Towards a New Grammar Checker for Bulgarian

The BulMMLU:gr Benchmark for Grammatical Error Detection and Correction

Svetla Koeva · Institute for Bulgarian Language, Bulgarian Academy of Sciences

LLMs edit text – but do they get Bulgarian grammar right?

Large Language Models can paraphrase, shorten, and simplify text with variable accuracy. AI writing tools add explicit grammar-checking – but coverage for Bulgarian remains thin and largely unverified.

- Grammar error detection & correction is well established for English
- Very few publications describe fine-tuning for Bulgarian-specific grammar
- No open, comprehensive Bulgarian grammar-checking service exists yet

**На 1 януари 2026 г. еврото ще влезе в обръщение...*

обръщение (address) used instead of обращение (circulation)

A paronym error that requires full semantic analysis to catch — not a spell-check.

Two aims of this study

- **Assess Detection Capability –** examine whether Bulgarian-supporting grammar checkers and selected LLMs can correctly identify specific spelling, grammatical, and punctuation errors

- **Chart the Path Forward –** analyse the current state of Bulgarian grammar checking and identify future directions for a dedicated AI-based grammar service

From rule-based checkers to neural GEC

Grammatical error detection (GED) and correction (GEC) have been studied extensively for English, with a defined error taxonomy and standardised evaluation (Bryant et al., 2017, 2019). Modern approaches increasingly formulate correction as sequence tagging or sequence-to-sequence generation.

GECToR	Tag, not rewrite — sequence-tagging GEC (Omelianchuk et al., 2020)
Multilingual expansion	Turkish (Kara et al., 2023), Chinese (Tian & Zhou, 2025), low-resource Indian languages (Sharma & Bhattacharyya, 2025)
Synthetic parallel data	LegalGEC: 350K synthetic Romanian legal-domain pairs (Timpuriu et al., 2026), synthetic Bulgarian error-correction datasets (Klouchek & Batista-Navarro 2024))
OmniGEC	Silver-standard GEC across 11 languages; fine-tunes Aya-Expanse & Gemma-3 (Kovalchuk et al., 2025)

Where BulMMLU:gr fits in

Two-stage training – pre-training on large synthetic datasets followed by fine-tuning on smaller gold-standard corpora (Ingolfsdóttir et al., 2023) – is now the dominant strategy for GEC. BulMMLU:gr is designed to serve precisely this second, gold-standard role for Bulgarian.

Synthetic Data

Rule-based corruption, back-translation, or LLM-based generation create large-scale noisy training pairs (Stahlberg & Kumar, 2021)

Gold-Standard Fine-tuning

Small, high-quality, human-annotated corpora refine the model – this is BulMMLU:gr's role

Proof of Concept

Gemma 3 27B fine-tuned on a BulMMLU:gr-derived set, published on GitHub (DCL-IBL/BG-LLMs)

Modelled on MMLU, built for Bulgarian

- 118 questions across 6 error types, 4 answer options each, 1 correct answer
- Component of the wider BulMMLU benchmark (Bulgarian history, literature, economics, language)
- Design principles
  - diverse difficulty, from semantic analysis to pure spelling
  - anguage-specific grammar, spelling, and punctuation rules
  - errors resolved via sentence analysis or external knowledge

Six types of Bulgarian grammatical error

Paronym confusion

One-letter-different word pairs, e.g. *обръщение* vs *обращение*

Comma with така / че / така че

Depends on whether *така че* is one conjunction or two

Adverb+adjective spelling

Fused (*светлозелен*) vs separate (*прясно боядисан*)

Adjective–abbreviation agreement

Gender/number must follow the head noun, e.g. *новият ДДС*

Fused/separate noun compounds

геймконзола (fused only) vs *еврозона*/*евро зона* (both allowed)

Foreign-origin plural forms

две уискита, not **две уиски*

Type 1: Paronym confusion

**На 1 януари 2026 г. еврото ще влезе в обръщение за парични плащания.*

Обръщение (address/salutation) → should be *обращение* (circulation)

The correct word can only be selected after analysing the syntactic and semantic structure of the sentence – no rule-based check alone can resolve it.

Detection difficulty

- **Spelling rule**
- **Grammar rule**
- **Sentence context**
- **External knowledge**

Type 2: Comma with така / че

*Направете, така че законът да бъде приет.

така as adverb + че conjunction, vs. така че as one compound conjunction

Correct punctuation depends entirely on the grammatical function of *така че* in context – identical surface string, two different structures.

Detection difficulty

- **Spelling rule**
- **Grammar rule**
- **Sentence context**
- **External knowledge**

Type 3: Fused vs. separate adverb+adjective

особено опасен (separate) ↔ *светлозелен (fused)*

Fused only when the pair forms a single lexicalised concept and cannot be reordered

Not all fused compound adjectives appear in orthographic dictionaries – rule-based checkers struggle here by design.

Detection difficulty

- **Spelling rule**
- **Grammar rule**
- **Sentence context**
- **External knowledge**

Type 4: Adjective–abbreviation agreement

**новото ДДС → новият ДДС (данък добавена стойност,* masc.*)*

Agreement follows the gender of the noun the abbreviation stands for – not the abbreviation's own surface form.

Requires knowledge of what the abbreviation actually stands for – a purely orthographic check cannot catch this.

Detection difficulty

- **Spelling rule**
- **Grammar rule**
- **Sentence context**
- **External knowledge**

Type 5: Fused / separate noun compounds

геймконзола (fused only) vs. *еврозона / евро зона* (both allowed)

гейм 'game' is not used independently; *евро* 'euro' is – hence the difference

Requires knowing which foreign-origin components function as independent words in Bulgarian – dictionaries lag behind new coinages.

Detection difficulty

- **Spelling rule**
- **Grammar rule**
- **Sentence context**
- **External knowledge**

Type 6: Foreign-origin plural forms

*две уиски → две уискита

*много билборди → много билбордове

Recently borrowed nouns whose plural endings are not yet firmly established

Some correct plurals are already dictionary-recorded; others are not, making this a moving target for any static rule set.

Detection difficulty

- **Spelling rule**
- **Grammar rule**
- **Sentence context**
- **External knowledge**

Structure, split, and provenance

- Format: CSV, MMLU-style, converted into a fine-tuning set
- Built manually, with semi-automatic support
- Released under CC BY 4.0

Six applications surveyed

Tool	Languages	Bulgarian?	Integration
Grammarly	24	No	Web, desktop, mobile
InstaText	40+	Yes	Web, editor plugins
LanguageTool	31	No	Web, browser, docs
QuillBot	5	No	Web, docs
Scribbr	5	No	Web (via QuillBot)
Scribens	40+	Yes	Web, browser

Bulgarian support relies almost entirely on multilingual coverage claims (40+ languages) rather than dedicated linguistic engineering.

Four possible outcomes per sentence

1. Correct detection + correct suggestion

Error identified and intended correction proposed

2. Correct detection + wrong suggestion

Error location right, replacement is not

3. Failure to detect

An existing error is missed entirely

4. False detection (overgeneration)

A correct language unit is incorrectly flagged

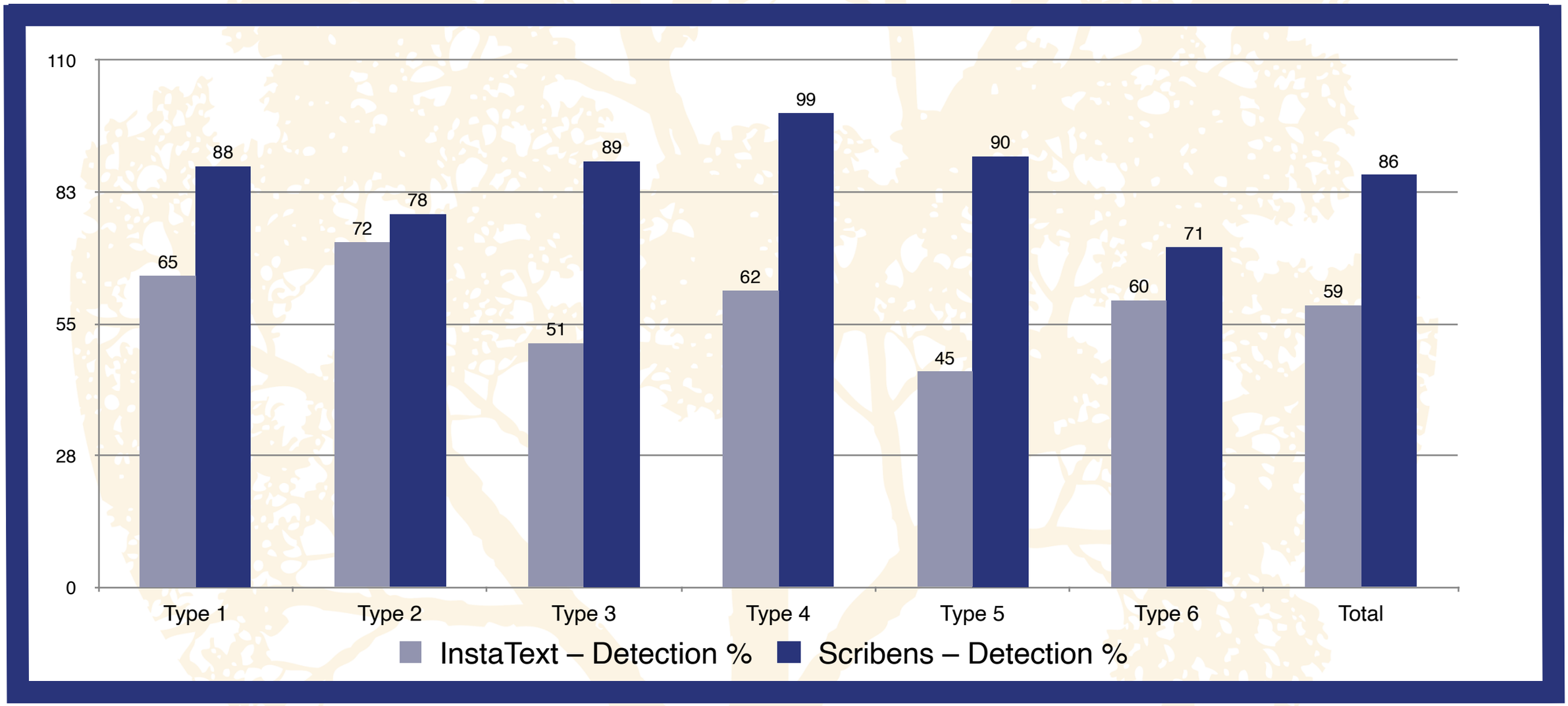
Metrics

Detection rate
= (1+2) / total

Correct-suggestion rate = 1 / total

Case 4 is reported separately and excluded from both rates.

Scribens outperforms InstaText



Where both applications fail

**в киносалона на гейм клуба прожектираха кино документалистика за гейм индустрията...*

****гейм клуб / *гейм индустрия***

Both apps treat these incorrect separate spellings
as correct.

****кино документалистика***

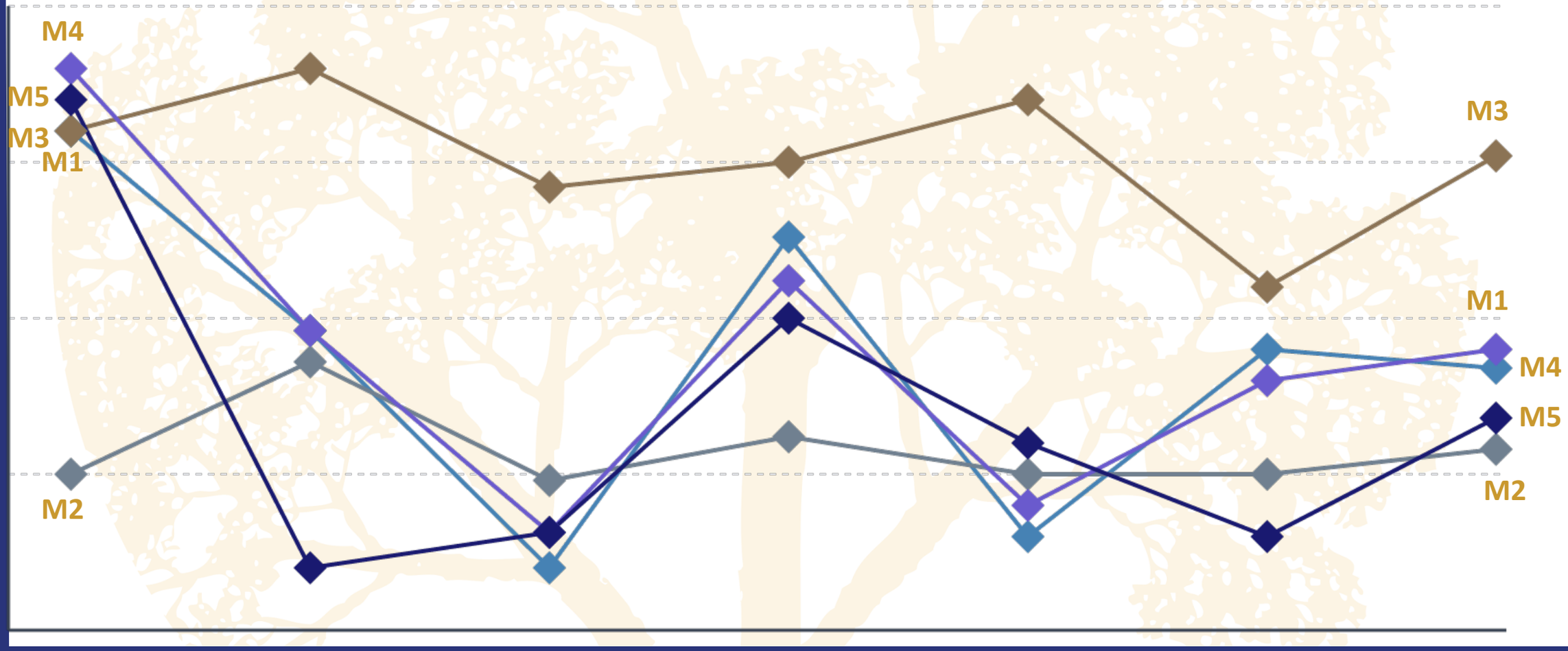
Should be one word. InstaText suggests a different phrase; Scribens flags it as a stylistic calque, not a spelling error.

Five models spanning open-weight and proprietary

M1	**Gemma 4 31B IT**	Google DeepMind · open-weight · 140+ languages · 256K context
M2	**BgGPT 3.0 27B**	INSAIT · open-weight · Bulgarian-specialised · 131K context
M3	**Claude Sonnet 4.6**	Anthropic · API · reasoning & agentic workflows · 1M context
M4	**Gemini 3.5 Flash**	Google DeepMind · API · low-latency · multimodal · 1M context
M5	**GPT-5.5**	OpenAI · API · general-purpose reasoning · 1.05M context

Prompted with the full question and four options; instructed to answer with a single letter (A–D). No few-shot examples shown.

M3 leads, but not everywhere



BulMMLU:gr – Grammar Checker for Bulgarian

Strength is task-specific, not general

M3 – most balanced

Best on 5 of 6 types (2–6); strong on Type 5 (0.85) and Type 2 (0.91).

M1 – sharp but narrow

Best on Type 1 (0.80) yet collapses on Type 5 (0.15) and Type 3 (0.10).

M2 – consistently weak

Below 0.50 on every single type; 0.29 overall.

Key findings

No single model achieves the best accuracy across all six error types.

General-purpose models are not pre-trained or fine-tuned for Bulgarian grammar – performance likely reflects training data and architecture, not language-specific design.

The McNemar test

Accuracy differences could arise from random variation. The McNemar test compares two classifiers on the same instances, focusing on disagreements – questions one model gets right and the other gets wrong – to test statistical significance.

We use the exact binomial form, computing the p-value directly from n = b + c disagreements with success probability 0.5.

Reading the test

- **$p < 0.05$**
  Statistically significant difference
- **$p \geq 0.05$**
  No significant difference – could be chance
- **No correction applied**
  Borderline p-values would not survive Bonferroni correction

M3's edge is statistically significant

	M1	M2	M3	M4	M5
M1	—	**0.0328***	**<0.0001***	0.6476	0.0872
M2	**0.0328***	—	**<0.0001***	**0.0094***	0.4709
M3	**<0.0001***	**<0.0001***	—	**<0.0001***	**<0.0001***
M4	0.6476	**0.0094***	**<0.0001***	—	**0.0146***
M5	0.0872	0.4709	**<0.0001***	**0.0146***	—

** significant at $p < 0.05$ · M3 differs significantly from every other model ($p < 0.0001$ in each case)*

Comparison by type

Type	M3 vs M1	M3 vs M2	M3 vs M4	M3 vs M5
Type 1	1.0000	**0.0034***	0.6875	1.0000
Type 2	**0.0039***	**0.0063***	**0.0039***	**<0.0001***
Type 3	**0.0010***	**0.0063***	**0.0034***	**0.0034***
Type 4	0.6875	**0.0391***	0.5078	0.3438
Type 5	**0.0005***	**0.0005***	**0.0010***	**0.0074***
Type 6	0.7744	0.1796	0.5811	**0.0386***

Types 2, 3, and 5 – the ones requiring context or external knowledge — show the most consistent, significant gaps in M3's favour.

Towards a dedicated Bulgarian grammar checker

- No existing tool – application or LLM – reliably handles all six Bulgarian error types
- Grammar-checking apps offer transparent, user-controlled correction that LLMs currently lack
- Open-weight models (e.g. M1) are promising fine-tuning candidates for a dedicated system

Next steps

- Systematic classification of frequent Bulgarian errors (syntax, punctuation, spelling)
- Extend BulMMLU:gr with paired incorrect/corrected sentences
- Fine-tune open-weight models (Gemma 3 27B, Gemma 4 27B) for Bulgarian GEC
- Explore quantised large open-weight models (DeepSeek-V3/V4 Pro, Kimi 3)

Thank you

- Svetla Koeva svetla@dcl.bas.bg

- Dataset: huggingface.co/datasets/DCL-IBL/BulMMLU-Dataset

- The work is part of the project *Infrastructure for Fine-tuning Pre-trained Large Language Models*, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

http://ifgpt.dcl.bas.bg

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria