Towards a new Grammar Checker for Bulgarian

Svetla Koeva
Institute for Bulgarian Language,
Bulgarian Academy of Sciences
52 Shipchenski prohod Blvd. Bldg. 17, Sofia 1113, Bulgaria
svetla@dcl.bas.bg

Abstract

This paper presents the dataset **Massive Multitask Language Understanding of Bulgarian Grammar** (**BulMMLU:gr**), modelled on the **MMLU** (Massive Multitask Language Understanding) benchmark. Its aim is to test the ability of AI grammar-checking applications, as well as different LLMs, to detect grammatical errors in Bulgarian and thus provide a basis for analysing the need for a specialised LLM-based tool to detect such errors and suggest corrections. Experiments conducted with two grammar-checking applications and five LLMs reveal limitations that highlight the need for a new, dedicated Bulgarian grammar-checking system. BulMMLU:gr can be further extended and used both for evaluation and for fine-tuning LLMs for grammatical error detection and correction suggestions.

Keywords: grammatical error detection and correction suggestions; grammar-checking applications; Large Language Models

1 Introduction

Large Language Models (LLMs) can edit and transform texts by paraphrasing, shortening, and simplifying them, with varying accuracy depending on the model's purpose, size, and language support. In addition, several artificial intelligence-based applications offer tools for improving the grammatical quality of written text, including grammar error detection and correction suggestions.

The **aims of this study** are: (i) to examine the extent to which applications with integrated grammar checkers that support Bulgarian, and selected LLMs, can correctly identify specific types of spelling, grammatical, and punctuation errors;[1] and (ii) to analyse the current state of grammar checking for Bulgarian and to identify future directions for developing a new AI-based grammar service dedicated to Bulgarian.

For this purpose, a dataset called **Massive Multitask Language Understanding of Bulgarian Grammar** (**BulMMLU:gr**), based on the **Massive Multitask Language Understanding** (**MMLU**) benchmark (Hendrycks et al., 2021), has been created. It is used to evaluate the ability of specialised AI-based grammar-checking applications and a set of LLMs (both proprietary and open-weight models) to detect grammatical errors in Bulgarian. This empirical evaluation is particularly important because very few publications describe the pre-training and fine-tuning procedures of state-of-the-art LLMs, and even fewer discuss specialised grammar-checking applications.

The main contributions of this work are as follows:

- **Development of the BulMMLU:gr** benchmark for Bulgarian, including questions designed to cover six types of grammatical, spelling, and punctuation errors in Bulgarian. The dataset has a standardised format, following the MMLU format.
- **Accuracy comparison** of two specialised Bulgarian grammar-checking applications and five LLMs (including proprietary and open-weight models) on the BulMMLU:gr benchmark.

The analysis indicates that proprietary LLMs and one existing grammar service demonstrate reasonably good, though imperfect, performance in detecting errors in Bulgarian grammar. However, there is currently no widely available open-source online service providing comprehensive Bulgarian grammar checking, nor any open-weight LLM that achieves comparably high accuracy.

[1] In this study, we count as grammatical not only purely grammatical errors but also spelling and punctuation errors that are related to grammar.

Therefore, it is necessary to expand the BulMMLU:gr dataset to cover a broader – ideally comprehensive – range of Bulgarian grammatical phenomena, and to use it for both testing and fine-tuning suitable language models to create a reliable tool for grammar checking in Bulgarian.

2 Related work

Traditional rule-based and statistical grammar checkers have gradually evolved into AI-based applications. They now offer a range of features that help users improve style and word choice, correct grammatical errors, and make content more readable and understandable. LLMs can naturally improve grammar, style, and fluency; however, they often alter the original text and, in some cases, the author's intended meaning may be lost, transformed, or even combined with incorrect statements. In contrast, AI-based grammar checkers improve clarity, readability, and flow by explicitly presenting each suggested edit and allowing authors to decide whether to accept or reject it.

Two essential elements of grammar checking have been studied for many years: grammatical error detection and grammatical error correction (Bryant et al., 2019; Rothe et al., 2021). Research on grammatical error detection and correction for English is well established, as are many other NLP tasks. In addition, a well-defined error taxonomy and standardised evaluation mechanism have been developed for English (Bryant et al., 2017). Error classification methods and approaches for transforming erroneous sequences into correct ones, based on the nature of the error, have also been proposed (Omelianchuk et al., 2020).

Recently, several publications have proposed approaches to grammatical error correction for languages other than English, including LLM-based methods for Turkish (Kara et al., 2023) and Chinese (Tian and Zhou, 2025), as well as synthetic data generation based on the Mask-Translate-and-Fill framework for low-resource Indian languages (Sharma and Bhattacharyya, 2025). For Bulgarian, mT5 was fine-tuned specifically for grammatical error correction (Klouchek and Batista-Navarro, 2024). The training data includes errors involving definite articles, pronouns, agreement, and verb morphology from a synthetic Bulgarian error-correction dataset. The fine-tuned model reportedly achieves an F0.5 score of 68.18% and outperforms the evaluated LLM baselines.

One of the most widely used modern approaches to training grammatical error correction models is to create synthetic parallel datasets containing pairs of erroneous and corrected sentences (Bondarenko et al., 2023). For example, RoLegalGEC, a synthetic Romanian-language parallel dataset for grammatical error detection and correction in the legal domain, contains 350,000 synthetically generated text pairs with error annotations (Timpuriu et al., 2026). The dataset was used to train and evaluate several grammatical error detection and correction models, including knowledge-distilled Transformer models, sequence-tagging architectures for error detection, and pre-trained sequence-to-sequence Transformer models for grammatical error correction.

Another example is OmniGEC, a collection of multilingual silver-standard datasets for grammatical error correction covering eleven languages: Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Slovene, Swedish, and Ukrainian (Kovalchuk et al., 2025). These datasets were developed to help adapt English grammatical error correction methods to multilingual settings. The source texts are mainly from Wikipedia and Reddit in the eleven target languages, together with an additional Ukrainian-only dataset sourced from social media. The Wikipedia edits were derived from human corrections, while the Reddit and Ukrainian-only data were automatically corrected using the GPT-4o-mini model. The quality of the corrections was evaluated both automatically and manually, and the data were further used to fine-tune two open-weight large language models, Aya-Expanse (8B) and Gemma-3 (12B), which were reported to achieve state-of-the-art results for paragraph-level multilingual grammatical error correction.

Another approach combines synthetic and gold data by either (i) fine-tuning a large language model in two stages – first on synthetic data and then on gold data – or (ii) training a small sequence-to-sequence (seq2seq) Transformer model that is first pre-trained on synthetic data and subsequently fine-tuned on gold data (Palma Gomez et al., 2023).

In summary, recent approaches to grammatical error detection and correction suggestions mainly rely on supervised fine-tuning using parallel corpora of erroneous and corrected sentence pairs. In this framework, grammatical error correction is typically formulated either as a sequence-to-sequence

task, in which models directly generate corrected text, or as a sequence-tagging task, in which models predict token-level edit operations (Omelianchuk et al., 2020). Owing to the scarcity of high-quality annotated data, synthetic data generation is widely employed. Errors are typically introduced using rule-based methods, back-translation, or, increasingly, LLM-based generation, enabling the creation of large-scale synthetic training datasets (Stahlberg and Kumar, 2021).

To further improve performance, many systems adopt a two-stage training strategy in which models are first pre-trained on large synthetic datasets and then fine-tuned on smaller, high-quality, human-annotated corpora (Ingólfsdóttir et al., 2023). The BulMMLU:gr dataset can be readily converted into a gold-standard fine-tuning dataset, as it contains erroneous sentences together with their corresponding corrections for the targeted error types. Its usefulness for this purpose could be further enhanced by expanding the dataset to cover a broader range of grammatical error types. As a proof of concept, Gemma 3 27B was fine-tuned using a version of BulMMLU:gr redesigned as a fine-tuning dataset.[2]

3 Description of the BulMMLU:gr dataset

The **Bulgarian Massive Multitask Language Understanding benchmark** (**MMLU-BG**) was developed to evaluate whether an LLM possesses general knowledge beyond simple text prediction in Bulgarian (Koeva et al., 2026). The original **MMLU** (**Massive Multitask Language Understanding**) (Hendrycks et al., 2021) covers 57 diverse subjects across multiple domains and difficulty levels. The questions are divided into a few-shot development set, a validation set for selecting hyperparameters, and a test set. Although **MMLU-BG** is well suited to assessing the general knowledge of LLMs, as it is a human translation of the original MMLU into Bulgarian, it does not reflect features specific to Bulgarian history, culture, and language.

In contrast, the newly developed dataset, **Massive Multitask Language Understanding for Bulgaria and Bulgarian** (**BulMMLU**), evaluates LLMs' knowledge of Bulgarian history, literature, economics, and the Bulgarian language. Its component, **Massive Multitask Language Understanding of Bulgarian Grammar** (**BulMMLU:gr**), provides a benchmark for evaluating specific aspects of Bulgarian grammar, spelling, punctuation, and context-dependent usage.

Six main types of grammatical error are included in the dataset for evaluation, selected according to the following principles:

- To provide diverse cases for evaluation, ranging from those that can be resolved by semantic analysis of the context and grammar rules to purely spelling-related errors.
- To test language-specific grammar, spelling, and punctuation rules.
- To evaluate correct grammar and spelling rules that depend on sentence analysis or external knowledge.

Type 1: Incorrect use of a paronym from a pair differing by only one letter. For example, in the Bulgarian sentence **На 1 януари 2026 г. еврото ще влезе в обръщение за парични плащания в страната*, 'On 1 January 2026, the euro will enter into circulation for cash payments in the country', the word *обръщение*, 'a word or expression with which the speaker addresses the listener', is incorrectly used instead of the word *обращение*, 'the process of use or exchange of money and goods; turnover'. The appropriate word can be selected only after analysing the syntactic and semantic structure of the sentence in which the target word is used and cannot be determined by applying rule-based grammar or spelling rules alone.

Type 2: Incorrect comma placement involving the adverb *така* 'in such a way', the subordinating conjunction *че* 'that', and the compound subordinating conjunction *така че* 'so'. For example: **Направете, така че законът да бъде приет* 'Do it in such a way that the law is passed'; **Не съм поканена на рождения ден така, че няма да дойда* 'I am not invited to the birthday party, so I will not come'. The correct punctuation depends on whether *така* functions as an adverb followed by the conjunction *че*, or whether *така че* functions as a compound conjunction. Correctly identifying these errors requires syntactic and semantic analysis of the sentence to determine the grammatical function of these expressions.

Type 3: Incorrect fused or separate spelling of an adverb and an adjective. Combinations of an adverb and an adjective (including adjectival participles) are written separately when they do not denote a single concept, for example, *особено*

[2] https://github.com/DCL-IBL/BG-LLMs/tree/feat/spelling/gemma3 (accessed on 15.07.26)

опасен ‘especially dangerous’ and *прясно боядисан* ‘freshly painted’. They are written as a single word when they form compound adjectives that conventionally denote a single concept, for example, *светлозелен* ‘light green’ and *световноизвестен* ‘world-famous’. In these cases, rearrangement of the components is not possible (e.g. **зелен светло*, **известен световно*). This type of error is difficult to detect because not all compound adjectives written as a single word are included in orthographic dictionaries,[3] and the choice between fused and separate spelling depends on whether the components denote a single lexicalised concept or form a phrase. Consequently, such errors are hardly detectable by rule-based grammar checkers.

Type 4: Incorrect agreement between an adjective and an abbreviation of a multi-word proper name. The adjective agrees in gender and number with the head noun of the multi-word proper name, even when the expression is abbreviated. For example, **новото ДДС* (данък добавена стойност, ‘Value Added Tax’, VAT) should be replaced with *новият ДДС*, because the noun *данък* is masculine. Detecting this type of agreement error requires knowledge of the multi-word expression represented by the abbreviation, including the grammatical gender and number of its head noun.

Type 5: Incorrect fused, semi-fused, or separate spelling of two nouns. Compound nouns in which one component is of foreign origin and not used as an independent word are written as one word. If the first component of the compound noun is of foreign origin and also used as an independent word, both fused and separate spellings are allowed. For example, the fused spelling *геймконзола* (‘gaming console’), where *гейм* denotes ‘relating to computer games and video games’, is the only prescribed form, whereas both *еврозона* and *евро зона* (‘euro zone’) are allowed because *евро* means ‘the monetary unit of the euro’ and is also used independently. Some fused words are listed in modern orthographic dictionaries, but not all new forms are included, as new words regularly enter the language. Thus, detecting this spelling error requires knowledge of which words are of foreign origin and not used separately in Bulgarian.

Type 6: Incorrect use of plural forms of masculine and neuter nouns. These nouns are of foreign origin and are either relatively rare or have recently entered the language, so their forms are not yet well established. For example, **две уиски* instead of *две уискита* ‘two whiskeys’, **много билборди* instead of *много билбордове* ‘many billboards’. The words whose plural forms might be mistaken are of foreign origin and relatively new to the language. For some of them, the existence of a plural form and its correct ending is already recorded in modern orthographic dictionaries, while others are still not included in the dictionaries.

The **BulMMLU:gr** (**Massive Multitask Language Understanding of Bulgarian Grammar**) dataset provides a new resource for evaluating state-of-the-art LLMs and grammar-checking applications by comparing their performance across different aspects of Bulgarian grammar, spelling, and punctuation. **BulMMLU:gr** follows the structure and data organisation of the original MMLU dataset (Hendrycks et al., 2021). Like MMLU, BulMMLU:gr consists of three parts: Development, Validation, and Test, and is intended to facilitate few-shot experiments and the reliable evaluation of LLMs. Each part comprises 5%, 15%, and 80% of the entire dataset, respectively. The dataset questions are designed to evaluate grammar, spelling, and punctuation phenomena that depend on sentence analysis or external knowledge; each question is followed by four answers, only one of which is correct, and the correct answer is provided as a separate field. All questions contain one or two example sentences with an intentionally inserted error of one of the described types. In some questions, two elements of the sentence are targeted. The dataset was developed manually by one person, with some semi-automatic support for detecting and extracting sentences with targeted correct words or phrases and for spell-checking. Additional semi-automatic validation procedures were performed to detect issues involving Cyrillic and Latin characters and incorrect punctuation inserted at the end of answers. The dataset contains 118 questions covering all selected error types. The dataset is maintained in CSV format and can be easily converted into a fine-tuning dataset.

The **BulMMLU:gr** dataset is publicly available under the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), facilitating the replication of experiments. It is distributed as part of the broader **Massive Multitask Language Understanding for Bulgaria and Bulgarian** (**BulMMLU**) benchmark, which evaluates LLMs’ knowledge of Bulgarian history, literature,

[3] https://beron.mon.bg (accessed on 15.07.26)

economics, and the Bulgarian language.[4]

Two types of experiments are conducted with the dataset. The first evaluates grammar-checking applications that support Bulgarian, while the second evaluates five LLMs that are either pre-trained on Bulgarian data or specifically trained and fine-tuned for Bulgarian.

4 Evaluation of grammar-checking applications

Contemporary grammar-checking applications such as Grammarly, QuillBot, and LanguageTool are AI-based, typically combining machine-learning methods with rule-based approaches to detect errors and suggest improvements to grammar, spelling, punctuation, and style, while providing real-time suggestions across browsers, mobile applications, and word processors. Modern tools go beyond correction by offering features such as paraphrasing, summarisation, and plagiarism and AI-generated content detection.

4.1 Brief description of grammar-checking applications

Here we present six publicly accessible grammar-checking applications.

Grammarly is an AI-based writing assistant that combines machine learning, natural language processing, and rule-based methods to provide grammar, spelling, punctuation, style, and fluency suggestions. It supports more than 20 languages, but not Bulgarian.[5] Grammarly is available as a web, desktop, and mobile application and integrates with tools such as Microsoft Word and Google Docs. Earlier work by Grammarly researchers introduced the GECToR grammatical error correction model, based on a Transformer encoder (Omelianchuk et al., 2020).

InstaText is an AI writing assistant that improves grammar, spelling, punctuation, clarity, and readability.[6] It integrates with popular writing platforms and supports editing in several major languages, while multilingual assistance extends to more than 40 languages, including Bulgarian.

LanguageTool is a grammar and style checker that combines rule-based and machine-learning techniques to detect grammatical, spelling, punctuation, and stylistic errors.[7] It also provides AI-based paraphrasing and supports 31 languages, although Bulgarian is not currently available.

Scribens is an online grammar and spell checker that combines rule-based and AI techniques to detect spelling, grammatical, and punctuation errors.[8] It supports more than 40 languages, including Bulgarian, and is available as both a free and a premium service.

QuillBot is an AI writing assistant offering paraphrasing, grammar-checking, summarisation, translation, plagiarism detection, and citation generation.[9] It integrates with Google Docs and Microsoft Word and supports several major European languages, but not Bulgarian.

Scribbr is an academic writing platform that provides plagiarism detection, citation generation, grammar-checking, and proofreading services.[10] Its Grammar Checker is powered by QuillBot and supports several major languages, but not Bulgarian.

In summary, there are various applications for checking and correcting grammar, which commonly provide a range of services, including grammar and style checking, improving fluency, text simplification, paraphrasing in different words or styles, plagiarism detection, and text summarisation. Most of these applications offer premium modes that use more advanced algorithms, perform better, allow checking larger volumes of text, and provide additional services. English is generally supported, and many applications also support widely used languages such as French, German, and Spanish. Less frequently, applications support other languages; among the six applications surveyed, only two support Bulgarian. Most applications also offer various options for use and integration with browsers, word processing programmes, and other platforms.

Table 1 presents a brief comparison of six AI-based applications that offer grammar-checking services. The comparison includes the number of supported languages, whether Bulgarian is among them, and whether the application supports various types of use: online writing or copying texts, integration in browsers and other online services, as well as in word processing tools, indicated with +. Support for Bulgarian mostly relies on multilingual

[4] https://huggingface.co/datasets/DCL-IBL/BulMMLU-Dataset (accessed on 15.07.26)

[5] https://support.grammarly.com/ (accessed on 15.07.26)

[6] https://instatext.io (accessed on 15.07.26)

[7] https://languagetool.org/ (accessed on 15.07.26)

[8] https://www.scribens.com/ (accessed on 15.07.26)

[9] https://quillbot.com/ (accessed on 15.07.26)

[10] https://www.scribbr.com/ (accessed on 15.07.26)

Tool	Languages	BG	Use
Grammarly	24	No	Online+
InstaText	40+	Yes	Online+
LanguageTool	31	No	Online+
QuillBot	5	No	Online+
Scribbr	5	No	Online+
Scribens	40+	Yes	Online+

Table 1: Grammar-checking applications

applications, which cover more than 40 languages but often provide only limited support for Bulgarian.

4.2 Evaluation, results and analysis

Only two grammar-checking applications that support Bulgarian are evaluated with BulMMLU:gr: **InstaText** and **Scribens**. For the evaluation, the example sentences from the dataset questions are extracted and submitted to each application. The results are recorded and analysed. For Scribens, however, manual inspection of the suggested replacements is required, as the application displays them only after the user clicks on the highlighted error. Four outcomes are distinguished:

1. **Correct error detection with a correct correction suggestion** – the application correctly identifies the error and proposes the intended correction.

2. **Correct error detection with an incorrect correction suggestion** – the application correctly identifies the location of the error but proposes an incorrect replacement.

3. **Failure to detect an error** – the application does not identify an existing error.

4. **False error detection** (overgeneration) – the application incorrectly flags a correct language unit as an error.

Table 2 presents the evaluation results for the two grammar-checking applications that support Bulgarian, InstaText and Scribens. The error detection rate (Det.) is calculated as the percentage of evaluated sentences that fall into cases 1 and 2, since the error is successfully detected regardless of the quality of the suggested correction. The correct correction rate (Sugg.) is calculated as the percentage of evaluated sentences that fall into case 1 only, as these are the instances in which both the error detection and the proposed correction are correct. Case 3 represents a failure of error detection and therefore lowers the error detection rate. Case 4 corresponds to overgeneration (a false positive) and is reported separately; it is not included in the calculation of either the error detection rate or the correct correction rate.

Appl	InstaText	InstaText	Scribens	Scribens
Mode	**Det.**	**Sugg.**	**Det.**	**Sugg.**
Type 1	0.65	0.63	0.88	0.82
Type 2	0.72	0.72	0.78	0.65
Type 3	0.51	0.48	0.89	0.84
Type 4	0.62	0.62	**0.99**	**0.99**
Type 5	0.45	0.42	0.90	0.78
Type 6	0.60	0.58	0.71	0.56
Total	0.59	0.57	**0.85**	**0.77**

Table 2: Detection and suggestion rates of grammar-checking applications

The results showed that, at the time of the experiment (July 2026), Scribens outperformed InstaText in detecting most error types, performing best with Type 4 errors (incorrect agreement between an adjective and an abbreviation of a multi-word proper name). However, both applications struggled to detect the remaining error types. InstaText achieved its lowest score on Type 5 (incorrect fused, semi-fused, or separate spelling of two nouns), whereas Scribens performed worst on Type 6 (incorrect use of plural forms of masculine and neuter nouns). For example, in the sentence:

В киносалона на ***гейм клуба*** *прожектираха* ***кино документалистика*** *за* ***гейм индустрията*** *и нейното влияние върху младите* 'In the game club's cinema, they screened a documentary about the game industry and its influence on young people',

both applications treat the incorrect separate spellings of **гейм клуб* and **гейм индустрия* as correct. Moreover, instead of detecting **кино документалистика* as an error and suggesting that it be written as a single word, both applications recommend replacing it with a different phrase. InstaText suggests *documentary film*, while Scribens provides the following explanation: *The phrase is stylistically incorrect. 'Documentary cinema' is the established term, whereas 'cinema documentary' is a literal and incorrect translation from a foreign language (a calque).*

In principle, Scribens' error explanation feature

often indicates that an error has been detected but not recognised as the correct error type. Consequently, it does not suggest the appropriate correction. This distinction is not taken into account in the error analysis. For example, the explanation provided for the plural form of the word *капучино* 'cappuccino' is incorrect: *The plural of капучино in Bulgarian is formed with the ending -a (капучина).*

These observations raise questions that cannot be answered because the training data, fine-tuning procedures, and underlying rules of both grammar-checking applications are not publicly documented. Consequently, it is only possible to hypothesise about the reasons for their behaviour. One possible explanation is that InstaText appears to rely primarily on AI-based methods, whereas Scribens appears to rely more heavily on rule-based error detection and correction.

Overall, although Scribens achieves relatively good error detection performance across the evaluated error types, neither application consistently provides the intended corrections. These findings suggest that both applications still have notable limitations in handling Bulgarian-specific grammar, spelling, and punctuation errors.

5 Evaluation of LLM performance in grammar checking

LLMs are used in grammar-checking applications because they learn language patterns from large-scale datasets, enabling them to detect and correct errors in a context-aware manner and potentially offering advantages over traditional rule-based systems.

5.1 Brief description of LLMs selected for the experiment

We present five LLMs selected to evaluate the performance of major proprietary models and open-weight models.

Gemma 4 31B IT is an open-weight multimodal LLM developed by Google DeepMind and released in 2026.Built on research related to the Gemini 3 series, it supports more than 140 languages (including Bulgarian), image inputs, and a 256K-token context window. The model is released under the Apache 2.0 licence and is publicly available on Hugging Face and Kaggle (Google DeepMind, 2026b).

BgGPT 3.0 27B is an instruction-tuned LLM developed by INSAIT and released in 2026.Based on the Gemma 3 architecture, it is specifically designed for Bulgarian-language understanding and generation, supporting Bulgarian and English, multimodal input, and a 131K-token context window. The model is freely available through downloadable weights and an API (INSAIT, 2026).

Claude Sonnet 4.6 is an instruction-tuned LLM developed by Anthropic and released in 2026.It is designed for reasoning, coding, and agentic workflows, supports a context window of up to one million tokens, and is available through the Anthropic API (Anthropic, 2026).

Gemini 3.5 Flash is a multimodal LLM developed by Google DeepMind and released in 2026.It is optimised for low-latency inference while maintaining strong reasoning and coding performance. The model supports text, image, audio, video, and PDF inputs, a one-million-token context window, and is available through the Gemini API, Vertex AI, Google AI Studio, and the Gemini application (Google DeepMind, 2026a).

GPT-5.5 is a general-purpose multimodal LLM developed by OpenAI and released in 2026.It is designed for reasoning, instruction following, coding, and agentic tasks. GPT-5.5 supports a 1.05 million-token context window and incorporates improvements in reasoning, coding, and tool use over previous GPT-5 models. The model is available through the OpenAI API (OpenAI, 2026).

Experiments conducted with different model families, and with different releases within the same model series, demonstrate a high degree of variability in performance. This suggests that solutions to specific tasks are influenced by multiple factors, many of which are not fully transparent or controllable. Consequently, our analysis focuses on identifying broader performance tendencies and patterns rather than drawing conclusions from the results of any single model or model release.

5.2 Evaluation, results and analysis

The BulMMLU:gr dataset is designed as a single prompt covering all six error types. The prompt asks the model to provide the correct answer to a given question by choosing from four options. No example is shown. The model is instructed to answer only with one of the letters designating each possible answer: A), B), C) or D). Three of the tested models produced outputs exactly as instructed, while two returned the letter indicating the correct answer together with an explanation, from which the selected option was extracted.

We evaluate model performance using accuracy, comparing each model prediction to the gold-standard answer and scoring it as correct or incorrect depending on whether it matches. The evaluation metric is binary accuracy, calculated as the proportion of instances in which the model's decision matches the gold standard out of the total number of answers. The results for five models, distributed among the six error types, are presented in Table 3.

Table 3: Accuracy of the five evaluated models

Type	**N**	**M1**	**M2**	**M3**	**M4**	**M5**
Type 1	20	0.80	0.25	0.80	**0.90**	0.85
Type 2	21	0.48	0.43	**0.90**	0.48	0.10
Type 3	21	0.10	0.24	**0.71**	0.19	0.19
Type 4	16	0.63	0.31	**0.75**	0.56	0.50
Type 5	20	0.15	0.25	**0.85**	0.20	0.30
Type 6	20	0.45	0.25	**0.55**	0.40	0.15
Total	118	0.42	0.29	**0.76**	0.45	0.34

The accuracy results reveal substantial performance differences across models and question types. Among the evaluated models, M3 achieves the highest overall accuracy (0.76), outperforming all others by a wide margin. The remaining models show markedly lower overall performance, with M4 achieving 0.45, M1 – 0.42, M5 – 0.34, and M2 – 0.29, and they exhibit more pronounced weaknesses for specific error types.

The results also show that model performance strongly depends on the type of question being addressed. No single model achieves the highest accuracy across all question types. For Type 1 questions, M4 performs best with an accuracy of 0.90, followed by M5 (0.85) and M1/M3 (0.80). In contrast, M4 performs substantially worse on Type 3 and Type 5 questions, indicating that its strengths may be limited to particular reasoning patterns or task formulations. M5 shows a similar pattern of narrow strength: it is the second-best model on Type 1 (0.85), but its accuracy drops sharply on Type 2 (0.10) and remains low for Types 3, 5, and 6, indicating that its performance is highly dependent on question type. The models are not specifically pre-trained or fine-tuned for Bulgarian grammar, spelling, or punctuation, and the strong performance of a particular model on a specific task may be due to several factors, such as training data, model architecture, Bulgarian coverage, and so on.

M3 shows a comparatively more consistent performance pattern across the evaluated categories. It achieves the best results for five of the six question types (Types 2–6), with Type 1 being the only category where another model (M4) performs better. M3's strongest results are on Type 5 (0.85) and Type 2 (0.90). M1, by contrast, is highly inconsistent: it performs reasonably well on Type 1 (0.80, on par with M3) but drops sharply on Type 3 (0.10), and remains weak on Type 5 (0.15) and Type 6 (0.45). This indicates that M1 lacks general robustness across error types. As a model with open weights, it nonetheless remains a good candidate for further fine-tuning to detect errors in Bulgarian grammar, spelling, and punctuation.

The lowest overall performance is observed for M2, which remains below 0.50 accuracy for all question types and achieves only 0.29 overall accuracy. This consistent underperformance suggests that the model may have difficulties with the specific reasoning requirements represented in the evaluation set. In comparison, M4 and M5 show intermediate performance but exhibit high variability between categories, indicating sensitivity to task formulation.

The observed variability confirms that model selection has a significant impact on task outcomes. Differences between models are not limited to average accuracy but also appear in their relative strengths across question types. Therefore, evaluating only aggregate performance may conceal important model-specific capabilities and limitations. These findings support the use of multiple evaluation categories and suggest that model choice should take into account the characteristics of the target task rather than relying solely on general performance rankings.

The McNemar test is a statistical test used to compare two classifiers on the same set of instances. Unlike tests based on independent samples, it evaluates paired predictions by focusing on cases where the two models disagree: instances correctly classified by one model but incorrectly classified by the other (McNemar, 1947). The test assesses whether the difference in error distributions between two models is statistically significant. In this work, we use the exact binomial form of the test rather than the chi-squared approximation, computing the p-value directly from the binomial distribution with $n = b + c$ trials and success probability 0.5. A low p-value indicates that the observed performance difference is unlikely to arise from random variation,

suggesting that one model achieves a statistically different level of accuracy compared with the other.

The pairwise McNemar test results for the complete set of questions (Table 4) show statistically significant differences between most model pairs. M3 exhibits the most distinctive performance profile, with statistically significant differences compared with every other model ($p < 0.0001$ in all comparisons). This confirms that the superior overall accuracy of M3 is not only a numerical improvement but also represents a statistically significant change in prediction behaviour compared with the remaining models.

The comparisons among M1, M2, M4, and M5 reveal a more mixed pattern. M1 differs significantly from M2 ($p = 0.0328$), but no significant difference is observed between M1 and M4 or M5 ($p = 0.6476$ and $p = 0.0872$, respectively). Similarly, M2 performs significantly differently from M4 ($p = 0.0094$), but its difference from M5 is not statistically significant ($p = 0.4709$). These results indicate that although the models have different average accuracies, some performance differences may not be sufficiently large or consistent across individual questions to reach statistical significance.

The results by question type (Table 5) provide additional insight into the source of these differences. M3 shows statistically significant differences compared with the other models across several question categories, particularly Types 2, 3, and 5. For Type 2 questions, M3 shows statistically significant differences in favour of M3 over all competing models ($p \leq 0.0063$), indicating a strong advantage in this category. Similarly, Type 3 and Type 5 questions show significant differences between M3 and all other models, suggesting that these categories contribute substantially to the overall performance gap.

Given the number of pairwise comparisons performed, results should be interpreted with awareness that no correction for multiple testing (e.g., Bonferroni) was applied; borderline p-values (e.g., 0.0328, 0.0391, 0.0386) would not remain significant under such correction.

Overall, the McNemar analysis confirms that M3 provides a statistically significant improvement over the other evaluated models for the majority of the benchmark. However, the category-level analysis shows that model superiority is task-dependent: some question types do not show significant differences, indicating that performance advantages arise primarily from specific capabilities rather than universal dominance across all tasks.

6 Conclusion and future work

Of the applications surveyed, only two supported Bulgarian and were included in the evaluation, specifically for tasks involving grammatical error detection and correction suggestions. Although large language models demonstrate strong capabilities in improving text fluency and correcting various types of errors, the evaluated LLM interfaces did not offer the same explicit correction-review workflow as grammar-checking applications, which allow users to review, select, and accept or reject individual corrections. This limitation highlights a potential direction for future research and development: creating a dedicated Bulgarian grammar-checking system that combines automatic error detection with user-controlled correction suggestions.

A key objective for future work is to develop a systematic description and classification of the most frequent errors in Bulgarian, covering syntax, punctuation, and spelling. This resource could support the extension of the BulMMLU:gr dataset by incorporating examples of incorrect sentences paired with the corresponding corrected versions. Further experiments will determine whether a synthetic dataset, symbolic approaches to data processing, or both are needed to achieve sufficient coverage of the different error types.

Future investigations may also explore the adaptation and fine-tuning of open-weight large language models for Bulgarian grammatical error detection and correction tasks. Potential candidate models include Gemma 3 27B and Gemma 4 27B, which could serve as foundations for developing specialised models capable of providing accurate, transparent, and user-controllable grammar correction assistance for Bulgarian. At present, the possibilities of quantising and running large open-weight language models (such as DeepSeek-V3, V4 Pro, or Kimi 3), or using pre-compressed or pruned community weights, remain largely unexplored.

7 Acknowledgments

The present study is carried out within the project Infrastructure for Fine-tuning Pretrained Large Language Models, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Anthropic. 2026. Claude Sonnet 4.6: Frontier performance for scale. Accessed: 15 July 2026.

Maksym Bondarenko, Artem Yushko, Andrii Shportko, and Andrii Fedorych. 2023. Comparative Study of Models Trained on Synthetic Data for Ukrainian Grammatical Error Correction. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 103–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Google DeepMind. 2026a. Gemini 3.5 Flash Model Card. Google AI.

Google DeepMind. 2026b. Gemma 4 Model Card. Accessed: 15 July 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, pages 1–27. OpenReview.net.

Svanhvít Lilja Ingólfsdóttir, Petur Ragnarsson, Haukur Jónsson, Haukur Simonarson, Vilhjalmur Thorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-Level Grammatical Error Correction Using Synthetic and Curated Corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

INSAIT. 2026. BgGPT 3.0: A new generation of Bulgarian AI models. Accessed: 15 July 2026.

Atakan Kara, Farrin Marouf Sofian, Andrew Bond, and Gözde Şahin. 2023. GECTurk: Grammatical Error Correction and Detection Dataset for Turkish. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 278–290, Nusa Dua, Bali. Association for Computational Linguistics.

Bozhidar Klouchek and Riza Batista-Navarro. 2024. Bulgarian Grammar Error Correction with Data Augmentation and Machine Translation Techniques. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 365–376, Trento. Association for Computational Linguistics.

Svetla Koeva, Ivelina Stoyanova, Dimiter Georgiev, Svetlozara Leseva, Valentina Stefanova, Maria Todorova, Tsvetana Ivanova Dimitrova, Hristina Kukova, Mihaela Moskova, and Tinko Tinchev. 2026. Bulgarian Massive Multitask Language Understanding Benchmark. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4658–4672, Palma de Mallorca, Spain. European Language Resources Association (ELRA).

Roman Kovalchuk, Mariana Romanyshyn, and Petro Ivaniuk. 2025. Introducing OmniGEC: A Silver Multilingual Dataset for Grammatical Error Correction. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 162–178, Vienna, Austria (online). Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12:153–157.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA – Online. Association for Computational Linguistics.

OpenAI. 2026. Introducing GPT-5.5. OpenAI Technical Report and API Documentation. Accessed: 15 July 2026.

Frank Palma Gomez, Alla Rozovskaya, and Dan Roth. 2023. A Low-Resource Approach to the Grammatical Error Correction of Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 114–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Ujjwal Sharma and Pushpak Bhattacharyya. 2025. IndiGEC: Multilingual Grammar Error Correction for Low-Resource Indian Languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22382–22396, Suzhou, China. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic Data Generation for Grammatical Error Correction

with Tagged Corruption Models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Wei Tian and Yuhao Zhou. 2025. Chinese Error Corrector 3-4B: State-of-the-Art Chinese Spelling and Grammar Corrector. *arXiv:2511.17562*.

Mircea Timpuriu, Mihaela-Claudia Cercel, and Dumitru-Clementin Cercel. 2026. RoLegalGEC: Legal Domain Grammatical Error Detection and Correction Dataset for Romanian. *arXiv:2604.19593*.

Appendix A McNemar Test Results

Model	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*	—

* Statistically significant difference at a significance level of $p < 0.05$.

Table 4: p-values from the McNemar test for pairwise comparisons between the models (total of 118 questions)

Question type	vs M1	vs M2	vs M4	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*

*Statistically significant difference compared to Model 3 at p < 0.05.

Table 5: p-values from the McNemar test: Comparison of Model 3 with the other models across question types