настройки и да изберат оптималната конфигурация за техния модел.

Test (Тестване) е най-голямата част за всяка тематична област, като включва поне 100 въпроса. Тази част се използва за финалната оценка на производителността на модела и генериране на официалния MMLU score.

Необходимо оборудване и среда за разработване на прототип на набора от данни MMLU-BG

Разработването на набора от данни MMLU-BG се извършва в споделено файлово пространство между членовете на екипа с регулиране на нивата на достъп и редакция в данните. Изисква се интернет достъп до споделеното пространство.

За целите на превода е възможно използване на инструменти за автоматичен превод (Google translate или AI системи), както и инструменти за компютърноподпомогнат превод.

Началната обработка на данните и крайната обработка и експортиране на завършените данни се осъществява посредством поредица от скриптове за валидиране на формата и съдържанието на данните, разделяне на частите в данните (Dev, Val, Test) и паралелизиране на българските и английските съответствия във въпросите.

Процедура за разработване на прототип на набора от данни MMLU-BG

Протоколът за разработване на прототип на набора от данни MMLU-BG включва следните процедури и правила за работа:

1. Като изходни данни е използван наборът от данни MMLU в неговата оригинална версия и обхват от 57 тематични области. Те са показани в Приложение 1, като са означени и тематичните области, обхванати в текущата версия на набора от данни.

Всяка тематична област се документира, като се посочват: експертът, който я превежда; експертът, който я проверява; състояние (разпределена, в процес, завършена); прогнозна дата на завършване. Така се проследява степента на завършеност на всяка тематична област и на целия набор от данни, както и разпределението на натовареност между експертите в екипа.

2. Изработен е подробен план за етапите на разработване на набора от данни чрез превод на данните от MMLU, редакция и оформление. Разпределението между членовете на екипа цели да постигне равномерно натоварване и осигуряване на експертиза при превода, като членове на екипа превеждат и проверяват тематични области, с чиято терминология са запознати (имат опит в разработването на лексикални ресурси в съответните области).

3. Процедурите за редакция и оформление на въпросите и отговорите имат за цел осигуряване на:

- последователност,
- правилно езиково оформление с оглед на правописа, пунктуацията и граматичната оформеност,
- езикова точност на формулировките,
- терминологична уместност и прецизност,
- фактологична точност на въпросите и дадените отговори, както и на посочения като верен отговор.

4. Тематичните области са организирани във файлове във формат CSV, като всяка тематична област е в отделен файл; за целите на разработката във файла са обединени въпросите от трите подразделения на оригиналния набор от данни – Dev, Val и Test. За съотнасяне с оригиналните въпроси и разпределението на частите са въведени идентификационни индекси.

В процеса на работа данните се представят в таблична форма. В таблицата се добавят нови колони за преводните еквиваленти на въпроса и дадените отговори. Включват се и колони за бележки по превода и редакцията, както и за означаване на проблемни въпроси с оглед на фактологичната точност и уместност.

5. Всяка тематична област се превежда от един експерт, за да се осигури последователно използване на терминология и формулировки, както и по-задълбочено запознаване с конкретната тематична област.

6. След превода и редакцията всяка тематична област се проверява от друг експерт, като се обсъждат грешки и проблемни въпроси и формулировки. Всички проблемни случаи се описват в общ документ и се обсъждат по време на редовните седмични срещи на екипа. Така се осигурява консистентност при оформяне на различните тематични области.

7. При завършване на дадена тематична област форматът на данните се проверява за валидност (разположение и съдържание на колоните) и тематичната област се документираща като завършена.

8. При публикуване на версия на набора от данни се обработват всички готови тематични области, като се разпределят обратно в трите подразделения – Dev, Val и Test. Експортира се наборът от данни за български език, заедно с неговото паралелно съответствие на английски език – извадката от MMLU, която съответства и се съотнася на завършената част от набора от данни за български.

9. В последващи версии на набора от данни е възможно да бъдат направени редакции и на вече публикувани версии, които отразяват поправки на идентифицирани неточности или цялостно прокарване на дадени принципи или формулировки.

Източници на данните:

MMLU: <https://huggingface.co/datasets/cais/mmlu/tree/main>

MMLU-Pro: <https://huggingface.co/datasets/li-lab/MMLU-ProX>

Global MMLU: <https://huggingface.co/datasets/CohereLabs/Global-MMLU>

Приложение 1.

Таблица 1. Разпределение на данните в MMLU-BG и покритие спрямо MMLU (текуща версия 1.0)

Файл	Тематична област	Ключови думи	Суперкатегория	Брой въпроси	Включен и във версия 1.0
abstract_algebra.csv	Абстрактна алгебра	Групи, пръстени, полета, векторни пространства	Наука, технология, инженерство, математика (STEM)	116	x
anatomy.csv	Анатомия	Централна нервна система, кръвоносна система	Наука, технология, инженерство, математика (STEM)	154	x
astronomy.csv	Астрономия	Слънчева система, галактики, астероиди	Наука, технология, инженерство, математика (STEM)	173	x
business_etiquettes.csv	Бизнес етика	Корпоративна отговорност, заинтересовани страни, регулации	Други	116	x

clinical_knowledge.csv	Клинични знания	Точкова диагностика, стави, коремен преглед	Други	299	х
college_biology.csv	Биология (университетско ниво)	Клетъчна структура, молекулярна биология, екология	Наука, технология, инженерство, математика (STEM)	171	
college_chemistry.csv	Химия (университетско ниво)	Аналитична химия, органична и неорганична химия, физикохимия	Наука, технология, инженерство, математика (STEM)	129	х
college_computer_science.csv	Компютърни науки (университетско ниво)	Алгоритми, системи, графи, рекурсия	Наука, технология, инженерство, математика (STEM)	253	х
college_mathematics.csv	Математика (университетско ниво)	Диференциални уравнения, реален анализ, комбинаторика	Наука, технология, инженерство, математика (STEM)	174	х
college_medicine.csv	Медицина (университетско ниво)	Въведение в биохимията, социология, логическо мислене	Други	251	х
college_physics.csv	Физика (университетско ниво)	Електромагнетизъм, термодинамика, специална теория на относителността	Наука, технология, инженерство, математика (STEM)	124	х
computer_security.csv	Компютърна сигурност	Криптография, зловреден софтуер, атаки по странични канали, техники за тестване с произволни данни	Наука, технология, инженерство, математика (STEM)	132	х
conceptual_physics.csv	Теоретична физика	Закони на Нютон, въртливо движение, гравитация, звук	Наука, технология, инженерство, математика (STEM)	266	
econometrics.csv	Иконометрия	Променливост, дългосрочни зависимости,	Социални науки	545	

		прогнозиране			
electrical_engineering.csv	Електротехника	Електрически вериги, енергийни системи, електрическо задвижване	Наука, технология, инженерство, математика (STEM)	166	
elementary_mathematics.csv	Елементарна математика	Текстови задачи, умножение, деление с остатък, закръгляване	Наука, технология, инженерство, математика (STEM)	434	x
formal_logic.csv	Формална логика	Твърдения, предикатна логика, логика от първи ред	Хуманитарни науки	333	x
global_facts.csv	Световни факти	Крайна бедност, ниво на грамотност, продължителност на живота	Други	115	
high_school_biology.csv	Биология (гимназиално ниво)	Естествен подбор, наследственост, клетъчен цикъл, цикъл на Кребс	Наука, технология, инженерство, математика (STEM)	347	x
high_school_chemistry.csv	Химия (гимназиално ниво)	Химични реакции, йони, киселини и основи	Наука, технология, инженерство, математика (STEM)	230	x
high_school_computer_science.csv	Информатика (гимназиално ниво)	Масиви, условия, итерации, наследяване	Наука, технология, инженерство, математика (STEM)	355	
high_school_european_history.csv	История на Европа (гимназиално ниво)	Ренесанс, реформация, индустриализация	Хуманитарни науки	1275	
high_school_geography.csv	География (гимназиално ниво)	Миграция на населението, ползване на селскостопански земи, урбанизация	Социални науки	225	

high_school_government_and_politics.csv	Държавно управление и политика (гимназиално ниво)	Клонове на държавното управление, гражданска свобода, политическа идеология	Социални науки	219	
high_school_macroeconomics.csv	Макроикономика (гимназиално ниво)	Икономически показатели, брутен вътрешен продукт, международна търговия	Социални науки	438	x
high_school_mathematics.csv	Математика (гимназиално ниво)	Основи на алгебрата, алгебра, тригонометрия, математически анализ	Наука, технология, инженерство, математика (STEM)	304	x
high_school_microeconomics.csv	Микроикономика (гимназиално ниво)	Търсене и предлагане, недостаъчна конкуренция, пазарна неефективност	Социални науки	269	
high_school_physics.csv	Физика (гимназиално ниво)	Кинематика, енергия, въртящ момент, налягане на флуидите	Наука, технология, инженерство, математика (STEM)	176	x
high_school_psychology.csv	Психология (гимназиално ниво)	Поведение, личност, емоции, учене	Социални науки	610	x
high_school_statistics.csv	Статистика (гимназиално ниво)	Случайни величини, разпределение на извадки, хи-квадрат тестове	Наука, технология, инженерство, математика (STEM)	248	
high_school_us_history.csv	История на САЩ (гимназиално ниво)	Гражданската война в САЩ, Голямата депресия, Великото общество	Хуманитарни науки	1118	
high_school_world_history.csv	Световна история (гимназиално ниво)	Османска империя, икономически империализъм,	Хуманитарни науки	2123	x

		Първа световна война			
human_aging.csv	Стареене при човека	Стареене, деменция, дълголетие, личностни промени	Други	251	x
human_sexuality.csv	Човешка сексуалност	Бременност, полова диференциация, сексуална ориентация	Социални науки	148	x
international_law.csv	Международно право	Човешки права, суверенитет, морско право, употреба на сила	Хуманитарни науки	139	
jurisprudence.csv	Юриспруденция	Естествено право, правен позитивизъм, правен реализъм	Хуманитарни науки	126	
logical_fallacies.csv	Логически противоречия	Похвати за манипулация на твърдения, надценяване на вероятността, погрешно пренасяне от частно към общо	Хуманитарни науки	190	x
machine_learning.csv	Машинно обучение	Машини с поддържащи вектори, размерност на Вапник-Червоненки с, архитектури за дълбоко учене	Наука, технология, инженерство, математика (STEM)	128	
management.csv	Стопанско управление	Организация, комуникация, организационна структура	Други	120	x
marketing.csv	Маркетинг	Сегментация, ценообразуване, пазарни проучвания	Други	264	x

medical_genetics.csv	Медицинска генетика	Гени и рак, често срещани хромозомни заболявания	Други	116	
miscellaneous.csv	Разни	Земеделие, оценка по метода на Ферми, поп култура	Други	874	х
moral_disputes.csv	Морални спорове	Свобода на словото, зависимости, смъртно наказание	Хуманитарни науки	389	х
moral_scenarios.csv	Морални казуси	Разкриване на физическо насилие, кражба, външни фактори	Хуманитарни науки	1000	х
nutrition.csv	Хранене	Метаболизъм, водоразтворими витамини, диабет	Други	691	х
philosophy.csv	Философия	Скептицизъм, фронеизис, експериментът „Давещото се дете“	Хуманитарни науки	350	х
prehistory.csv	Праистория	Неандерталци, Мезоамерика, измиране, каменни инструменти	Хуманитарни науки	364	х
professional_accounting.csv	Счетоводство (професионално ниво)	Одит, финансови отчети, регулация, оценка	Други	332	х
professional_law.csv	Право (професионално ниво)	Гражданско право, наказателно право, договори, собственост, доказателства	Хуманитарни науки	1709	
professional_medicine.csv	Медицина (професионално ниво)	Диагностика, фармакотерапия, превенция на заболяванията	Други	308	х
professional_psychology.csv	Психология (професионално ниво)	Диагностика, биология и поведение, човешко	Социални науки	686	

		развитие			
public_relations.csv	Връзки с обществеността (пиар)	Теория на медиите, кризисен пиар, разузнаване	Социални науки	127	x
security_studies.csv	Сигурност	Екологична сигурност, тероризъм, оръжие за масово унищожение	Социални науки	277	x
sociology.csv	Социология	Социализация, градове и общности, неравенство и богатство	Социални науки	229	x
us_foreign_policy.csv	Външна политика на САЩ	Мека сила, външна политика по време на Студената война, изолационизъм	Социални науки	120	x
virology.csv	Вирусология	Епидемиология, коронавируси, ретровируси, херпес вируси	Други	189	x
world_religions.csv	Световни религии	Юдаизъм, християнство, ислям, будизъм, джайнизъм	Хуманитарни науки	195	x

История на редакциите на протокола

Версия	Дата	Автор	Промени
1.0	15.09.2025	С. Коева и екип	Първоначален протокол
