Набори от данни на български език за оценка на големи езикови модели

Представяне на резултати от проекта
„Инфраструктура за фина настройка на предварително обучени големи езикови модели“

Валентина Стефанова
Секция по компютърна лингвистика
Институт за български език, Българска академия на науките

valentina@dcl.bas.bg

Данни за оценка на големи езикови модели

Разработени са четири набора от данни за оценка на големи езикови модели. Тук ще представим два от тях за две различни задачи.

MMLU-BG – набор от данни на български език за оценка на възможностите на големите езикови модели да „разбират“ и да прилагат знание от различни области. Създаден е от експерти посредством превод и адаптация на Measuring Massive Multitask Language Understanding (MMLU).

Reasoning-BG – набор от данни на български език за оценка на възможностите на големите езикови модели да „разсъждават“ и да достигат до заключения.

MMLU-BG

Данните в MMLU-BG са организирани в 56 тематични области. Включва общо 15,000 въпроса, всеки с по четири отговора, от които един е правилен.

Създаването на данните за български език включи:

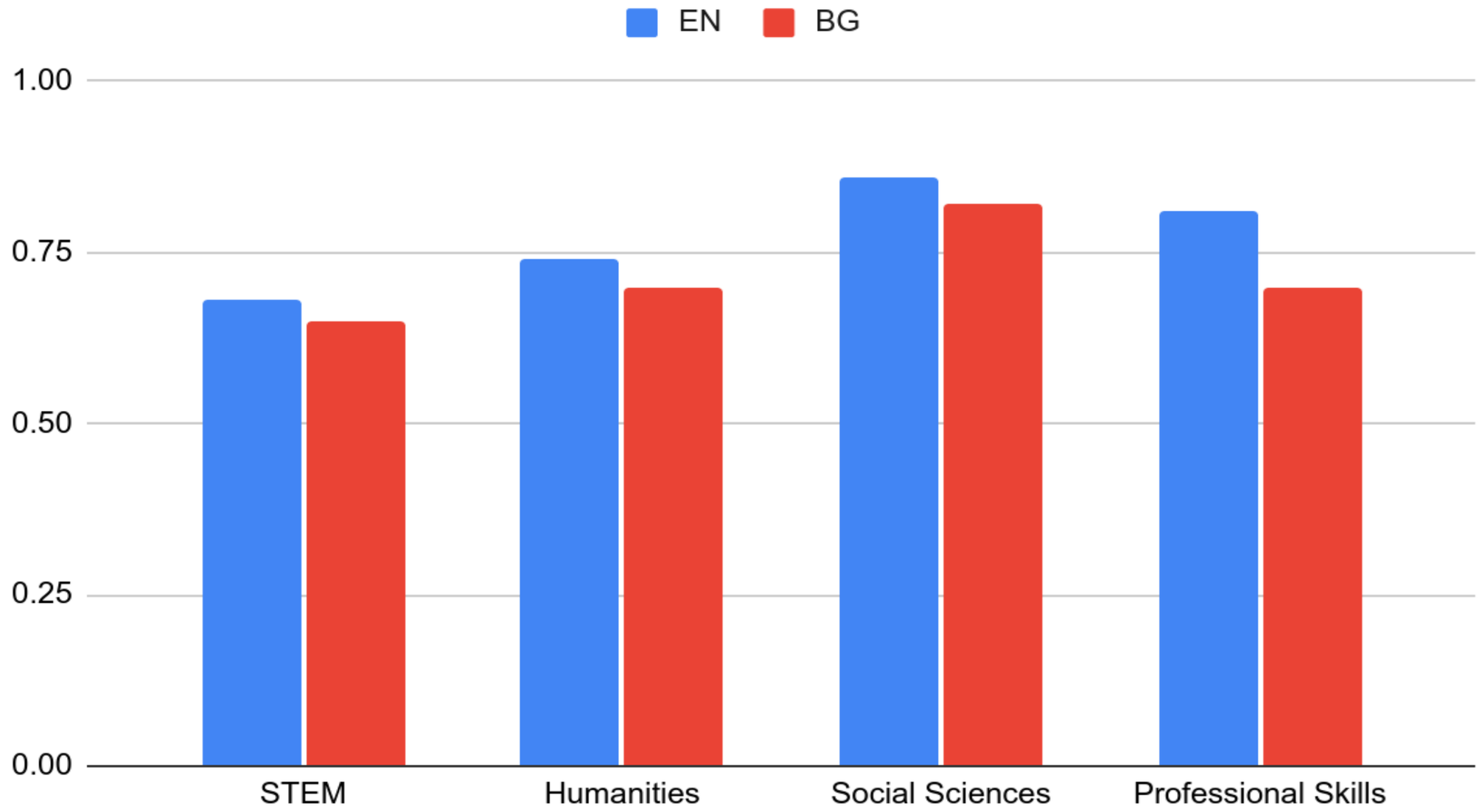
- комплексна терминологична и смислова адаптация (не просто превод);
- запазване на степента на трудност и логическата структура на въпросите;
- коректност на научната терминология;
- смислова и граматическа правилност на български език.

MMLU-BG

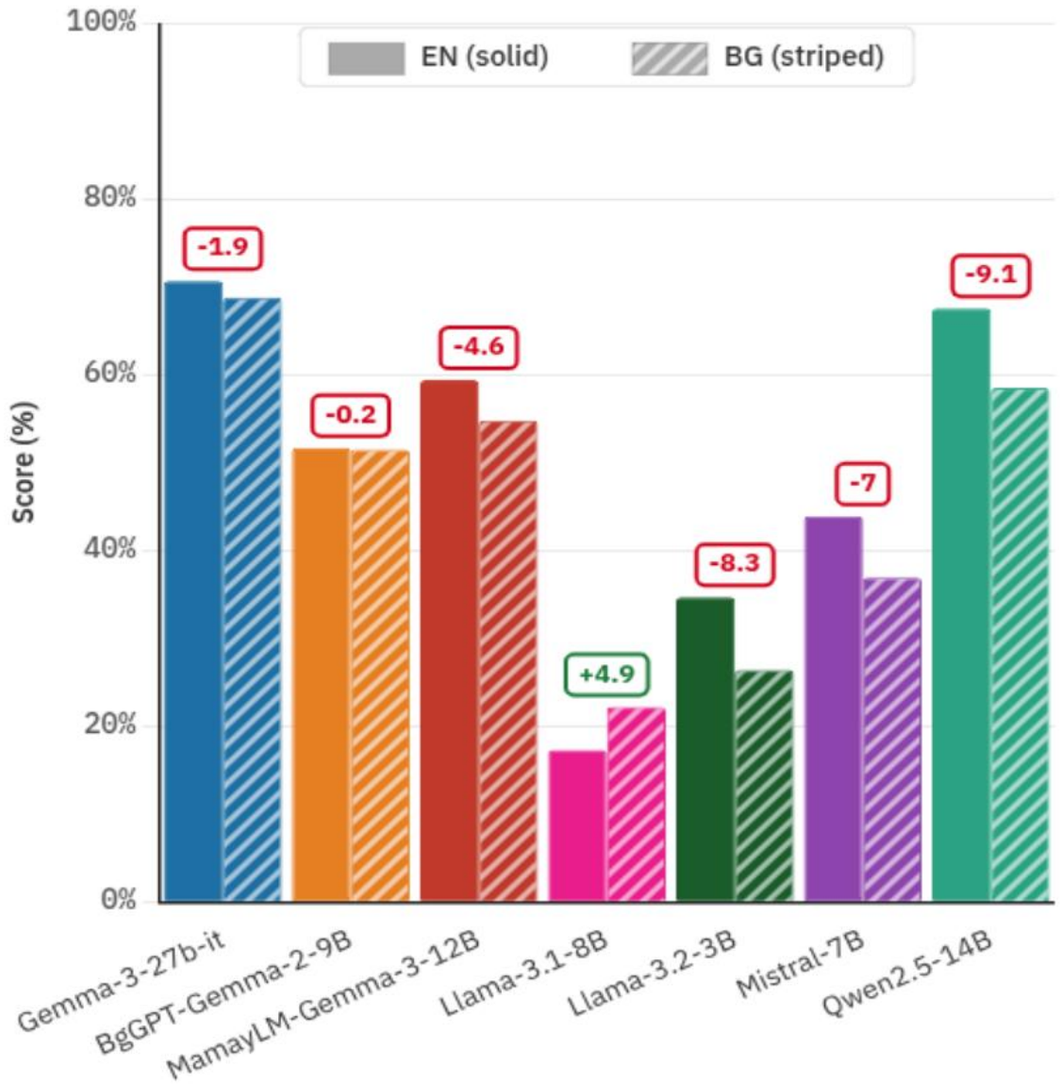
Какви последици биха могли да възникнат вследствие на разработването и използването на генетично модифицирани растения?

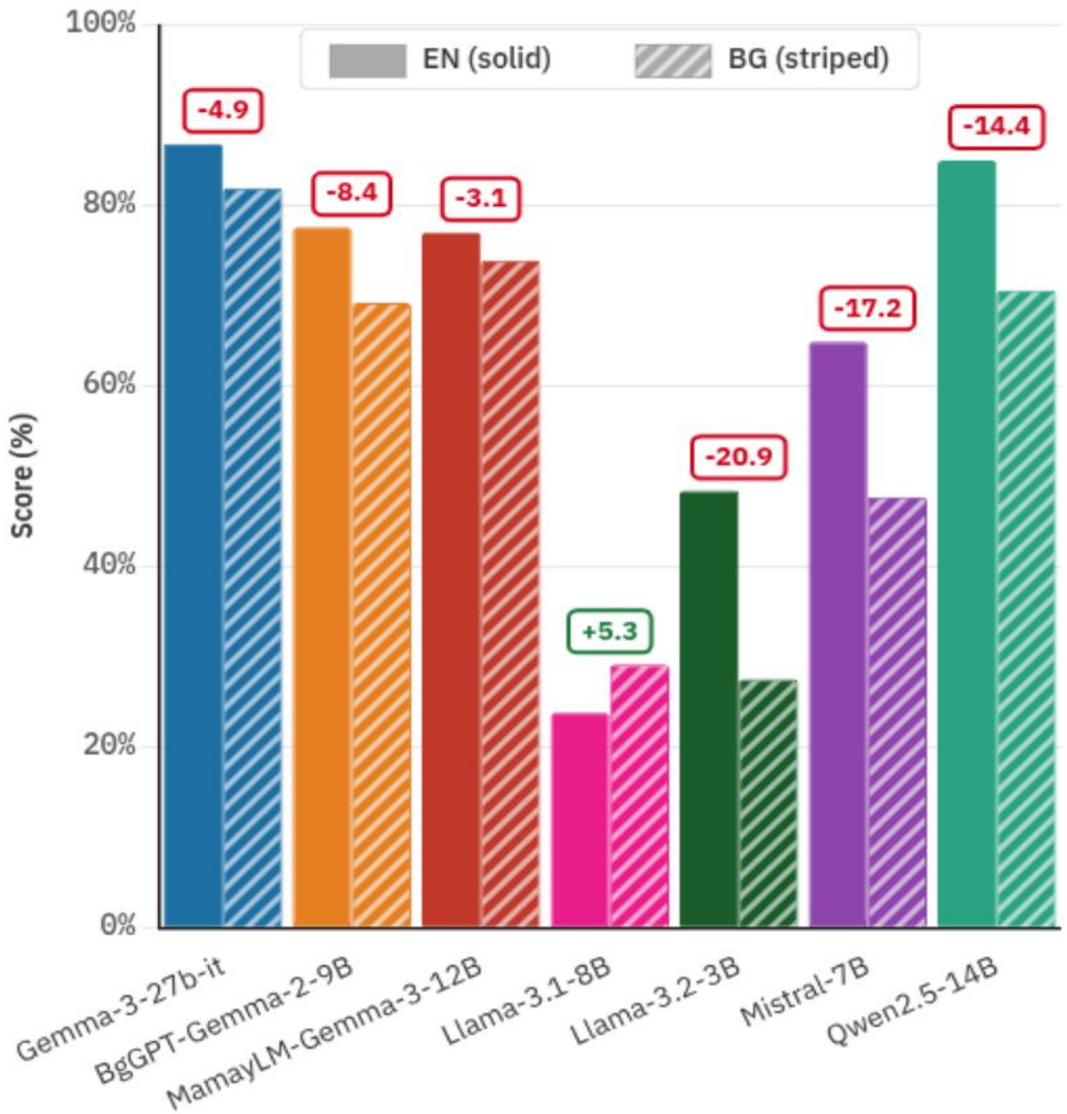
A) поевтиняване на семената
B) патентоване на нови сортове растения
C) повишаване на генетичното разнообразие в засетите площи
D) прилагане на по-слаб контрол и по-малко законови регулации, отколкото по отношение на немодифицираните култури

MMLU-BG


EN
BG
1.00
0.75
0.50
0.25
0.00
STEM
Humanities
Social Sciences
Professional Skills

MMLU-BG


EN (solid)
BG (striped)
Score (%)
100%
80%
60%
40%
20%
0%
-1.9
-0.2
-4.6
+4.9
-8.3
-7
-9.1
Gemma-3-27b-it
BgGPT-Gemma-2-9B
MamayLM-Gemma-3-12B
Llama-3.1-8B
Llama-3.2-3B
Mistral-7B
Qwen2.5-14B


EN (solid)
BG (striped)
Score (%)
100%
80%
60%
40%
20%
0%
-4.9
-8.4
-3.1
+5.3
-20.9
-17.2
-14.4
Gemma-3-27b-it
BgGPT-Gemma-2-9B
MamayLM-Gemma-3-12B
Llama-3.1-8B
Llama-3.2-3B
Mistral-7B
Qwen2.5-14B

Набор от данни за оценка на разбиране и аргументиране

Наборите от данни за оценка на извличането на информация и логически изводи върху текст съдържат подходящи текстове и въпроси към тях, предназначени:

- да оценяват способността на езиковите модели да извършват смислов анализ;
- за извличане на информация;
- за намиране на логически връзки и интерпретиране на текстово съдържание.

Набор от данни за оценка на разбиране и аргументиране

Включва 232 текста с по 10 въпроса към всеки текст.

Процесът по създаване на набора от данни включи:

- подбор на подходящи научнопопулярни текстове;
- предварителна езикова обработка;
- генериране на въпроси и отговори чрез свободен голям езиков модел, работещ локално;
- ръчна редакция и проверка на въпросите и отговорите;
- проверка на еднозначността и коректността на отговорите;
- оценка на смисловото съответствие между текст и въпрос.

Набор от данни за оценка на разбиране и аргументиране

Типология и класификация на текстовете:

- Информативни текстове, извлечени от научни статии и биографии, ориентирани към проверка на фактологична точност и времева (хронологична) ориентация.
- Научнопопулярни текстове, заимствани от списания и енциклопедии, ориентирани към откриване на причинно-следствени връзки.

Набор от данни за оценка на разбиране и аргументиране

Разпределение на корпуса по тематични области

Тематична област	Текстове	Изречения	Думи
Природни науки	60	2,535	40,462
Наука и технологии	42	1,511	26,418
Литературни изследвания	38	1,292	26,658
История	37	1,279	25,312
География	23	1,108	19,816
Психология	13	474	8,459
Култура	7	313	5,780
Медицина	7	253	4,347
Философия	5	206	4,103
Общо	**232**	**8,971**	**157,355**

Общо 232 текста, 8,971 изречения и 157,355 думи.

Набор от данни за оценка на разбиране и аргументиране

Парадоксално е, че в стремежа ни към по-здравословни социални навици и отношения днес живеем във видимо за всеки по-болно и нещастно и все по-дехуманизирано общество. Основната причина, разбира се, е откъсването от реалността на човешката ни същност и неадекватното идеологизиране на взаимоотношенията ни. Ницше има едно важно предупреждение: „Никакви скокове в добродетелта!“. Защото накъдето и да скочиш, все пропадаш. Екстремният стремеж към добродетелност е един от основните деструктивни фактори в живота ни днес... Борбата винаги е била за съзнанието ни и „моралът“ винаги е бил поводът, който е оправдавал всичко.

Какво според автора е отношението между „морал“ и „борба“?

A) Моралът винаги е средство за оправдаване на борбата.

B) Борбата винаги е за постигане на морални цели.

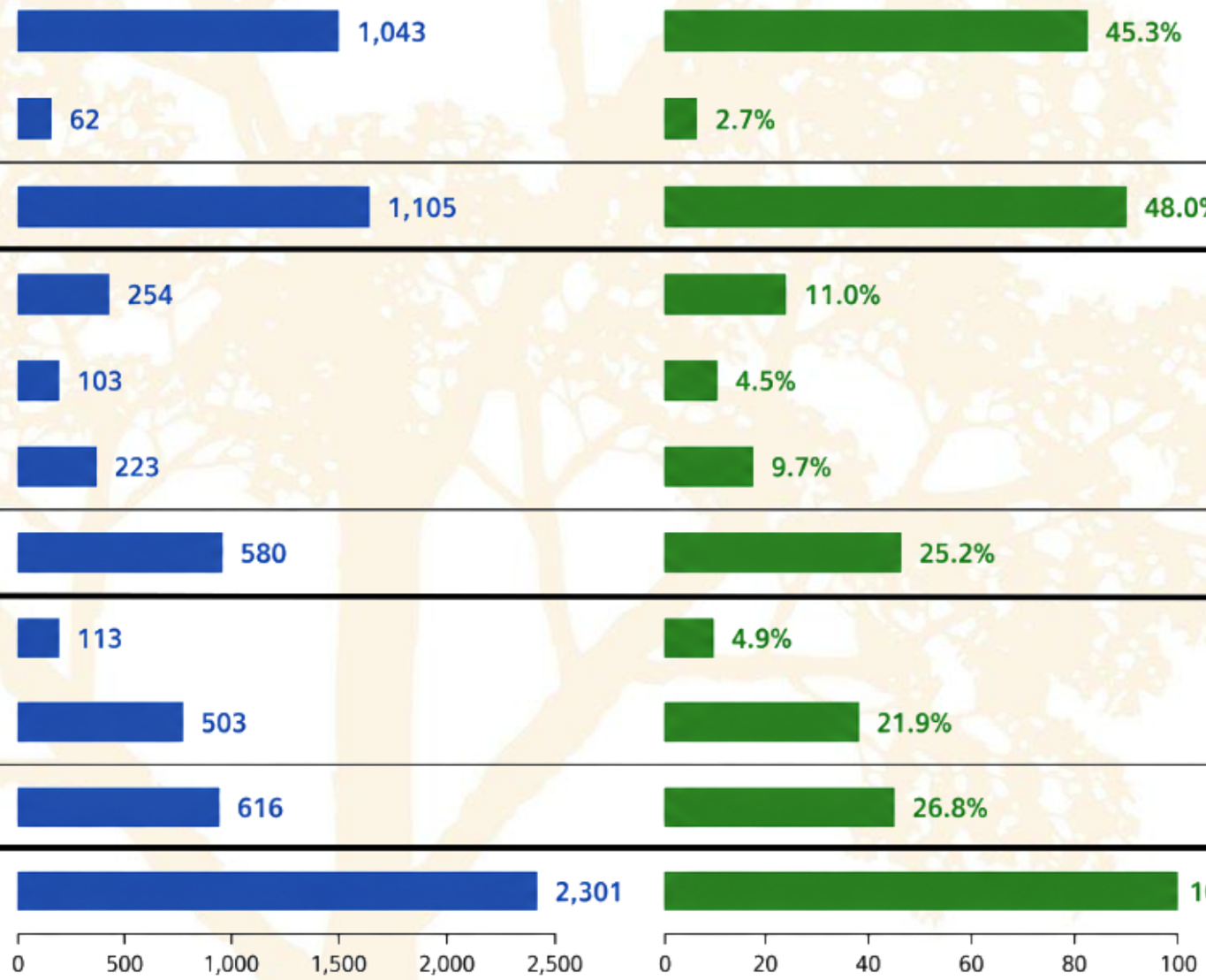
C) Моралът и борбата са взаимно изключващи се понятия.

D) Моралът и борбата са нямат връзка помежду си.

Набор от данни за оценка на разбиране и аргументиране

Разпределение на въпросите според категориите

Категория въпроси	Брой (N)	Процент (%)
Фактологични	1,043	45.3%
Лексикални и интерпретативни	62	2.7%
Извличащи	1,105	48.0%
Причинно-следствени	254	11.0%
Значимост	103	4.5%
Сравнения и отношения	223	9.7%
Абстрактни	580	25.2%
Условни	113	4.9%
Достигане до заключение	503	21.9%
Логически	616	26.8%
Общо	2,301	100.0%



Благодарности

⊙ Работата е част от проекта „Инфраструктура за фина настройка на предварително обучени големи езикови модели“ по договор номер ПВУ – 55 от 12.12.2024 /BG-RRP-2.017-0030-C01/.

Финансирано от Европейския съюз
NextGenerationEU

Национален план за възстановяване и устойчивост
НА РЕПУБЛИКА БЪЛГАРИЯ

⊙https://ifgpt.dcl.bas.bg