Serbian SuperGLUE: Towards an Evaluation Benchmark for South Slavic Language Models

Mitar Perović[1] **Teodora Mihajlov**[1]
Recrewty
{mitar, teodora.mihajlov}@recrewty.com

Abstract

We introduce Serbian SuperGLUE, a comprehensive benchmark for evaluating natural language understanding in Serbian, adapted from the English SuperGLUE benchmark. The benchmark comprises seven tasks spanning question answering, natural language inference, and coreference resolution, created through a combination of LLM-based translation with automatic post-editing and native data generation. We evaluate seven encoder-based language models, including both Serbian-specific (BERTić, Jerteh) and multilingual models (mmBERT, XLM-RoBERTa variants). Our results reveal that multilingual models remain competitive with language-specific alternatives, with mmBERT achieving the best performance on RTE (75.7%) and XLM-R-BERTić leading on BoolQ (82.0%). We observe significant training variance on smaller datasets, with standard deviations exceeding 10% in some configurations, highlighting the importance of multi-seed evaluation for low-resource benchmarking. We release the benchmark, evaluation code, and model checkpoints to facilitate reproducible research on South Slavic language understanding.[1]

1 Introduction

Recent proliferation of language models brought about rapid development on different language understanding tasks. It is no surprise that most of this has happened with the English language in focus, due to its wide use, but also the availability of data, and easy access to leaderboards and benchmarks. Unfortunately, the similar coverage for smaller languages is lacking. In this work, we focus on the Serbian language, and more broadly on South-Slavic languages, to develop datasets and tools to facilitate evaluation for complex language understanding tasks.

[1]Resources available at: [URL]

Serbian is spoken by approximately 12 million people and, together with Croatian, Bosnian, Montenegrin, Macedonian, and Slovene, forms the South Slavic language group. Scarcity of resources (Marovac et al., 2023) and substantial morphological complexity (Pakoci et al., 2019) pose significant challenges for developing and evaluating language models for these languages. To date, BENCHić is the only benchmark created specifically for the South Slavic languages, focusing on relatively simple language understanding tasks such as named entity recognition, sentiment analysis, Choice of Plausible Alternatives, and discrimination between closely related languages. Within BENCHić, Serbian and Croatian are included across all tasks, while Bosnian, Macedonian, and Montenegrin are present only in some, due to a lack of data (Rupnik et al., 2023).

To date, no benchmark exists for evaluating complex NLU tasks for BCMS languages. Moreover, resource development within the region remains fragmented, with limited collaboration and few shared infrastructures, hindering the establishment of unified evaluation frameworks and the comparability of model performance.

To address these gaps, we introduce Serbian SuperGLUE - a SuperGLUE-based (Wang et al., 2019) benchmark for higher-level language understanding in Serbian. We begin with Serbian due to its comparatively better resource availability and the relative ease of future localization to other BCMS languages. Our motivation for expanding this benchmark lies in the fact that higher-level tasks are not covered by existing benchmarks (apart from SuperGLUE for Slovenian), while smaller languages in the group, such as Bosnian, Montenegrin, and Macedonian, generally lack systematic coverage. At this initial stage, the development of data for all South Slavic languages was constrained by both limited data availability and the absence of established collaborations with domain experts from

other countries. Nevertheless, we see the development of the Serbian SuperGLUE-style benchmark as a catalyst for broader collaboration, enabling the formation of an expert network and providing a shared framework for the future development of language-specific resources.

The initial release consists of automatically translated and post-edited versions of the English SuperGLUE tasks, with additional manual verification for diagnostic subsets. Our contributions are twofold: (1) we provide the first SuperGLUE-style benchmark tailored to encoder-based models for Serbian, designed to extend to other BCMS varieties; and (2) we encourage greater resource integration and reuse by offering a unified, openly available evaluation framework for the BCMS NLP community.

The remainder of the paper is structured as follows: Section 2 reviews recent language understanding benchmarks (2.1) and summarizes available resources for South Slavic languages (2.2). Section 3 describes the construction of the Serbian SuperGLUE benchmark, including the translation pipeline (3.1), the evaluation datasets (3.2), and the diagnostic datasets (3.3). In Section 4, we present evaluation results for the most relevant models for the Serbian language, then proceed to discussion (5). Finally, we end with conclusions and plans for future work (6, and address the limitations of the presented approach (7).

2 Related Work

2.1 Benchmarking language models

Building evaluation datasets and benchmarks for language models is often a demanding and time-consuming process. Nevertheless, well-designed benchmarks drive model improvement and facilitate informed model selection for different applications. One of the first initiatives for a unified evaluation of language models was the General Language Understanding (GLUE) benchmark (Wang et al., 2018). It covers sentiment analysis, paraphrase detection, linguistic acceptability, and natural language inference (NLI) tasks. In addition, the benchmark includes two manually constructed, expert-annotated diagnostic datasets for probing specific linguistic features present in natural language.

However, with the rapid development of language models such as ELMo (Peters et al., 2018), OpenAI's GPT (Radford et al., 2018), and BERT (Devlin et al., 2019), model performance on GLUE quickly reached saturation. Consequently, SuperGLUE benchmark was introduced (Wang et al., 2019). It builds upon the design principles of GLUE but increases task complexity and extends beyond the sentence-pair classification format to include question answering and coreference resolution. SuperGLUE retains two most difficult tasks from its predecessor, the Winograd Schema Challenge (WNLI) and Recognizing Textual Entailment (RTE). Additionally, SuperGLUE includes human baselines across all tasks.

A key limitation of notable benchmarks is their focus on English, which prevents their use for multilingual and monolingual evaluation in other languages. To address this issue, several multilingual evaluation suites have been developed, such as MMLU-ProX (Xuan et al., 2025), XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020), which offer cross-lingual transfer tasks in a wide range of languages. However, of these resources, only MMLU-ProX includes Serbian, leaving the language underrepresented in standardized evaluation settings.

To further extend coverage for other languages, a growing body of work has focused on translating and adapting English-based benchmarks. Examples inspired by GLUE include BelarusianGLUE (Aparovich et al., 2025), BasqueGLUE (Urbizu et al., 2022), Russian SuperGLUE (Shavrina et al., 2020), SwedishGLUE (Adesam et al., 2020), and others (e.g., Tran et al., 2024; Osório et al., 2024). Additionally, some projects have introduced entirely new benchmark suites tailored to specific language and cultural contexts, namely LiRo for Romanian (Dumitrescu et al., 2021), HuLU for Hungarian (Ligeti-Nagy et al., 2024) and CLUE for Chinese (Xu et al., 2023).

2.2 Resources for South Slavic languages

In our work, we focus on the South Slavic language group. Within this region, Slovenian currently leads in benchmark development, with Slovene SuperGLUE (Žagar and Robnik-Šikonja, 2022), SloBench (DRAGAR, 2022) and BENCHić (Rupnik et al., 2023), the latter covering Bosnian, Croatian, Montenegrin, and Serbian as well. Resources for BCMS languages are developed largely through regional initiatives and organizations. The most frequent are plain-text corpora, but task-specific resources are also available.

Most existing plain-text corpora are derived from

the web, with only a few exceptions being literary or mixed-origin (Škorić and Janković, 2024; Ćavar and Brozović Rončević, 2012). Large-scale web corpora include srWac (Serbian), meWac (Montenegrin), hrWac (Croatian) and bsWac (Bosnian), produced through web crawls of top-level domains (Ljubešić and Klubička, 2014; Ljubešić and Erjavec, 2021), and their subsequent extensions (Ljubešić and Kuzman, 2024). Another valuable resource is the OSCAR 23.01 common-crawl corpus (Ortiz Su'arez et al., 2019), available for Bosnian, Croatian, Macedonian, and Serbian. The only sizable literary corpus available online is SrpELTeC, which contains historical Serbian texts (Stanković et al., 2022).

A notable example of parallel corpora that includes BCMS languages is the MaCoCu dataset which is available for Serbian, Croatian, and Bosnian (Bañón et al., 2022). It available in an aggregated form for the three languages format (Kuzman et al., 2023), combined with other datasets (Ljubešić et al., 2024a), and as a deduplicated as version (Ljubešić, 2021).

In addition to plain-text corpora, a range of task-specific annotated resources exists for Serbian. Most resources target Universal POS tagging, named entity recognition (NER), and sentiment analysis. Representative datasets include SrpKor4Tagging for POS tagging (Stankovic et al., 2020), SETimes_sr for combined POS and NER (Batanović et al., 2023; Samardžić et al., 2017), and WikiAnn-sr for NER (Rahimi et al., 2019). Comparable datasets have been developed for Croatian, Bosnian, and Serbo-Croatian, with *hr500k* (Ljubešić et al., 2016) being a notable Croatian resource covering both NER and POS annotations. For sentiment analysis, ParlaSent is a multilingual corpus of parliamentary speeches, manually annotated for sentiment, available for Bosnian, Croatian, Serbian, and four additional languages (Mochtak et al., 2024). Despite this coverage, annotated corpora remain limited in scale relative to the extensive plain-text collections (Škorić and Janković, 2024).

Support for higher-level semantic tasks is less developed. Beyond tagging, only a handful of corpora provide material for reasoning or question answering. Some examples include synthetic datasets *m_mmlu-sr*[2] and *m_hellaswag*[3], and a combination of synthetic data and natural language - *serbian_qa*.[4] Recent efforts have also focused on alignment and instruction-tuning for LLMs (e.g., *airoboros-3.0-serbian*[5]; *open-orca-slim-sr*[6], etc.) (Škorić and Janković, 2024). These emerging datasets mark an important shift toward a more complex, task-oriented evaluation for BCMS languages.

3 Serbian SuperGLUE

To pave the way for the evaluation of language models for more complex NLU tasks for BSCM languages, we translate the SuperGLUE benchmark from English to Serbian. We opt for SuperGLUE due to a higher task complexity, and some of the easier tasks, like sentiment analysis, and Named Entity Recognition (NER) being covered as a part of BENCHić (Rupnik et al., 2023).

In this paper, we present an initial version of the benchmark. We automatically translate and post-edit most of the datasets within the benchmark, with the exception of COPA, ReCoRD, and WiC datasets. COPA was readily available in Serbian as a part of BENCHić, while the other two could not be automatically translated, due to morphological and semantic differences between English and Serbian. We build ReCoRD adapting the existing resources. WiC is omitted from our Serbian benchmark because it requires models to assess whether a target word carries the same meaning in two sentences, and no publicly available Serbian lexicon currently exists.

The tasks covered in the benchmark are question answering (QA), natural language inference (NLI), and coreference resolution. In the next sections, we present the process of automatic translation and automatic post-editing (APE) of the SuperGLUE benchmark, including a brief translation analysis, and describe each dataset within the benchmark.

3.1 Dataset translation

For dataset translation, we selected Gemini Flash 2.5 following a comparative evaluation of five large language models (LLMs). Four annotators scored the generated translations from each model on a scale from 1 to 5. GPT-4o-mini and Gemini Flash

[2] https://huggingface.co/datasets/alexandrainst/m_mmlu

[3] https://huggingface.co/datasets/alexandrainst/m_hellaswag

[4] https://huggingface.co/datasets/smartcat/serbian_qa

[5] https://huggingface.co/datasets/draganjovanovich/airoboros-3.0-serbian

[6] https://huggingface.co/datasets/datatab/open-orca-slim-serbian

Original Translation	APE
question_srb: Da li *je* Himalaj najviši planinski venac na svetu?	**question_srb:** Da li *su Himalaji* najviši planinski venac na svetu?
passage_srb: Taken (TV serija) – Taken je kriminalistička dramska serija zasnovana na istoimenoj filmskoj trilogiji. (...) **question_srb:** Serija *"Is-Taken"* je zasnovana na istoimenom filmu.	**paragraph_srb:** Taken (TV serija) — Taken je kriminalistička dramska serija zasnovana na istoimenoj filmskoj trilogiji. (...) **question_srb:** *Da li je serija "Taken"* zasnovana na istoimenom filmu?
sentence1_srb: *Iz brze pretrage na Guglu*, Bitcoin Cash je nastao kao tvrdaviljuška Bitcoina i navodno je brži i održiviji.	sentencel_corr: *Prema brzoj pretrazi na Google-u*, Bitcoin Cash je nastao kao hard fork Bitcoina i navodno je brži i održiviji.

Table 1: Comparison of original translations and Automated Post-Editing (APE) results. Some items shown as excerpts (...). Bold text presents dataset rows. Text in italic is part of the corrected part of the text.

2.5 achieved the highest average ratings (4.321 and 4.250, respectively). Despite GPT-4o-mini's slightly higher score, we opt for Gemini due to its greater output consistency and GPT's occasional production of non-existent Serbian word forms, making Gemini's translations more reliable for large-scale automatic processing. An example of model translation comparison and scoring is presented in Table 2.

After translating all instances with Gemni, we carry out an in-depth manual evaluation of randomly sampled translations. The analysis indicates that, in spite preserving the original meaning and outputting all content in Serbian, the model exhibits several systematic errors. First, the model frequently mishandles named entities, despite explicit instructions. Notably, it often fails to transcribe personal names into Serbian, which is possible in all instances due to phonology of Serbian. Also, in paired inputs (e.g., passage–question pairs), named entities are sometimes translated in one segment but left in English in the other, likely due to an initial methodological issue in which paired components were processed separately. Second, the model frequently assigns incorrect grammatical gender to agent nouns, defaulting to masculine forms. Additional translationese phenomena are common, including unnatural phrasing, English-like syntactic structures, and incorrect or fabricated terminology in domain-specific contexts. Finally, although instructed to use Latin script, the model occasionally outputs Cyrillic text. Representative examples of each error type are provided in Table 6 in Appendix A.

Although the automatically translated data was largely fluent, we opt for improved, mainly due to grammatical agreement errors, which could affect text comprehension. Fully manual post-editing of the translations would be time-consuming and costly. To address the detected issues in a scalable manner, we introduce Automatic Post-Editing (APE) as a refinement step.

As with the translation, we evaluated multiple candidate models for APE. Based on this evaluation, we selected GPT-5-mini for the post-editing stage. To avoid the model's tendency to revert translations back to English when provided both the English source and the Serbian text, we restricted the input to the Serbian translation alone.

The APE step primarily resolved grammatical and agreement errors while enhancing terminology and named entity consistency through joint processing of paired items. Notably, despite lacking English source access, the model occasionally improved semantic alignment in the Serbian output. Conversely, APE introduced specific new errors: reverting transliterated names to English, disrupting gender agreement for named individuals, and omitting appositions. Table 1 provides comparative evidence: (1) correcting the auxiliary *je* to the plural *su* for *Himalaji*; (2) resolving a parsing error where a question was misinterpreted as a headline; and (3) an instance of stylistic over-correction where the semantic change was unnecessary.

To quantitatively evaluate the quality of the APE results, three annotators evaluated 100 translated examples on a 1–5 scale using criteria based on prior error analysis. The annotation instructions are available in A.1.

Annotator 1 gave the lowest and most variable scores (Mo = 4; 3.83 ± 0.89). In Annotator 1 first misinterpreted dataset column names, lowering scores, which required re-annotation. Annotator 2 higher but similarly variable scores (Mo = 5; 3.44 ± 1.21), and Annotator 3 the highest and most consistent scores (Mo = 5; 4.66 ± 0.74). Fleiss' Kappa ($\kappa = 0.1226$) indicates low inter-annotator agreement. Inter-annotator agreement analysis reveals that Annotators 1 and 2—both trained linguists—achieved a higher level of con-

English	Serbian translation	Model	Avg. score
The June Movement has always been opposed to giving EU cooperation a military dimension.	Junski pokret je oduvek bio protiv davanja vojnoj dimenziji saradnji u EU.	claude-sonnet-4	1.75
	Junski pokret se uvek protivio davanju vojne dimenzije saradnji EU.	gemini-2.5-flash	**4.25**
	Junijski pokret je oduvek bio protiv davanja vojne dimenzije saradnji u okviru EU.	gpt-4o	3.25
	Junski pokret se oduvek protivio davanju vojnog aspekta saradnji sa EU.	gpt-4o-mini	**4.25**
	Junijaški pokret je oduvek bio protiv toga da se saradnji u EU daje vojna dimenzija.	gpt-5	3.75
	Junski pokret je uvek bio protiv davanja vojne dimenzije saradnji sa EU.	google-translate	4.20

Table 2: Comparison of Serbian translations for a sample sentence across models. The *Avg. score* column reports the average rating assigned by the four annotators, with best scores presented in **bold**.

sensus ($\kappa = 0.2282$) compared to their agreement with Annotator 3 (ranging from $\kappa = 0.0485$ to $\kappa = 0.0829$). While these values reflect the inherent subjectivity of the task, the relative consistency between the experts suggests a more stable quality signal. Consequently, our quality assessment primarily relies on the consensus between Annotators 1 and 2. We acknowledge these figures as a baseline and plan to conduct a comprehensive re-annotation phase using a larger pool of linguistic experts and more granular annotation guidelines.

Translation quality varies across datasets. AX_g achieves the highest mean score (4.80 ± 0.25), indicating consistently high-quality translations, likely due to its simpler structure. AX_b also scores highly (4.20 ± 0.72), though with moderate variability, suggesting that while most translations are good, some examples are rated lower. RTE shows high average quality (4.14 ± 0.47) but with noticeable inconsistency. BoolQ (3.88 ± 0.48), CB (3.48 ± 0.58), and MultiRC (3.74 ± 0.56) show moderate quality, with CB and MultiRC displaying more variable scores. WSC has the lowest mean (3.52 ± 0.77) and high variability, reflecting generally poor translations despite being fully manually corrected. Scores may have been affected by the dataset's complex structure, highlighting the need to improve annotation guidelines in future iterations. Overall, translation quality depends on both dataset complexity and the consistency of automatic post-editing, which can influence downstream evaluation.

3.2 Evaluation datasets

Serbian SuperGLUE comprises six datasets for model evaluation, and two diagnostic datasets for probing knowledge of specific lingustic features in models. The evaluation datasets cover three tasks: question answering (QA), natural language inference (NLI), and coreference resolution. Examples of all evaluation datasets are presented in Table 5. Unless stated otherwise, dataset was automatically translated and post-edited using the aforementioned procedure. Tasks, corpus sizes per split, and evaluation metrics for each task are presented in Table 3.

BoolQ (Boolean Questions, Clark et al., 2019) is a question answering (QA) task consisting of a paragraph and a yes/no question related to it. The questions are sourced from Google searches, while the paragraphs are excerpts from Wikipedia articles containing the answer.

COPA (Choice of Plausible Alternatives, Roemmele et al., 2011) is a causal reasoning task. Each example consists of a premise, a question asking if something is a cause or if it is an effect of the premise. The dataset was already translated into Serbian as a part of the BENCHić benchmark (Rupnik et al., 2023), using the XCOPA methodology (Ponti et al., 2020) . The Serbian version, COPA-sr, contains 1,000 examples, divided into training (400 instances), validation (100 instances), and test (500 instances) splits (Ljubešić et al., 2022).

MultiRC (Multi-Sentence Reading Comprehension, Khashabi et al., 2018) is a multiple choice question-answering (MCQs) task with multiple correct answers. Each example is a paragraph, a question, and a list of possible answers. To find the

Corpus	Train	Dev	Test	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1a/EM	various
ReCoRD*	5607	1869	1869	QA	F1/EM	Wikipedia, news, scientific articles
RTE	2490	277	300	NLI	acc.	news, Wikipedia
WSC	554	104	146	coref.	acc.	fiction books

Table 3: The tasks included in the Serbian SuperGLUE. *QA* stands for Question Answering, *NLI* stands for Natural Language Inference, and *coref.* for Coreference Resolution. The table displays sizes per split, as well as whether a corpus was translated from English. For corpora translated from English, text sources are derived from Wang et al. (2019). Datasets not translated from english are marked with an asterix (*).

correct answer(s) the model has to extract information from several paragraph sentences.

ReCoRD (Reading Comprehension with Commonsense Reasoning; Zhang et al., 2018) is a multiple-choice question (MCQ) task derived from online English news. Each example consists of a passage and a cloze-style (fill-in-the-blank) question in which named entities are masked. Direct translation of the dataset into Serbian is challenging due to difficulties in aligning noun forms between passages and queries. To address this, we first derive passages from the *smartcat/serbian_qa* dataset,[7] which contains natural Serbian text from Wikipedia, news, and scientific articles. We then extract entities from these passages and provide passage–entity tuples to GPT-4o-mini to generate corresponding cloze questions. The entities in the generated questions are subsequently masked, and checks are performed to ensure that all masked entities appear in the passage.

CB (CommitmentBank, De Marneffe et al., 2019) is a corpus of short texts in which at least one sentence contains an embedded clause. Each of these embedded clauses is annotated with the degree to which it appears the person who wrote the text is committed to the truth of the clause. Each example consists of a premise containing an embedded clause and the corresponding hypothesis is the extraction of that clause.

RTE (Recognizing Textual Entailment) is a natural language inference (NLI) task formulated as a binary classification. The dataset examples are derived from Wikipedia and news sources. RTE used in Xu et al. (2023) is a collection of datasets that emerged as a part of a shared task. We translate and post-edit the same versions of the datasets.

[7] https://huggingface.co/datasets/smartcat/serbian_qa

WSC (Winograd Schema Challenge, Levesque et al., 2012) is a coreference resolution dataset that requires commonsense reasoning. Each example contains a sentence with a pronoun and one or more candidate noun phrases, and the task is to determine whether the pronoun refers to the given noun. The WSC dataset was automatically translated to Serbian but required extensive manual correction. Coreference in Serbian is often resolved differently than in English, primarily through verb inflections, which carry information on tense, grammatical number, and grammatical gender. In addition, it can be resolved through relative pronouns and relative clauses. The English dataset does not cover these cases. Thus, we manually correct to include both pronoun, and verb resolutions. In the future, we plan to expand the dataset with cases covering relative pronouns and clauses.

3.3 Diagnostic datasets

In addition to the primary evaluation tasks, SuperGLUE includes two diagnostic datasets designed for error analysis, qualitative comparison, and assessment of model-level world knowledge, linguistic competence, and gender bias. AX_b provides broad-coverage diagnostics and is evaluated using Matthews correlation coefficient (MCC), while AX_g focuses on gender-bias detection in NLI, with performance measured by accuracy and the Gender Parity Score (GPS).

In our benchmark, both diagnostic sets were automatically translated and refined with APE. The AX_g dataset required manual correction because agent nouns were predominantly assigned masculine gender, which would have invalidated the gender-bias evaluation for Serbian. We corrected the examples so that agent nouns appear both masculine and feminine form to mirror the structure of

the English dataset. Finally, the diagnostic datasets were manually annotated by three annotators, two master's students in linguistics and one engineer, to establish human baseline performance. Average MCC for the AX_b dataset is 0.557, while average accuracy on AX_g is 0.858. In this stage, the models were not evaluated on the diagnostic datasets, thus, the human baseline will be used in future work for comparison with model performance.

4 Results

We evaluate seven encoder-based language models on Serbian SuperGLUE: BERTić (Ljubešić and Lauc, 2021), CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020), Jerteh (Škorić, 2024a), mmBERT (Marone et al., 2025), mmBERT-small (Marone et al., 2025), XLM-R-BERTić (Ljubešić et al., 2024b), and TeslaXLM (Škorić, 2024b). All models are fine-tuned on each task using five random seeds (42-46), and we report mean accuracy with standard deviation. Training hyperparameters are consistent across models: learning rate of 1×10^{-5}, batch size of 16, up to 30 epochs with early stopping (patience of 5), warmup ratio of 0.1, and weight decay of 0.01. All experiments use FP16 mixed precision training on NVIDIA GPUs (RTX 4090). As the official SuperGLUE test labels are not publicly available, we report performance on our internal validation split. To ensure the long-term utility of the benchmark, the final test set will be released with locked labels obtained through coordination with the original SuperGLUE authors. This will allow for a standardized "blind" evaluation in future iterations of the benchmark.

4.1 Main Benchmark Results

Table 4 presents the main benchmark results across all evaluation tasks. We report accuracy for BoolQ, RTE, WSC; accuracy and macro-F1 for CB; $F1_a$ (answer-level F1) and EM (exact match) for MultiRC; and F1/EM for ReCoRD.

Question Answering Tasks. For BoolQ, XLM-R-BERTić achieves the highest accuracy (82.0%), followed by mmBERT (77.9%) and BERTić (77.1%). TeslaXLM shows notably high variance ($\sigma = 8.1\%$), with one seed achieving 80.7% while others hover around 62%, highlighting the sensitivity of larger models to random initialization on this task.

MultiRC proves challenging for all models, with BERTić achieving the best $F1_a$ of 65.8%. Notably, Jerteh, despite being the largest model (355M parameters), significantly underperforms with an $F1_a$ of only 18.5%, suggesting potential issues with multi-sentence reasoning in Serbian or overfitting on the relatively small training set.

ReCoRD shows uniformly high performance across all models (94.6-95.9% F1/EM), with minimal differentiation. This near-saturation suggests that the task, generated from Serbian Wikipedia and news sources, may be too straightforward for current encoder models, or that the entity extraction formulation does not fully capture the commonsense reasoning challenge of the original English dataset.

Natural Language Inference. On CB, Jerteh achieves the best accuracy (88.6%) and F1 (88.5%), outperforming other models by a substantial margin. This strong performance on three-way classification contrasts with its weak MultiRC results, suggesting that Jerteh excels at shorter inference tasks but struggles with longer contexts.

RTE presents a more varied picture. mmBERT leads with 75.7% accuracy, while TeslaXLM performs near random chance (53.5%). XLM-R-BERTić shows extremely high variance ($\sigma = 12.7\%$), with individual seeds ranging from 52.7% to 79.8%, indicating training instability on this small dataset (2,490 training examples).

Coreference Resolution. The WSC dataset exhibits a striking ceiling effect: all models converge to approximately 63-65% accuracy regardless of architecture or pretraining. mmBERT achieves exactly 63.5% across all five seeds ($\sigma = 0.0$), suggesting that models may be learning to predict the majority class rather than performing genuine coreference resolution. This pattern likely stems from the small training set size (554 examples) and class imbalance in the evaluation split.

Comparison with English SuperGLUE. To contextualize the Serbian benchmark results, we compare model performance against the best English results reported in the original SuperGLUE paper (Wang et al., 2019), presented together with our results in Table **??**. To maintain a controlled comparison, we reference the results from the original SuperGLUE publication; despite the availability of newer models online, the published paper offers greater clarity on the underlying training setups.

Notably, ReCoRD, the only dataset built from

Model	BoolQ	CB		COPA	RTE	MultiRC		ReCoRD	WSC
	acc.	acc.	F1	acc.	acc.	$F1_a$	acc.	F1	acc.
BERTić	77.1±1.1	75.4±2.3	57.1	63.4±2.5	70.7±2.1	**65.8**	**72.0**	95.6	**64.4**±1.2
CroSloEngual	71.3±0.5	80.4±2.8	80.1	57.4±3.6	63.4±3.4	51.6	64.3	94.6	63.7±0.8
Jerteh	73.0±0.7	**88.6**±3.0	**88.5**	**64.6**±2.3	59.9±1.5	18.5	58.9	94.6	64.2±0.4
mmBERT	77.9±0.5	83.9±2.5	79.1	64.0±6.0	**75.7**±1.7	48.4	66.2	95.5	63.5±0.0
mmBERT-small	72.9±1.4	76.4±2.3	68.3	59.8±3.1	59.4±5.0	46.8	64.5	94.7	63.1±0.9
TeslaXLM	66.2±8.1	87.9±3.4	82.1	59.8±2.8	53.5±1.8	64.6	69.8	**95.9**	63.7±0.4
XLM-R-BERTić	**82.0**±0.4	78.9±7.3	66.2	64.4±4.8	68.8±12.7	62.6	71.4	**95.9**	**64.4**±0.8
Best English	80.4	90.4	84.7	84.4	82.7	70.4	24.5	74.8	64.3

Table 4: Main benchmark results on Serbian SuperGLUE. Values represent mean performance across 5 seeds, with standard deviation in subscript where applicable. Best results per column are in **bold**. MultiRC reports F1 over answer-options ($F1_a$) and per-option accuracy (acc.). Note that WSC and COPA results exhibit ceiling effects around 63-65% due to limited training data (554 and 400 examples respectively). Best English row presents the best results on the English SuperGLUE benchmark as reported in Wang et al. (2019).

scratch in Serbian, significantly outperforms the English baseline. This suggests that "native-first" construction avoids "translationese" and artifacts, offering a more authentic measure of linguistic reasoning. While CB and RTE results are generally consistent with English, monolingual Serbian models underperform on RTE compared to mmBERT. As RTE source texts are English-news-centric, the gap likely stems from latent knowledge of English-specific named entities, highlighting a cultural bias in translated benchmarks. The lower scores on COPA likely reflect a data-sparse environment where the small training set is insufficient for encoder-only models to master complex causal reasoning. Conversely, WSC performance remains on par with English despite Serbian's morphological complexity. However, the manual adaptation of WSC omitted certain structures (e.g., clausal coreference), which may have inadvertently reduced task difficulty.

4.2 Model Comparison

Table 4 summarizes model performance across task categories.

Language-Specific vs. Multilingual Models. BERTić, a Serbian-specific BERT model, shows competitive performance on most tasks, particularly excelling on MultiRC ($F1_a$=65.8%) and BoolQ (77.1%). However, mmBERT, despite being trained on 104 languages, achieves the best RTE accuracy (75.7%) and competitive results elsewhere, suggesting that massive multilingual pretraining can compensate for lack of language-specific focus.

Model Scale. Larger models do not consistently outperform smaller ones. Jerteh (355M parameters) achieves state-of-the-art on CB but severely underperforms on MultiRC and RTE. XLM-R-BERTić (270M parameters) leads on BoolQ and ties for best on ReCoRD but shows high training variance on RTE and CB. These results suggest that model scale alone is insufficient; architecture and pretraining data composition play equally important roles.

Training Stability. We observe substantial variance across random seeds for several model-task combinations:

- TeslaXLM on BoolQ: $\sigma = 8.1\%$ (range: 62.2-80.7%)
- XLM-R-BERTić on RTE: $\sigma = 12.7\%$ (range: 52.7-79.8%)
- XLM-R-BERTić on CB: $\sigma = 7.3\%$ (range: 73.2-84.6%)

This variance is particularly pronounced on smaller datasets (CB: 250, RTE: 2,490 training examples) and highlights the importance of reporting confidence intervals rather than single-seed results.

5 Discussion

5.1 Task Difficulty Hierarchy

Our results reveal a clear hierarchy of task difficulty. ReCoRD emerges as the easiest (94-96% for all models), likely because our generation methodology produces more straightforward entity-matching questions than the original English dataset. BoolQ and CB show meaningful model differentiation with moderate difficulty. RTE and MultiRC prove more challenging, with substantial

performance gaps between models. WSC, despite its conceptual difficulty, fails to differentiate models due to dataset limitations rather than task complexity.

5.2 Cross-lingual Transfer

A key finding is that language-specific pretraining does not guarantee superior performance. mmBERT achieves the best RTE accuracy despite lacking Serbian-specific pretraining. We attribute this to: (1) larger pretraining corpora providing more diverse linguistic patterns; (2) beneficial transfer from related Slavic languages in the multilingual training data; and (3) universal NLI patterns that transcend language boundaries.

However, Serbian-specific models like BERTić show advantages on tasks requiring deeper linguistic understanding, such as MultiRC, where paragraph-level comprehension benefits from native language modeling.

6 Conclusion

We presented Serbian SuperGLUE, the first comprehensive benchmark for complex language understanding tasks in Serbian. Our benchmark includes seven tasks covering question answering, natural language inference, and coreference resolution, created through LLM-based translation and native data generation.

Evaluation of seven encoder models reveals several key findings: (1) multilingual models like mmBERT remain competitive with Serbian-specific alternatives; (2) model scale does not guarantee better performance-Jerteh (355M) excels on CB but fails on MultiRC; (3) small datasets exhibit high training variance, necessitating multi-seed evaluation; and (4) the WSC coreference task requires native data augmentation to enable meaningful model comparison.

We release Serbian SuperGLUE as an open resource to support the development of NLU systems for Serbian and, by extension, other South Slavic languages. Future work will expand the benchmark with native coreference data, instruction-tuned LLM evaluation, and diagnostic task completion.

7 Limitations

7.1 Dataset translation limitations

While translating established English evaluation frameworks such as SuperGLUE enables the reuse of well-defined tasks and closed test set labels, this approach has several limitations. First, translation, particularly automatic machine translation, can introduce lexical, syntactic, or semantic errors that may hinder task completion and affect the reliability of evaluation on the translated data. Additionally, the translation model itself may introduce biases, reflecting the data it was trained on and potentially amplifying or distorting patterns in the original dataset. Second, translating datasets from English may result in the loss of cultural context relevant to the target language community. Finally, not all linguistic phenomena that are important for the target language are necessarily present in English. As a result, translated benchmarks may fail to fully capture the linguistic challenges posed by the target language.

7.2 Reproducibility Concerns

The high variance on several tasks raises reproducibility concerns. When XLM-R-BERTić achieves 52.7% to 79.8% on RTE depending solely on random seed, single-seed evaluations become unreliable. We recommend:

- Reporting mean and standard deviation across $\geq$5 seeds
- Using validation F1 rather than loss for early stopping
- Employing lower learning rates (5×10^{-6}) for datasets with $<$1,000 examples
- Considering ensemble methods for production deployments

7.3 Limitations of Coreference Evaluation

The WSC ceiling effect (all models at $\sim$63-65%) renders this task uninformative for model comparison. Contributing factors include:

- **Insufficient training data**: 554 examples cannot support fine-tuning for commonsense reasoning
- **Label imbalance**: $\sim$65% positive labels make majority-class prediction a strong baseline
- **Translation artifacts**: Serbian coreference relies on verb inflection and grammatical gender, patterns not fully captured in translated data

We recommend treating WSC results with caution and prioritizing native Serbian coreference data collection in future work.

Acknowledgments

We would like to sincerely thank Daria Milošević, Nina Škoro, and Ivan Đukanović for their valuable contributions to data annotation. We also extend our thanks to Daria Milošević and Nina Škoro for the careful review and correction of the automatic translations of the diagnostic datasets.

References

Yvonne Adesam, Aleksandrs Berdicevskis, and Felix Morger. 2020. Swedishglue–towards a swedish test set for evaluating natural language understanding models.

Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. Belarusianglue: Towards a natural language understanding benchmark for belarusian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527.

Marta Bañón, Miquel Esplà-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik Van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, and 1 others. 2022. Macocu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *23rd Annual Conference of the European Association for Machine Translation, EAMT 2022*, pages 303–304. European Association for Machine Translation.

Vuk Batanović, Nikola Ljubešić, Tanja Samardžić, and Tomaž Erjavec. 2023. Serbian linguistic training corpus SETimes.SR 2.0. Slovenian language resource repository CLARIN.SI.

Damir Ćavar and Dunja Brozović Rončević. 2012. Riznica: the croatian language corpus. *Prace filologiczne*, 63:51–65.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 107–124. Data can be found at `https://github.com/mcdm/CommitmentBank/`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

FRENK DRAGAR. 2022. *SloBench: Slovenian Natural Language Processing Benchmark*. Ph.D. thesis, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, and 1 others. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2023. Get to know your parallel data: Performing English variety and genre classification over MaCoCu corpora. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103, Dubrovnik, Croatia. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. *KR*, 2012(13th):3.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and 1 others. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. 2024. Hulu: Hungarian language understanding benchmark kit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8360–8371.

Nikola Ljubešić. 2021. Text collection for training the BERTić transformer model BERTić-data. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Tomaž Erjavec. 2021. Montenegrin web corpus meWaC 1.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Nikola Ljubešić and Taja Kuzman. 2024. Classla-web: comparable web corpora of south slavic languages enriched with linguistic and genre annotation. *arXiv preprint arXiv:2403.12721*.

Nikola Ljubešić and Davor Lauc. 2021. BERTić – the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42. Association for Computational Linguistics.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022. Choice of plausible alternatives dataset in serbian COPA-SR. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Vít Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024a. Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 189–203, Torino, Italia. ELRA and ICCL.

Nikola Ljubešić, Vít Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024b. Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 189–203, Torino, Italia. ELRA and ICCL.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Ulfeta A Marovac, Aldina R Avdić, and Nikola Lj Milošević. 2023. A survey of resources and methods for natural language processing of serbian language. *arXiv preprint arXiv:2304.05468*.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16024–16036, Torino, Italia. ELRA and ICCL.

Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Tomás Freitas Osório, Bernardo Leite, Henrique Lopes Cardoso, Luís Gomes, João Rodrigues, Rodrigo Santos, and António Branco. 2024. Portulan extraglue datasets and models: Kick-starting a benchmark for the neural processing of portuguese. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pages 24–34.

Edvin Pakoci, Branislav Popović, and Darko Pekar. 2019. Using morphological data in language modeling for serbian large vocabulary speech recognition. *Computational intelligence and neuroscience*, 2019(1):5072918.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. Benchić-lang: A benchmark for discriminating between bosnian, croatian, montenegrin and serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal Dependencies for

Serbian in comparison with Croatian and other Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 39–44, Valencia, Spain. Association for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 4717–4726.

Mihailo Škorić. 2024a. Novi jezički modeli za srpski jezik. *Infoteka*, 24. ArXiv:2402.14379.

Mihailo Škorić. 2024b. Novi jezički modeli za srpski jezik. *Infoteka*, 24.

Mihailo Škorić and Nikola Janković. 2024. New textual corpora for serbian language modeling. *arXiv preprint arXiv:2405.09250*.

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Dusko Vitas, Mihailo Skoric, and Milica Ikonić Nešić. 2022. Distant reading in digital humanities: Case study on the Serbian part of the ELTeC collection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3337–3345, Marseille, France. European Language Resources Association.

Ranka Stankovic, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for Serbian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3954–3962, Marseille, France. European Language Resources Association.

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dien Dinh. 2024. ViGLUE: A Vietnamese general language understanding benchmark and analysis of Vietnamese language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4174–4189, Mexico City, Mexico. Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: Less is more in multilingual models. In *Text, Speech, and Dialogue (TSD 2020)*, volume 12284 of *Lecture Notes in Computer Science*, pages 104–118. Springer.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. Basqueglue: A natural language understanding benchmark for basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, and 1 others. 2025. Mmluprox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene superglue benchmark: translation and evaluation. *arXiv preprint arXiv:2202.04994*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

A Appendix

BoolQ	**Passage:** *Stanica 19 – Stanica 19 je američka akciono-dramska televizijska serija koju je kreirala Stejsi MekKee za ABC. MekKee, Šonda Rajms, Betsy Beers i Paris Barclay su izvršni producenti serije, koja je drugi spin-of serije Uvod u anatomiju. Smeštena u Sijetl, serija se fokusira na živote muškaraca i žena u Vatrogasnoj stanici 19 Sijetl. Seriju produciraju Shondaland i ABC Studios, a MekKee je izvršni producent.* **Question:** *is da li je stanica 19 spinoff serije uvod u anatomiju* **Answer:** `Yes`
CB	**Text:** *To je deo njihove religije, religije kojoj se ne podsmevam jer sadrži mnoge elemente koji se poklapaju sa našim, iako joj nedostaje naša istina. Na jednom od njihovih velikih festivala izvode ritual isterivanja đavola iz tela. Prvo dolaze bubnjari – moram reći da ženama nije dozvoljeno da učestvuju u ovom ritualu, i možda će se dame ovde složiti sa mnom da su zbog toga srećne.* **Hypothesis:** *nijedna žena ne sme da učestvuje u ovom ritualu* **Entailment:** `True`
COPA	**Premise:** *Utrnulo mi je stopalo.* **Question:** *What's the EFFECT of this?* **Alternative 1:** *Obuo sam cipele.* **Alternative 2:** *Protresao sam stopalo.* **Correct Alternative:** `2`
MultiRC	**Paragraph:** *(...) U januaru 2000. godine, pomoćnik državnog sekretara Karl Inderfart i koordinator Stejt departmenta za borbu protiv terorizma Majkl Šihan sastali su se sa generalom Mušarafom u Islamabadu, nudeći mu kao nagradu za pakistansku saradnju mogućnost predsedničke posete u martu. Takvu posetu je Mušaraf žarko želeo, delimično kao znak legitimnosti njegove vlade. On je dvojici izaslanika rekao da će se sastati sa Mula Omarom i izvršiti pritisak na njega u vezi sa Bin Ladenom. Međutim, oni su se vratili u Vašington izveštavajući da Pakistan „u stvari verovatno neće ništa učiniti, s obzirom na ono što vidi kao prednosti talibanske kontrole nad Avganistanom." Predsednik Klinton je trebalo da putuje u Indiju. Stejt department je smatrao da on ne bi trebalo da poseti Indiju, a da ne poseti i Pakistan. (...)* **Question:** *Koga je Mušaraf nameravao da sretne kako bi razgovarao o Bin Ladenu, nakon sastanka sa pomoćnikom američkog državnog sekretara i koordinatorom Stejt departmenta za borbu protiv terorizma?* **Candidate answers:** *Rumunsku policiju ili tajnu službu* (`F`), *Predsednik Klinton* (`F`), *Majkl Šihan* (`F`), *Odeljenje za borbu protiv terorizma* (`F`), *Muli Omaru* (`T`)
ReCoRD	**Paragraph:** *Posle svrgavanja Mihajla V (1041—1042) vlast prelazi na ostarele ćerke Konstantina VIII caricu Zoju i monahinju Teodoru. Međutim njih dve nisu imale sposobnosti za vladanje državom, a pored svega je među njima vladala velika netrpeljivost zbog čega je vrlo brzo postalo jasno da njih dve neće biti u stanju da vode državu zbog čega je šezdesetčetvorogodišnja carica Zoja 11. 06. 1042. godine stupila u svoj treći brak, a njen novi muž i car postao je senator Konstantin Monomah, predstavnik civilnog plemstva, dok su ćerke Konstantina VIII zvanično bile njegove savladarke dobivši odrešene ruke za trošenje državnog novca radi sopstvenog zadovoljstva* **Query** Posle svrgavanja @placeholder @placeholder je stupila u brak sa @placeholder **Correct Entities:** `Mihajla V, carica Zoja, Konstantinom Monomahom`
RTE	**Text:** *Indijski proizvođač čelika Tata Steel, iz grupe Tata, kupuje anglo-holandskog čeličnog giganta Corus Group za 6,7 milijardi funti (12 milijardi dolara), čime postaje peti najveći proizvođač čelika na svetu. 70-godišnji predsednik grupe Tata, Ratan Tata, iz jedne od najpoznatijih indijskih poslovnih porodica, pobedio je u trci protiv 52-godišnjeg Benjamina Steinbrucha, poznatog brazilskog direktora koji je glavni i najveći vlasnik kompanije Companhia Siderúrgica Nacional (CSN). (...)* **Hypothesis:** *Tata grupa je osnovana pre 70 godina.* **Entailment:** `False`
WSC	**Text:** *Naučnici proučavaju tri vrste riba koje su nedavno pronađene kako žive u Indijskom okeanu. Počeli su pre dve godine.* **Coreference:** `True`

Table 5: Training set examples from the tasks in Serbian SuperGLUE. **Bold** text represents part of the example format for each task. Text in *italics* is part of the model input. *Underlined* text is specially marked in the input. Text in a `monospaced font` represents the expected model output. Some text parts are shown as excerpts (...).

	English example	Serbian translation	Error type	Comment
AXB	**sentence1**: Most of the graduates of my program have moved on to other things *because the jobs suck*.	**sentence1**: Većina diplomaca mog smera je nastavila dalje *jer poslovi su užasni*.	word order	*jer poslovi su užasni* → *jer su poslovi užasni*
AXG	**premise**: The investigator wanted to interview the witness in person, but he was *unreachable*.	**premise**: Istražitelj je želeo da intervjuiše svedoka lično, ali on je bio *nedostižan*.	wrong word meaning	*nedostižan* → *nedostupan*
CB	A: Uh, well, that gets (...) , B: I don't think they have a whole lot to say, even.	А: Ух, па, то добија (...) В: Не мислим да они имају много тога да кажу, чак и.	cyrillic script	/
RTE	**premise**: (...) Coughlin is America's fastest swimmer ever ... **hypothesis**: Coughlin is the fastest swimmer in the world.	**premise**: (...) Koflin je najbrža američka plivačica ... **hypothesis**: Koflin je *najbrži plivač* na svetu.	wrong gender	Coughlin (female), therefore → (*najbrža plivačica*)
WSC	**text**: By rolling over in her upper berth, Tatyana could look... **span1**: Tatyana **span2**: *Her*	**text**: Dok se prevrtala u svom gornjem krevetu, Tatjana je mogla... **span1**: Tatjana **span2**: *Njena*	wrong coreference resolution	*span2* (*Njena*) does not exist in *text*

Table 6: Error analysis of automatic translation from English to Serbian with Gemini Flash 2.5. The table summarizes representative examples of each error type, with some items shown as excerpts (. . .). English example presents the original text, which was the model input. Serbian translation is model output. **Bold** text presents different parts of model input. Text in *italic* is part of the text where error occurs.

A.1 Annotation Instructions

You will be presented with examples from different datasets, where each example comprises the original English text and its translation to Serbian.

Your task is to rate the translation quality on a scale of 1 to 5 by comparing the Serbian output to the English original. For each example, consider the provided context (column names and full text) to ensure the translation accurately conveys the original meaning and maintains stylistic consistency. Prioritize the correct translation or transcription of named entities, and ensure the final text is both grammatically correct and natural-sounding in Serbian.

Score descriptions:

- **1** - the translation is not at all related to the original English text
- **2** - the translation alters the original meaning or omits key information
- **3** - the translation conveys the original meaning, but there are minor mistakes in named entity handling
- **4** - the translation is almost perfect: it conveys the original meaning, it is complete, and named entities and personal names are properly handled, but there are minor stylistic or syntactic errors
- **5** - the translation is completely accurate and fluent in Serbian

Pay special attention to:

- **Word order** - is the word order correct in Serbian?

- **Noun words** - is the grammatical number, gender and case correct?
- **Verbs** - are verbs used in proper tense and form in relation to the English text?
- **Named entities** - are named entities translated correctly and consistently across all text parts (columns)?
- **Personal names** - are personal names properly and consistently transcribed across all text parts (columns)?

Examples

English source	Serbian translation	Score	Explanation
question: do you have to use a designated hitter in the american league	*question:*da li morate da koristite određenog udarača u američkoj ligi	4	Meaning of the word "designated" is incorrectly translated given the context
passage: Egyptian television is preparing to film a series that highlights the unity and cohesion of Moslems and Copts as the single fabric of the Egyptian society (...).	*passage:* Egipatska televizija se priprema da snima seriju koja ističe jedinstvo i koheziju muslimana i kopta kao jedinstvene celine egipatskog društva (...).	2	The translation of the phrase "the unity and cohesion" is incorrectly translated and the translation is in the wrong case, altering the original meaning.
paragraph: Following his meeting with US Assistant Secretary of State and US State Department's counterterrorism coordinator, who did Musharraf intend to meet to discuss Bin Laden?	*paragraph:* Nakon sastanka sa pomoćnikom američkog državnog sekretara i koordinatorom Stejt departmenta za borbu protiv terorizma, sa kim je Mušaraf nameravao da se sastane kako bi razgovarao o Bin Ladenu?	3	Syntactic error: main clause should come before the subordinate clause.

Additional instructions

- Different examples may include different text parts (columns) because they come from different datasets. For instance, one example may have *sentence_1* and *sentence_2*, while another may have *passage*, *question*, and *answer*. This is not an error. Column names should **not** be translated into Serbian; only evaluate the text following *column_name:*. All texts should be evaluated using the same criteria, regardless of which parts they contain.
- Focus on the meaning of the translation, not on text formatting.
- Check how named entities and personal names should be translated or transcribed into Serbian (e.g., using Wikipedia). Pay particular attention to consistency of translation and transcription across the entire example.
- Score each example independently.