https://doi.org/10.37708/ezs.swu.bg.v24i3.11

ГОЛЕМИТЕ ЕЗИКОВИ МОДЕЛИ И БЪЛГАРСКАТА ЛИТЕРАТУРА (ЕКСПЕРИМЕНТАЛНО ИЗСЛЕДВАНЕ ВЪРХУ ЗАДАЧИ ОТ НВО ЗА 7. КЛАС)

Владислав Маринов

Великотърновски университет „Св. св. Кирил и Методий“, България
Email: w.marinov@ts.uni-vt.bg

ORCID ID: 0000-0003-4477-5783

LARGE LANGUAGE MODELS AND BULGARIAN LITERATURE (AN EXPERIMENTAL STUDY ON 7TH-GRADE NATIONAL EXTERNAL ASSESSMENT TASKS)

Vladislav Marinov, St Cyril and St Methodius University of Veliko Tarnovo, Bulgaria

ABSTRACT: This study compares the performance of BgGPT, Gemini, Claude, and ChatGPT on standardized tasks concerning Bulgarian literature. The corpus consists of literature tasks 18–24 from the Bulgarian national external assessment for 7th grade over five consecutive years (2022–2026). Each test was solved three times by each system in a new session, producing 60 sessions in total. Multiple-choice tasks were scored against the official answer keys, while the four short answers in task 24 were scored separately and normalized so that each test task had equal weight. The results do not support the working hypothesis that a model adapted to Bulgarian would outperform general multilingual systems. Gemini achieved the highest overall score (88.1%), followed by ChatGPT (86.0%), BgGPT (68.8%), and Claude (61.4%). All four systems performed less accurately on the short free-response task than on the multiple-choice tasks, although the present design does not isolate response format from content. Performance also varied across types of literary work and across cumulative year samples. The comparison is limited to the freely available versions of the four systems accessed in July 2026 and does not isolate the effects of their technical characteristics. The findings indicate that language adaptation alone is insufficient to predict performance on subject-specific and culturally localized content and support the use of authentic local educational tasks when evaluating language models for educational contexts.

KEYWORDS: large language models, BgGPT, Bulgarian literature, standardized assessment, language adaptation

1. Въведение

Големите езикови модели (LLM) все по-широко навлизат в образователната практика, включително в обучението по български език и литература. Български изследвания от последните години разглеждат възможностите за педагогическо интегриране на генеративния изкуствен интелект (GenAI), развитието на умения за работа с езикови модели и необходимостта от критична оценка и проверка на генерираното съдържание (Ангелова/Angelova, 2026; Трайчева/Traycheva, 2026). При работата с български езиков материал трябва да се отчита и ролята на

контекста за реализацията на определени езикови явления (Osenova, 2025). Експериментални данни за работата на LLM с български език показват, че дори предоставянето на точното езиково правило не гарантира устойчивото му прилагане от моделите при промяна на контекста (Todoranova & Todoranova, 2026). Това насочва вниманието не само към възможностите за използване на тези технологии в обучението, но и към необходимостта емпирично да се проверява доколко самите модели се справят с конкретно предметно съдържание. Българската литература е подходяща област за подобна проверка, тъй като работата със задачи върху нея предполага разбиране на български език, знания за автори, произведения, персонажи, жанрови характеристики, теми и мотиви и способност за интерпретация на художествен текст. Значението на езиковия и културния контекст за представянето на LLM се изследва и чрез CulturalBench – специално разработен набор от въпроси за оценяване на знанията на езиковите модели за различни култури. Резултатите показват случаи, в които силни общи модели се представят по-добре от модели, чието обучение се извършва чрез данни, съобразени с езика и с културата на определена държава. (Chiu et al., 2025). Това е така, тъй като представянето на езиковите модели при работа с отделни езици зависи от редица фактори, сред които обемът на наличните обучителни данни за съответния език, близостта му до други езици и капацитетът на модела (Chang et al., 2024). Същевременно специализирани модели, при които отделни компоненти са предназначени за обработка на определени езици или групи езици (cross-lingual expert models), могат да постигат по-добри резултати от модели, обучени да работят с множество езици (Blevins et al., 2024). Тези резултати показват, че специализацията към определен език може да бъде свързана с предимство, но не позволява то да се приема като нейна задължителна последица.

Въпросът за ефекта от езиковата специализация може да бъде разгледан конкретно за българския език чрез BgGPT – модел, разработен с акцент върху българския език и локалното знание (Alexandrov et al., 2024). В настоящото изследване BgGPT не се разглежда като принципно различен тип система спрямо общите многоезични модели, а като подходящ случай за проверка дали целенасочената адаптация към българския език е свързана с по-добро представяне при работа с българско литературно съдържание.

За подобна проверка е важно представянето на моделите да бъде оценявано върху задачи от съответния езиков, културен и образователен контекст. Изследването със SeaExam и SeaBench показва, че представянето на моделите при автентични задачи от местния образователен и културен контекст може да се различава от резултатите им при задачи, първоначално създадени на друг език и впоследствие преведени (Liu et al., 2025). Bulgarian Massive Multitask Language Understanding Benchmark предлага задачи от различни области за оценяване на модели на български език (Koeva et al., 2026), но не включва отделен набор от задачи по литература. Така остава

открит въпросът за представянето на LLM при стандартизирани задачи върху българската литература.

Целта на настоящото изследване е да се установи дали адаптираният към българския език BgGPT показва по-добри резултати от три общи многоезични модела – Gemini, Claude и ChatGPT – при решаването на стандартизирани задачи върху българската литература. Използван е корпус от литературните задачи от националното външно оценяване по български език и литература за 7. клас за пет последователни години (2022 – 2026). Освен общото представяне на моделите се анализират различията в отговорите им според типа на задачите, както и доколко резултатите се променят с разширяването на изследвания корпус. Работната хипотеза е, че BgGPT, който е адаптиран спрямо българския език, ще се представи по-добре от трите общи многоезични модела.

2. Методология

Изследвани са безплатно достъпните към 4 юли 2026 г. версии на BgGPT (INSAIT), Gemini (Google), Claude (Anthropic) и ChatGPT (OpenAI), чрез корпус, който включва свързаните с българската литература задачи № 18 – 24 от националното външно оценяване по български език и литература за 7. клас за пет последователни години (2022 – 2026). Задача № 25 е изключена, тъй като изисква кратък интерпретативен текст, за който официалният ключ предлага примерни насоки, а не еднозначен верен отговор. С изключването на тази задача се ограничава субективната изследователска преценка при оценяването. Официалните тестове и ключове са взети от архива „Тестове и верни отговори от НВО за VII клас по години“ на Министерството на образованието и науката (Министерство на образованието и науката/Ministry of Education and Science, б.д.).

За всяка година са включени шест задачи с избираем отговор (№ 18 – 23) и задача № 24 с четири кратки свободни отговора. Корпусът обхваща 35 задачи – 30 с избираем и 5 със свободно генериран кратък отговор. Четирите отговора в задача № 24 се оценяват поотделно.

Петте теста са предназначени за едно и също образователно равнище и имат съпоставима структура. Задачите се разглеждат като елементи на един и същ тип изпитен инструмент, без да се предполага еднаква емпирична трудност през отделните години. Включването на пет последователни години цели резултатите да не зависят от конкретния подбор на задачи в една изпитна година. Изследването оценява представянето на LLM при стандартизирани задачи върху българската литература, като получените резултати се анализират самостоятелно, без съпоставка с постиженията на учениците през съответните години.

За експеримента от всеки оригинален тест е създаден минитест в отделен файл, който съдържа само задачи № 18 – 24. Премахнати са останалите задачи, ключовете с верните отговори, годината на провеждане и други елементи, които могат пряко да идентифицират конкретния изпит. Процедурата цели да ограничи влиянието на допълнителната информация

при подаването на задачите към моделите, но не може със сигурност да се установи дали задачите или техни версии са присъствали в данните, използвани при обучението на моделите или последващото им дообучение. Възможното предварително познаване на публично достъпни тестови материали е известен методологически проблем при изследването на LLM (Zhu et al., 2024).

Всеки от петте минитеста е решен трикратно от всеки модел, като всяко решаване е проведено в нова сесия при изключена постоянна памет, което осигурява процедурна независимост, използвана и в предходно изследване при сравнителна работа с няколко LLM (Маринов/Marinov, 2026). Така са получени 15 сесии на модел, или общо 60, а за всяка задача са налични по три отделни отговора. Трикратното решаване е избрано, за да не се основава оценката на единичен отговор на генеративния модел и същевременно да се запази управляем обем на корпуса. Всички получени отговори са включени при изчисляването на показателите за представянето на моделите. Устойчивостта на резултатите между трите отделни решавания е предмет на самостоятелно изследване и не се анализира тук.

При експеримента с BgGPT не беше установена възможност за деактивиране на достъпа до външни източници, а при част от отговорите системата визуализираше източници. Затова не може да се гарантира, че всички отговори са генерирани само въз основа на знанията, придобити при обучението, без външна информация. По време на експеримента не е извършвано целенасочено търсене и не са давани инструкции за използване на външни източници.

При всички модели и сесии е използвана една и съща инструкция: „Реши задачи № 18 – 24 от приложения файл. За всяка задача посочи отговора, който според теб е верен, и го обоснови накратко.“ Оценяван е първият генериран отговор, без последващи уточняващи или коригиращи въпроси. Също така не са давани допълнителни указания за вида на изпита, учебното равнище, литературния контекст или начина на решаване. Изискването за кратка обосновка е включено с оглед на предходно наше изследване, установило случаи на правилен краен отговор, придружен от неправилно обяснение (Marinov, in press). Връзката между правилността на отговора и валидността на генерираната обосновка ще бъде разгледана в отделно изследване и не е предмет на настоящия анализ.

За задачи № 18 – 23 всеки отговор е оценен бинарно спрямо официалния ключ на МОН: 1 за правилен и 0 за неправилен. Задача № 24 изисква четири кратки отговора (А – Г), всеки от които е оценен отделно като правилен или неправилен спрямо предварително подготвения ключ. След трикратно решаване на тестовете за всяка от петте години резултатът е общо 60 оценени кратки отговора за всеки модел. За да има задача № 24 същата тежест в общия резултат като всяка от останалите задачи, резултатът от четирите ѝ отговора при всяко решаване е преобразуван в стойност от 0 до 1: 0 правилни отговора = 0; 1 = 0,25; 2 = 0,50; 3 = 0,75; 4 = 1. Така всяка

от седемте задачи в един тест има максимална стойност 1, независимо от формата си. Получената стойност по-нататък се означава като общ (нормализиран) резултат.

При трикратно решаване на всеки от петте теста максималният резултат за един модел е 105 точки (35 задачи × 3), а общият процент се изчислява като отношение между получените и максимално възможните точки. При отделния анализ на задача № 24 резултатът се изчислява върху 60-те кратки отговора за всеки модел, а при изчисляването на общия резултат четирите отговора носят общо максимум 1 точка, така че задача № 24 има същата тежест като всяка от задачи № 18 – 23. Разликите в получените проценти се използват за описание и сравнение на представянето на моделите при решаването на задачите в изследвания корпус, без да се прави статистически извод дали установените различия са устойчиви и извън него.

За допълнителния анализ затворените задачи № 18 – 23 са групирани според това какъв тип работа с литературното съдържание изисква задачата. Разграничени са пет обобщени категории: разпознаване и възпроизвеждане на конкретни знания за произведението; определяне на жанр или литературно понятие; характеристика на герой или друг елемент от произведението; разпознаване на тема или мотив; интерпретация на смисъла на конкретен текстов елемент. Групирането има аналитична функция и не представлява самостоятелна таксономия на литературните компетентности или йерархия на когнитивна сложност.

3. Резултати

Резултатите показват различия в точността на отговорите както между четирите модела, така и между задачите с избираем и със свободен отговор. При затворените задачи № 18 – 23 най-висок резултат постига ChatGPT (91,1%), следван от Gemini (90,0%), BgGPT (74,4%) и Claude (64,4%). При задача № 24, която изисква свободно генериране на четири кратки отговора, резултатите са по-ниски при всички модели: Gemini достига 76,7%, ChatGPT – 55,0%, Claude – 43,3%, а BgGPT – 35,0%. Разликата спрямо затворените задачи е най-малка при Gemini (13,3 процентни пункта), следван от Claude (21,1), ChatGPT (36,1) и BgGPT (39,4 процентни пункта).

Таблица 1. Представяне на моделите според формата на задачите и общ резултат

Модел	**Задачи № 18 – 23 (%)**	**Задача № 24 (%)**	**Общо (%)**
BgGPT	74,4	35,0	68,8
Gemini	90,0	76,7	88,1
Claude	64,4	43,3	61,4
ChatGPT	91,1	55,0	86,0

Общият резултат се изчислява върху всички еднакво претеглени задачи и затова не представлява средно аритметично на двата процента в таблицата. За получаването му процентът за задачи № 18 – 23 се умножава по 6, към него се прибавя процентът за задача № 24, а полученият сбор се

разделя на общия брой задачи – 7. Най-висок резултат постига Gemini (88,1%), следван от ChatGPT (86,0%), BgGPT (68,8%) и Claude (61,4%). Тези резултати не потвърждават работната хипотеза за предимство на BgGPT – в изследвания корпус адаптираният към българския език модел се представя по-слабо от Gemini и ChatGPT, но по-добре от Claude.

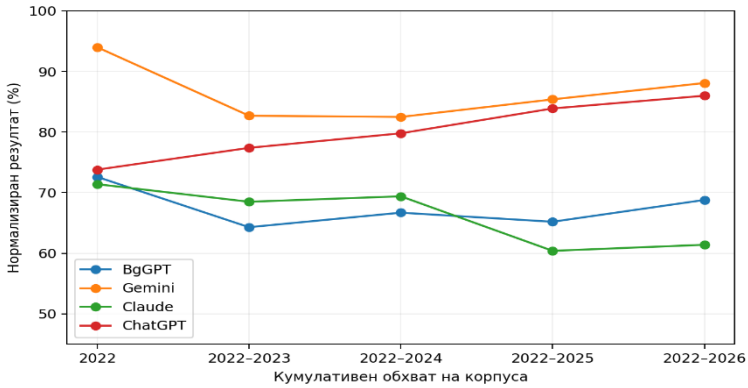
Групирането на затворените задачи според типа работа с литературното съдържание показва, че различията между моделите не са еднакви във всички категории. За по-детайлен анализ резултатите от затворените задачи № 18 – 23, представени обобщено в *Таблица 1*, са разпределени в *Таблица 2* според типа работа с литературното съдържание. При задачите за определяне на жанр или литературно понятие всички модели достигат 100%. Тази категория обаче включва само две задачи, поради което резултатът не се интерпретира самостоятелно. При задачите за характеристика на герой или друг елемент от произведението Gemini и ChatGPT достигат 100%, BgGPT – 95,8%, а Claude – 87,5%.

Таблица 2. Представяне на моделите според типа работа с литературното съдържание

Тип работа	**Брой задачи (n)**	**BgGPT**	**Gemini**	**Claude**	**ChatGPT**
Жанр/понятие	2	100,0	100,0	100,0	100,0
Характеристика	8	95,8	100,0	87,5	100,0
Разпознаване/възпроизвеждане	6	72,2	100,0	55,6	94,4
Тема/мотив	8	70,8	87,5	58,3	87,5
Интерпретация	6	44,4	66,7	38,9	77,8

Най-ниски стойности и при четирите модела се наблюдават при задачите, изискващи интерпретация на конкретни елементи от художествения текст: BgGPT – 44,4%, Gemini – 66,7%, Claude – 38,9% и ChatGPT – 77,8%. Тъй като броят на задачите в категориите е малък и неравномерен, различията са описателни и не дават основание за общи изводи относно трудността на различните типове работа с литературното съдържание.

За да се провери доколко общото представяне зависи от включените изпитни години, общият резултат за задачи № 18 – 24 е преизчислен при постепенно разширяване на корпуса. Първо е изчислен процентът само за задачите от 2022 г., след това общо за 2022 – 2023 г., 2022 – 2024 г., 2022 – 2025 г. и накрая за целия период 2022 – 2026 г. Така се проследява как се променят резултатите и подреждането на моделите с разширяването на корпуса.



Фигура 1. Промяна на общия (нормализирания) резултат за задачи № 18 – 24 при последователното включване на изпитните години

При използване само на теста от 2022 г. Gemini постига най-висок резултат (94,0%), докато останалите три модела са значително по-близки помежду си. С включването на следващите изпитни години различията между тях се очертават по-ясно, като резултатите на ChatGPT се доближават до тези на Gemini, а BgGPT и Claude остават на по-ниски позиции. Същевременно подреждането не е постоянно през целия период: при по-ограничен корпус отделни модели разменят местата си, което показва влияние на конкретния подбор на включените задачи. След добавянето на данните за 2025 г. моделите се подреждат в реда Gemini, ChatGPT, BgGPT и Claude и това подреждане се запазва и при включването на 2026 г. Анализът потвърждава значението на обхвата на корпуса – при последователното му разширяване резултатите и подреждането на моделите се променят. Това е особено важно при близки резултати, тъй като малък корпус може да придаде прекомерна тежест на особеностите на задачите от конкретна година.

4. Дискусия

Основният извод от изследването е, че адаптацията на BgGPT към българския език не му осигурява предимство спрямо всички изследвани общи многоезични модели при стандартизирани задачи върху българската литература. Gemini и ChatGPT се представят по-добре от BgGPT в общото оценяване, докато Claude остава след него, което показва, че разграничението между адаптиран към българския език и общ многоезичен модел не съответства пряко на установеното подреждане на четирите системи. Този резултат съответства на предходни наблюдения, че предимството на модели, специализирани към определен език, не се проявява последователно при различни задачи, включително при оценяване на културно специфично знание (Blevins et al., 2024; Chang et al., 2024; Chiu et al., 2025). Настоящият дизайн позволява да се сравни представянето на четирите реално достъпни системи при еднаква процедура, но не и да се

изолира влиянието на отделните им технически характеристики и динамиката в развитието им.

Приносът на изследването е в експерименталната проверка на връзката между езиковата адаптация на големите езикови модели и представянето им при стандартизирани задачи по българска литература, включени в корпус от тестове от НВО за 7. клас. От образователна гледна точка установените различия показват необходимостта възможностите на LLM да се оценяват чрез задачи, създадени в конкретния езиков и предметен контекст, в който се предвижда тяхното използване. Адаптацията към даден език или доброто общо представяне на един модел не са достатъчни основания да се предполага, че той ще работи еднакво успешно с конкретно учебно съдържание. При използването на LLM в обучението по български език и литература това предполага точността на генерираното съдържание да се проверява спрямо съответните предметни знания и изисквания, а не да се приема като даденост въз основа на характеристиките или заявената специализация на модела.

5. Заключение

Преди да бъдат направени изводите, трябва да се отбележат и няколко ограничения при интерпретацията на получените резултати. Корпусът включва стандартизирани задачи от пет последователни изпитни години и не обхваща всички възможни форми на работа с българската литература. Използвани са публично достъпни тестови материали, поради което не може да се изключи предварително познаване на задачите от моделите. Сравнението се отнася до безплатно достъпните към 4 юли 2026 г. версии на системите, при BgGPT не беше установена възможност за изключване на достъпа до външни източници, а различията между затворените задачи и задача № 24 не могат да бъдат приписани единствено на формата на отговора. Анализът е ограничен до точността на отговорите – валидността на обосновките и устойчивостта на представянето при повторни решавания са предмет на отделни изследвания.

В резултат от проведеното изследване, в рамките на посочените ограничения, могат да бъдат направени следните основни изводи:

1) Работната хипотеза не се потвърди – адаптираният към българския език BgGPT не показва предимство пред трите общи многоезични системи при решаването на стандартизирани задачи върху българската литература. Gemini и ChatGPT се представят по-добре от BgGPT, докато Claude остава след него.

2) Представянето на моделите се различава, а подреждането им се променя в зависимост от обхвата и състава на задачите. Това показва, че сравнителната оценка на LLM може да зависи от конкретния обем на използвания корпус.

3) Езиковата адаптация сама по себе си не е достатъчна, за да се предвиди представянето на LLM при конкретно предметно и културно специфично съдържание. Затова възможностите на големите езикови

модели трябва да бъдат проверявани чрез задачи от езиковия, културния и образователния контекст, в който ще се използва съответният LLM.

4) Към момента на изследването (юли 2026 г.) използването на големи езикови модели в обучението по българска литература изисква критична оценка и проверка на генерираното съдържание, тъй като нито един от изследваните модели не показва напълно надеждно представяне при всички разглеждани типове задачи. Дори високият общ резултат на конкретен модел не изключва допускането на грешки в конкретни случаи.

БИБЛИОГРАФИЯ:

Ангелова, Т. (2026). Как да помогнем на учениците умно да използват изкуствен интелект на занятията по български език. *Български език и литература, 68(2s),* 20 – 33. (**Angelova, T.** How to help students use artificial intelligence wisely in Bulgarian language classes. *Balgarski ezik i literatura [Bulgarian Language and Literature], 68(2s),* 20 – 33. https://doi.org/10.53656/bel2026-2s-3)

Маринов, В. (2026). Координирано използване на големи езикови модели в академичната работа. *Медии и език, 00,* 1 – 28. (**Marinov, V.** Coordinated integration of large language models in academic work. *Medii i ezik*, 00, 1 – 28. https://doi.org/10.58894/dvzte852)

Министерство на образованието и науката. (б.д.). *Тестове и верни отговори от НВО за VII клас по години.* Посетено на 4 юли 2026 г. от https://www.mon.bg/obshto-obrazovanie/natsionalno-vanshno-otsenyavane-nvo/nvo-za-vii-klas/testove-i-verni-otgovori-ot-nvo-za-vii-klas-po-godini/ (**Ministry of Education and Science. (n.d.).** *Tests and answer keys from the National External Assessment for Grade 7, by year*. Retrieved July 4, 2026, from https://www.mon.bg/obshto-obrazovanie/natsionalno-vanshno-otsenyavane-nvo/nvo-za-vii-klas/testove-i-verni-otgovori-ot-nvo-za-vii-klas-po-godini/)

Трайчева, Б. (2026). Изследване на нагласи и умения за работа с езиков изкуствен интелект в обучението по български език. *Български език и литература, 68(2s),* 34 – 46. (**Traycheva, B.** A study of attitudes and skills for working with linguistic artificial intelligence in Bulgarian language education. *Balgarski ezik i literatura [Bulgarian Language and Literature], 68(2s)*, 34 – 46. https://doi.org/10.53656/bel2026-2s-4)

Alexandrov, A., Raychev, V., Dimitrov, D. I., Zhang, C., Vechev, M., & Toutanova, K. (2024). *BgGPT 1.0: Extending English-centric LLMs to other languages.* arXiv. https://arxiv.org/abs/2412.10893

Blevins, T., Limisiewicz, T., Gururangan, S., Li, M., Gonen, H., Smith, N. A., & Zettlemoyer, L. (2024). Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 10822–10837). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.604

Chang, T. A., Arnett, C., Tu, Z., & Bergen, B. K. (2024). When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4074–4096). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.236

Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., & Choi, Y. (2025). CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 25663–25701). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1247

Koeva, S. P., Stoyanova, I., Georgiev, D., Leseva, S., Stefanova, V., Todorova, M., Dimitrova, T. I., Kukova, H., Moskova, M., & Tinchev, T. (2026). Bulgarian Massive Multitask Language Understanding Benchmark. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (pp. 4658–4672). ELRA Language Resource Association. https://doi.org/10.63317/57jy4fetjmc4

Liu, C., Zhang, W., Ying, J., Aljunied, M., Luu, A. T., & Bing, L. (2025). SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 6134–6151). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.341

Marinov, V. (in press). *Educational potential and constraints of large language models in Bulgarian language education*. Unpublished manuscript.

Osenova, P. (2025). Contemporary Bulgarian grammar and some context-related usage features. *Journal of Global Sociolinguistics, 1(2),* 1–7. https://hdl.handle.net/1887/4290853

Todoranova, A., & Todoranova, L. (2026). Assessing AI's Performance: Implications for Educational Use. *Baltic Journal of Modern Computing, 14(2)*, 402–416. https://doi.org/10.22364/bjmc.2026.14.2.07

Zhu, Q., Cheng, Q., Peng, R., Li, X., Peng, R., Liu, T., Qiu, X., & Huang, X. (2024). Inference-time decontamination: Reusing leaked benchmarks for large language model evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 9113–9129). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.532

Copyright © 2026 Marinov. This is an open-access publication distributed under the terms of the Creative Commons Attribution 4.0 International Licence CC BY 4.0