



Финансирано от
Европейския съюз

Следващо Поколение ЕС



План за възстановяване и
устойчивост



Република България

ВЪТРЕШЕН ДОКЛАД

от гл. ас. д-р Ивелина Стоянова, гл. ас. д-р Светлозара Лесева, гл. ас. д-р Цветана Димитрова, гл. ас. д-р Валентина Стефанова, гл. ас. д-р Мария Тодорова, гл. ас. д-р Христина Кукова, гл. ас. д-р Михаела Московска
до проф. д-р Тинко Тинчев, ръководител на работния пакет,

за текуща самооценка на техническата готовност на **продукта**

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

Вътрешен доклад ID: IFGPT-2025-REPORT-004

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

Тип на продукта: **Набор от езикови данни**

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

Дата: **10.09.2025 г.**

Извършени дейности по разработване на набора от данни MMLU-BG

1. Подбрани са тематични области от набора от данни MMLU и са определени правила и принципи за техния превод и представяне. Преводът е извършен от експерт и са предприети стъпки за осигуряване на качеството чрез допълнителна проверка да данните от втори експерт. До момента са преведени 39 тематични области (от общо 57).

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

Файл	Тематична област	Ключови думи	Суперкатегория	Брой въпрос	Включен и във
------	------------------	--------------	----------------	-------------	---------------

				и	версия 1.0
abstract_algebra.csv	Абстрактна алгебра	Групи, пръстени, полета, векторни пространства	Наука, технология, инженерство, математика (STEM)	116	х
anatomy.csv	Анатомия	Централна нервна система, кръвоносна система	Наука, технология, инженерство, математика (STEM)	154	х
astronomy.csv	Астрономия	Слънчева система, галактики, астероиди	Наука, технология, инженерство, математика (STEM)	173	х
business_ethics.csv	Бизнес етика	Корпоративна отговорност, заинтересовани страни, регулации	Други	116	х
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, прогнозиране	Социални науки	545	
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	х
formal_logic.csv	Формална логика	Твърдения, предикатна логика, логика от първи ред	Хуманитарни науки	333	х
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	х

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_macroconomics.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, недостатъчна 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	х
high_school_psychology.csv	Психология (гимназиално ниво)	Поведение, личност, емоции, учене	Социални науки	610	х
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, Първа световна война	Хуманитарни науки	2123	х
human_aging.csv	Стареене при човека	Стареене, деменция, дълголетие, личностни промени	Други	251	х
human_sexuality.csv	Човешка сексуалност	Бременност, полова диференциация, сексуална ориентация	Социални науки	148	х
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,	Хуманитарни науки	190	х

		надценяване на вероятността, погрешно пренасяне от частно към общо			
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	x
moral_disputes.csv	Морални спорове	Свобода на словото, зависимости, смъртно наказание	Хуманитарни науки	389	x
moral_scenarios.csv	Морални казуси	Разкриване на физическо насилие, кражба, външни фактори	Хуманитарни науки	1000	x
nutrition.csv	Хранене	Метаболизъм, водоразтворими витамини, диабет	Други	691	x
philosophy.csv	Философия	Скептицизъм, фронеизис, експериментът „Давещото се дете“	Хуманитарни науки	350	x

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 развитие	Социални науки	686	
public_relations.csv	Връзки с обществеността (пиар)	Теория на медиите, кризисен пиар, разузнаване	Социални науки	127	х
security_studies.csv	Сигурност	Екологична сигурност, тероризъм, оръжие за масово унищожение	Социални науки	277	х
sociology.csv	Социология	Социализация, градове и общности, неравенство и богатство	Социални науки	229	х
us_foreign_policy.csv	Външна политика на САЩ	Мека сила, външна политика по време на Студената война, изолационизъм	Социални науки	120	х
virology.csv	Вирусология	Епидемиология, коронавируси, ретровируси, херпес	Други	189	х

		вируси			
world_religions.csv	Световни религии	Юдаизъм, християнство, ислям, будизъм, джайнизъм	Хуманитарни науки	195	x

2. Данните са публикувани в хранилища за данни и могат да се използват за оценка на големи езикови модели:

<https://huggingface.co/datasets/DCL-IBL/MMLU-BG-Dataset>

<https://github.com/DCL-IBL/MMLU-BG-Dataset>

3. Проведени са функционални тестове в контролирана среда, потвърждаващи работоспособността на интегрираната система, като данните са приложени в експерименти за оценка на големи езикови модели. Експериментите показват, че данните са приложими, като позволяват сравнителни анализи както за производителността на различни модели, така и за двата езика – български и английски.

4. Идентифицирани са технологични ограничения и критични точки:

- Въпроси с фактологични неточности.
- Проблеми с липса на терминологични съответствия на български език.

Тези проблеми се преодоляват чрез стратегии, обсъдени на срещите с екипа и имат за цел подобряването на качеството на ресурса. Новите версии на ресурса ще съдържат и подобрения по вече публикуваните тематични области с цел да се осигури последователност и точност на данните.

Получени резултати

1. Използван е цялостният концептуален модел на продукта MMLU и е създаден протокол за неговия превод и представяне на български език.

2. Приложени са ключовите технически параметри на продукта, стандартни за извършване на оценка на големи езикови модели – формат на данните. Данните са представени паралелно на български и английски език, с което се позволяват директни съпоставителни експерименти за двата езика.

3. Набор от данни MMLU-BG на български език за оценка на продуктивността на големи езикови модели. Текущата версия включва 39 тематични области, получени чрез превод от оригиналния набор от данни MMLU. Данните са достъпни тук: <https://huggingface.co/datasets/DCL-IBL/MMLU-BG-Dataset>, <https://github.com/DCL-IBL/MMLU-BG-Dataset>

Доказателства и аргументиране на ниво на техническа готовност TRL 4

Конкретните доказателства за постигнато ниво **TRL 4** включват:

- Настоящият вътрешен доклад документиращ извършените дейности.

- Данните са достъпни в хранилища за данни:
<https://huggingface.co/datasets/DCL-IBL/MMLU-BG-Dataset>
<https://github.com/DCL-IBL/MMLU-BG-Dataset>
- Данните са приложени в експерименти за оценка на големи езикови модели, които предстои да бъдат описани в публикация.
- Представени са резултати от начални тестове (test report) за приложението на MMLU-BG за целите на оценката на големи езикови модели, който е в процес на разработване (Приложение 1).

Приложение 1.

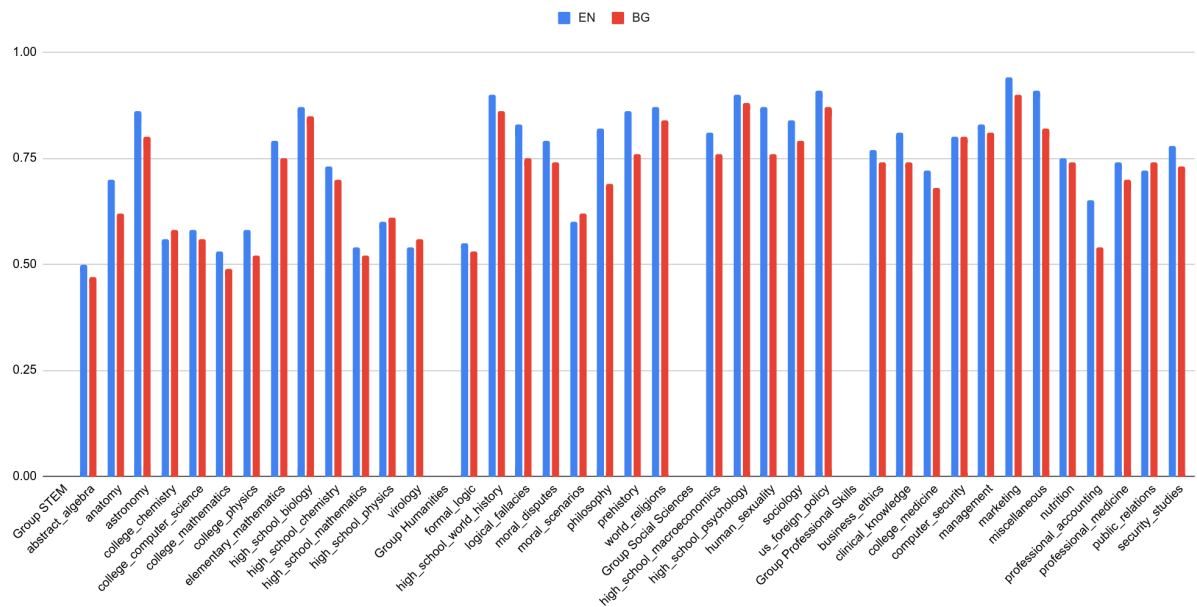
Таблица 1. Примерен резултат от сравнителна оценка за модела **Гемма 3 (27 млрд. параметъра)** върху MMLU и MMLU-BG (жълто – под 0.8; оранжево – под 0.65; червено – под 0.5).

Език и модел	EN (MMLU)	google/gemma-3-27b-it		BG (MMLU-BG)	google/gemma-3-27b-it	
Group STEM						
abstract_algebra	50	100	0.5	47	100	0.47
anatomy	95	135	0.7	84	135	0.62
astronomy	130	152	0.86	121	152	0.8
college_chemistry	56	100	0.56	58	100	0.58
college_computer_science	58	100	0.58	56	100	0.56
college_mathematics	53	100	0.53	49	100	0.49
college_physics	59	102	0.58	53	102	0.52
elementary_mathematics	297	378	0.79	282	378	0.75
high_school_biology	269	310	0.87	263	310	0.85
high_school_chemistry	148	203	0.73	142	203	0.7
high_school_mathematics	145	270	0.54	140	270	0.52
high_school_physics	91	151	0.6	92	151	0.61
virology	89	166	0.54	93	166	0.56

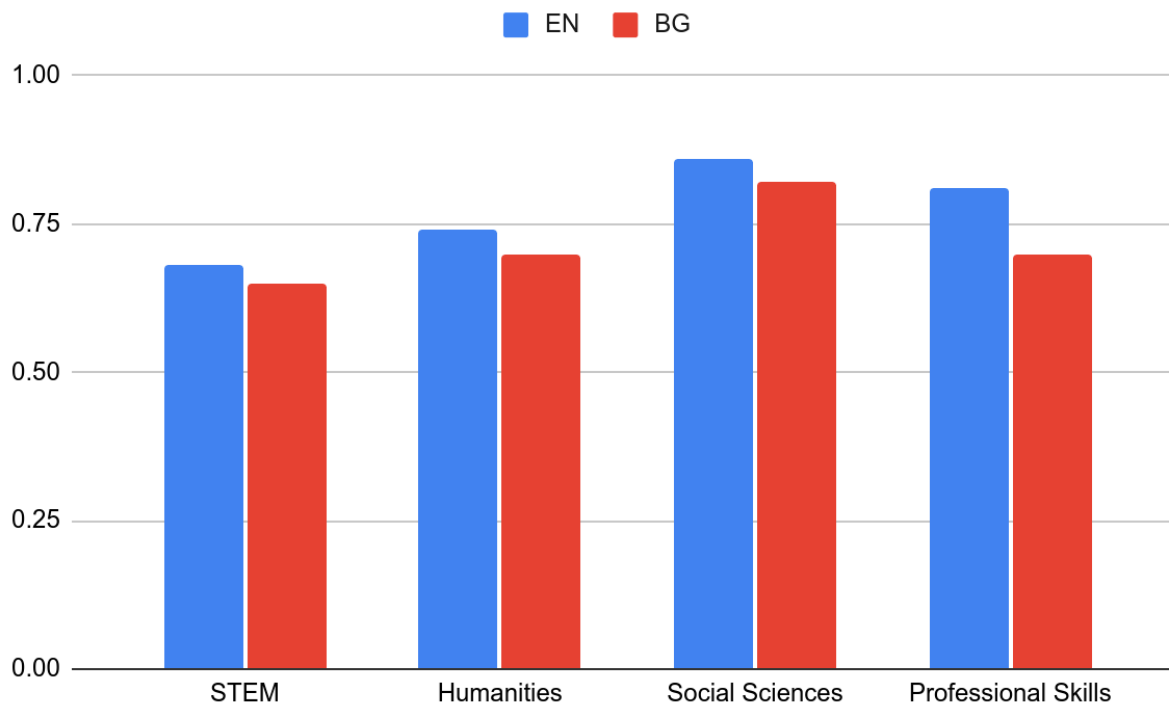
Group Humanities						
formal_logic	69	126	0.55	67	126	0.53
high_school_world_history	213	237	0.9	204	237	0.86
logical_fallacies	136	163	0.83	123	163	0.75
moral_disputes	274	346	0.79	257	346	0.74
moral_scenarios	540	895	0.6	552	895	0.62
philosophy	254	311	0.82	215	311	0.69
prehistory	278	324	0.86	247	324	0.76
world_religions	148	171	0.87	144	171	0.84
Group Social Sciences						
high_school_macro_economics	315	390	0.81	296	390	0.76
high_school_psychology	490	545	0.9	482	545	0.88
human_sexuality	114	131	0.87	99	131	0.76
sociology	168	201	0.84	158	201	0.79
us_foreign_policy	91	100	0.91	87	100	0.87
Group Professional Skills						
business_ethics	77	100	0.77	74	100	0.74
clinical_knowledge	214	265	0.81	197	265	0.74
college_medicine	124	173	0.72	118	173	0.68
computer_security	80	100	0.8	80	100	0.8
management	86	103	0.83	83	103	0.81
marketing	220	234	0.94	211	234	0.9
miscellaneous	715	783	0.91	642	783	0.82
nutrition	231	306	0.75	225	306	0.74
professional_accounting	184	282	0.65	153	282	0.54
professional_m	200	272	0.74	191	272	0.7

edicine						
public_relations	79	110	0.72	81	109	0.74
security_studies	190	245	0.78	178	245	0.73
ОБОБЩЕНИ РЕЗУЛТАТИ						
STEM	1540	2267	0.68	1480	2267	0.65
Humanities	1912	2573	0.74	1809	2573	0.7
Social Sciences	1178	1367	0.86	1122	1367	0.82
Professional Skills	2577	3196	0.81	2233	3195	0.7

Фигура 1. Сравнителен анализ между български и английски по тематични области.



Фигура 2. Обобщен сравнителен анализ между български и английски по суперобласти.



Предал: Ивелина Стоянова
Дата: 10.09.2025

Приел: Тинко Тинчев