Computer Science and Information Systems 00(0):0000–0000 https://doi.org/10.2298/CSIS251215056L

Paper Type: Research Paper

Efficient Serbian Language and Question Answering Adaptation of LLaMA 3.2 model ⋆

Adela Ljajić[1], Marko Zubac[1], Gorana Gojić[1], Miloš Košprdić[1], Nikola Milošević[1,2], and Dragiša Mišković[1]

[1] The Institute for Artificial Intelligence Research and Development of Serbia
Fruškogorska 1, 21000 Novi Sad
[adela.ljajic (corresponding author), marko.zubac, gorana.gojic, milos.kosprdic, nikola.milosevic, dragisa.miskovic]@ivi.ac.rs
[2] Bayer A.G.
Muellerstrasse 178, 13353 Berlin, Germany
nikola.milosevic@bayer.com

Abstract. This paper presents a two-stage, parameter-efficient adaptation of the LLaMA 3.2 11B Vision model for Serbian language generation and question answering. In the first stage, a Serbian-language adapter is trained with QLoRA on a large web-scale Serbian corpus and a curated collection of formal academic texts, adapting the model to Serbian using web-scale and formal academic texts. In the second stage, a task-specific QA adapter is stacked atop the language-adapted model and fine-tuned on a Serbian translation of the Natural Questions benchmark, with evaluation on both in-domain and out-of-domain Serbian QA datasets. The approach remains feasible on a single 40 GB GPU. The Serbian language adapter improves accuracy on a Serbian Cloze-style grammatical completion benchmark by 10 percentage points, while the full stacked model achieves strong gains in question answering, improving out-of-domain performance by up to 39.08% in F1, 4.31% in BERTScore, and 29.35% in ROUGE-L.

Keywords: Causal Language Modeling, Supervised Fine-Tuning, Serbian Language, LLaMA, QLoRA.

1. Introduction

The rapid progress of large language models (LLMs) has redefined the boundaries of natural language understanding and generation. Models such as GPT-5, PaLM 2, and LLaMA 3 demonstrate remarkable reasoning and compositionality capabilities across a wide range of tasks. However, the benefits of this progress are unevenly distributed: most models are optimized for English or other high-resource languages, while languages with smaller digital footprints, such as Serbian, remain underrepresented in both datasets and model evaluation benchmarks.

⋆ This article is an extended and significantly revised version of the abstract previously published in the Artificial Intelligence conference 2025, Book of Abstracts.

This linguistic imbalance poses challenges not only for model fairness and inclusivity but also for the practical deployment of AI systems in multilingual regions. Training large models from scratch for low-resource languages (LRL) is often infeasible due to the immense computational and data requirements. Consequently, efficient adaptation techniques have become essential for enabling equitable access to LLM technology.

Serbian presents several linguistic challenges for large language models. It is a morphologically rich South Slavic language with an extensive case system, productive inflection, and free word order, which increases the difficulty of next-token prediction. The coexistence of digraphs, phonetic spelling conventions, and two scripts (Latin and Cyrillic) further complicates tokenization and representation learning. These characteristics, combined with the scarcity of high-quality annotated resources, make Serbian a representative and demanding example of an LRL for model adaptation.

Recent work on multimodal pretraining shows that visual and textual modalities are largely complementary and that incorporating visual data does not degrade language modeling performance while enabling richer representations and emergent capabilities through cross-modal learning. This suggests that multimodal foundation models may offer greater representational capacity even when applied to text-only tasks, motivating exploration of multimodal architectures for low-resource language generation [31]. This motivates the use of a multimodal foundation model as a basis for low-resource language adaptation, while the present study focuses exclusively on text-based Serbian language generation and QA.

The choice of the LLaMA-3.2-11B-Vision model was guided by both methodological and practical considerations. Larger models consistently demonstrate stronger cross-lingual transfer and improved performance on morphologically rich languages, as shown in prior work on multilingual and cross-lingual language models [7, 36, 11]. At the same time, parameter-efficient fine-tuning methods such as QLoRA make it feasible to adapt models of this scale on a single GPU, enabling a balance between model capacity and computational efficiency. Alternative open models such as Mistral or Gemma were also considered. However, LLaMA-3.2 offered more mature and stable support for parameter-efficient adaptation at the time of experimentation, which allowed us to maintain a controlled and reproducible experimental setup. We selected the Vision variant to ensure compatibility with future multimodal extensions, while using it in a text-only setting in this work.

In this study, we investigate Quantized Low-Rank Adaptation (QLoRA), a parameter-efficient fine-tuning (PEFT) strategy for adapting a large multimodal language model to an LRL setting. QLoRA has emerged as a promising solution for scaling model adaptation to low-resource settings. This technique allows the reuse of the expressive power of pretrained models while updating only a fraction of their parameters. In this study, we explore the potential of such an approach by fine-tuning a state-of-the-art LLaMA 3.2 11B Vision [3] (L3.2-11B-V) model on a Serbian text corpus, aiming to enhance its capacity for monolingual text generation while maintaining feasible training costs.

The proposed framework comprises two stages. In the first stage, we fine-tune a Serbian language adapter for causal language modeling (CLM) using the SrWaC and STARS datasets. In the second stage, the adapter is extended with a task-specific adapter for question-answering (QA), trained on a Serbian translation of the Natural Questions

[3] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision

dataset (NQsr). Both stages apply the same QLoRA fine-tuning strategy with 4-bit quantization, enabling the adaptation of the L3.2-11B-V model under modest computational constraints.

The contributions of this work are threefold:

- we present an efficient fine-tuning methodology that enables the adaptation of large multimodal LLMs to LRL using limited hardware resources
- we develop a Serbian language adapter fine-tuned on large-scale Serbian web and academic corpora and evaluate its grammatical competence using a Cloze-style benchmark, and
- we build a QA adapter on top of the Serbian-language adapter using a Serbian translation of the Natural Questions dataset, and we evaluate it on the QAsr benchmark using F1 and BERTScore.

These contributions provide evidence that quantized, adapter-based fine-tuning offers a practical and extensible pathway for adapting large language models to underrepresented languages and associated downstream tasks.

2. Related Work

LLMs such as GPT-4, Claude 3 and Claude 4, PaLM 2, and LLaMA have advanced the state of the art in natural language processing by enabling general-purpose reasoning and generation capabilities [23, 2, 3, 1, 32]. However, their training data remain dominated by English and a few other high-resource languages, resulting in uneven performance across linguistic communities [15]. This imbalance is particularly evident for morphologically rich LRLs such as Serbian, which are underrepresented in both pretraining data and evaluation benchmarks. At the same time, recent survey work documents a growing ecosystem of Serbian NLP corpora, lexical resources, and tools, while initiatives such as ReLDI have contributed to the development and dissemination of open language resources for Serbian and related South Slavic languages [22, 27]. These considerations further motivate the need for scalable and resource-efficient methods for language adaptation. Recent multilingual benchmark studies such as Global-MMLU and MMLU-ProX further demonstrate that LLM performance remains uneven across low-resource and morphologically rich languages, motivating continued research on efficient language adaptation methods [29, 35].

The challenge of extending LLM capabilities to LRLs has been approached from several directions. Multilingual models such as mBERT [9], XLM-R [7], and BLOOM [4] attempt to cover hundreds of languages jointly, but their performance degrades significantly on morphologically rich or syntactically distant languages [7]. For South Slavic languages, regional models such as BERTić [20] and CroSloEngual BERT [33] have improved representation. However, these models remain encoder-only and are therefore primarily suited to representation learning and extractive tasks, whereas our work targets generative question answering using a decoder-only architecture.

Recent work on document-based question answering further emphasizes the importance of combining language models with task- and domain-specific adaptation. Hrvačević et al. [13] provide a review of LLM-based document QA systems and highlight that fine-tuning and domain adaptation remain important for improving QA performance, particularly in multilingual and low-resource settings. Related university-focused DBQA

work [10] also shows that combining retrieval mechanisms with fine-tuned transformer-based QA models improves answer accuracy on domain-specific data. Although these studies focus on document-based QA pipelines rather than low-resource language adaptation, they support the broader motivation for our two-stage approach, where Serbian language adaptation is followed by task-specific QA fine-tuning.

The emergence of PEFT has provided a practical route for adapting large models in constrained computational environments. LoRA [14] introduced a low-rank adaptation mechanism that updates only a small number of parameters while freezing the base model. Subsequent variants, such as QLoRA [8], further reduced memory requirements by using 4-bit quantization, enabling fine-tuning of 10B–70B models on a single GPU. Other related approaches include prefix-tuning [19], adapter fusion [24], and prompt-tuning [18], all of which aim to achieve effective adaptation with minimal training cost.

PEFT methods improve the performance of encoder-only multilingual models when adapting to different LRLs and downstream tasks [26, 25, 24]. The work in [26] proposes a scheme that uses adapters to separate language-specific and task-specific knowledge and shows that a task adapter trained on a high-resource language can be combined with language adapters for low-resource or even unseen languages to achieve strong cross-lingual transfer. The subsequent work in [25] establishes a unified methodology for using adapters in encoder-based models, supporting the fast adaptation of pretrained multilingual encoders and improving the reproducibility of adapter-based experiments. Building on the availability of task adapters, [24] introduces an attention-based fusion mechanism that combines multiple task adapters and yields the largest gains on target tasks with small training datasets, a common situation in LRL settings. Together, these contributions indicate that adapters are an effective tool for extending multilingual encoder-based models to LRLs with limited data and compute by reusing and composing knowledge learned from high-resource languages.

While most adapter-based PEFT work focuses on encoder-only architectures and task adapters, recent work has started to explore decoder-only models. In [30], task adapters are trained for LLaMA 3.1 and Mistral v0.3 on QA, NLI, reading comprehension, and reasoning as downstream tasks. The experimental results include an evaluation of Serbian language as a low-resource language in the NLI setting, where adapter-based methods improve F1 over the base model by 23–33.2%, depending on the PEFT method. However, systematic evaluation of language adapters in decoder-only architectures for low-resource languages remains limited, as most recent research has focused on adapters in encoder-only architectures.

Our work builds on the approach in [26] by using separate language and task adapters to specialize a multilingual model for Serbian. Unlike [26], which evaluates encoder-based architectures, we apply a two-stage PEFT pipeline to fine-tune a decoder-only architecture, following the direction in [30]. We first train a Serbian-language adapter, then specialize the model for the downstream QA task with a task adapter, using quantized adapters to improve training efficiency. We use the multilingual L3.2-11B-V model as the base architecture, since prior work indicates that higher-capacity multilingual models better support adaptation to LRLs [7]. To enable future extensions of Serbian-language evaluation to multimodal tasks, we select a multimodal architecture. Our results indicate that quantized adapters provide a practical pathway for extending multimodal LLMs to Serbian and Serbian-language QA.

Table 1. Overview of datasets used for CLM fine-tuning (in tokens) and SFT (in samples).

Language Adapter (SrLangA)			
Dataset	**Use**	**Tokens**	**Notes**
SrWaC	Train	∼1.1 B	Web text, diverse registers
STARS	Train	∼1.8 B	Academic Serbian

QA Adapter (SrQAA)			
Dataset	**Use**	**Samples**	**Notes**
NQsr_train	Train	90k	Serbian NQ translation
NQsr_eval (5%)	Eval	5k	In-domain evaluation
QAsr	Eval	100	Out-of-domain, Custom QA

3. Datasets

The fine-tuning and evaluation of our SrLangA and SrQAA adapters were conducted using several datasets covering diverse linguistic and functional domains. To ensure both in-domain learning and generalization, we combined publicly available corpora with custom-curated datasets for evaluation.

3.1. Language Adapter (SrLangA) Training Datasets

SrWaC [21] served as the first training corpus for fine-tuning the language adapter. It consists of web-scraped text from a variety of domains, including news portals, blogs, and discussion forums, making it representative of modern Serbian usage. SrWaC is made up of 555M linguistic tokens (words) and around 1.1B L3.2-11B-V tokens.

STARS is a Serbian subset of the ZNANJE dataset [37], a curated collection of Serbian texts covering formal academic style. This dataset was used to enhance stylistic diversity and lexical richness. The inclusion of STARS aimed to expose the adapter to more formal and domain-specific expressions that are less common in web text.

The combination of SrWaC and STARS provided a balanced source of Serbian language data suitable for CLM, helping the adapter generalize across both formal and informal registers.

Tokenization, text cleaning, and the chunking strategy. Both datasets were preprocessed and tokenized uniformly before fine-tuning, as expected to ensure linguistic coherence and compatibility with the architecture of the L3.2-11B-V model. For the STARS dataset, additional text cleaning was done to clear the automatically generated section numbering patterns typically present in academic texts (e.g., 1.1, 1.2.3, 4.5.6.7). Such patterns were removed using a script that identified numerical sequences corresponding to hierarchical section markers and removed them from the text. This normalization step

prevented overfitting to non-linguistic structural tokens and ensured that the model focused on linguistic content during fine-tuning rather than on document formatting artifacts. After text normalization, all data were converted to UTF-8-encoded CSV format, with missing entries filled out and coerced to string type.

During a preliminary inspection, Cyrillic text was segmented into more tokens per word than its Latin transliteration, resulting in less efficient use of context. All texts in both SrWaC and STARS were normalized to the Serbian Latin script prior to tokenization. Texts containing Serbian Cyrillic characters were transliterated using a deterministic mapping between the Serbian Cyrillic and Latin alphabets, while texts already written in Latin script required no transliteration. This normalization reduced script-related variability across the corpora and enabled more efficient tokenization while preserving the original linguistic content. We retain the original LLaMA 3.2 tokenizer for compatibility with the pretrained model and QLoRA setup. On the normalized corpus, it achieves 2.78 bytes per token and 2.72 characters per token (approximately 2–3 characters per token), indicating a reasonably compact subword representation for Serbian.

Afterward, a sentence-based chunking strategy was implemented to segment long documents into sequences compatible with a context window. The approach utilized the PunktSentenceTokenizer, implemented in NLTK [5], with an abbreviation set specific for Serbian (including npr., itd., dr., prof., mr., gdin., and gdja.) to achieve accurate sentence boundary detection. Each document was first divided into sentence-level units, which were progressively accumulated into clusters of up to 2048 tokens. If an individual sentence exceeded this limit, it was further split into smaller word-level segments to avoid overflow. The resulting chunks were then tokenized with the L3.2-11B-V tokenizer, using the end-of-sequence token for padding, while sequences shorter than 15 tokens were filtered out during preprocessing to exclude trivial or low-information samples. For SrWaC, out of 10,543,879 examples, 2,347,095 (22.26%) contained 15 or fewer tokens. For STARS, only 107 out of 1,065,779 examples (0.01%) fell into this category. These differences reflect the distinct characteristics and domain-specific structures of the two corpora, resulting in a substantially different filtering effect.

3.2. SrQAA Adapter Training Datasets

For task-specific fine-tuning, we used the NQsr dataset, a Serbian translation of the NQ dataset [17]. The NQsr dataset comprises approximately 90,000 question–answer pairs used for SFT, while an additional 5,000 examples are held out for in-domain evaluation, as summarized in Table 1. The original dataset contains Google search queries paired with both long and short versions of the answer. We translated the queries and long answers, as the short answers are already included in the long ones. In the initial experiments, we used only a neural machine translation (NMT) model via the Google Translate paid API. We observed that entities with non-Serbian names, which naturally appear in the English dataset, were occasionally translated through transliteration rather than through phonetic adaptation or by preserving the original form, resulting in linguistically incorrect translations. To address this issue, we implemented a multi-stage translation pipeline. First, named entity recognition (NER) was applied to the English text using the spaCy NER model [12] to identify entities likely to contain non-native Serbian names, such as personal names, locations, and product names. The surrounding context was translated using the Google Translate API, while the extracted entities were translated separately using the

Table 2. Descriptive statistics of adequacy, fluency, and grammar scores across four quartiles (Q1-Q4) and all examples in evaluation subset (All).

	Adequacy			**Fluency**			**Grammar**		
	Mean	**Std**	**Median**	**Mean**	**Std**	**Median**	**Mean**	**Std**	**Median**
Q1	4.604	0.218	4.670	4.395	0.400	4.500	4.530	0.306	4.500
Q2	4.480	0.327	4.570	4.191	0.467	4.200	4.367	0.436	4.500
Q3	4.674	0.287	4.800	4.265	0.528	4.400	4.312	0.484	4.430
Q4	4.494	0.373	4.500	4.182	0.581	4.170	4.239	0.575	4.330
All	4.563	0.313	4.600	4.258	0.498	4.400	4.362	0.466	4.500

GPT-4o model [23], leveraging its broader contextual knowledge to generate more linguistically appropriate Serbian translations. Finally, contextual translations and corrected entity translations were merged, and an LLM-based post-processing step was applied to adjust grammatical inflections and ensure consistency with Serbian morphology.

Validation of NQsr translations. We evaluate translation quality on a sample of 100 question-answer pairs from the NQsr dataset. The sample is constructed using stratified sampling based on named entity density (NED), defined as the ratio of named entity words to total words in a sample. Higher NED means less surrounding context for the NMT model, which is observed to lead to lower translation quality. We compute NED on the English source samples using a spaCy NER model, since English NER models achieve higher detection accuracy than those available for Serbian [34]. Samples are divided into four NED quartiles and 25 samples are drawn uniformly at random from each quartile, yielding the evaluation subset that spans the full range of translation difficulty.

Annotation is conducted per sentence in a sample to account for variation in translation quality within a single sample. We use a semi-manual procedure that involves an LLM[4] assigning the initial annotations later reviewed by a human reviewer to reduce annotation time and cost. First, an LLM assigns scores to each sentence following the annotation rules described in Appendix C. A human rater then reviews and corrects scores assigned by the LLM for full compliance with the annotation rules, with the sample score being calculated as the average of sentence-level scores. Each sample is assessed on adequacy, fluency, and grammatical correctness. Adequacy measures how well the translation preserves the meaning of the source, while fluency measures how natural the translation sounds to a native speaker. These two dimensions are well established in translation quality assessment and follow the 5-point scale introduced by [16]. To address annotation inconsistencies reported for the original scale [6], we provide more detailed, proportional score descriptions. We also introduce grammatical correctness as an additional dimension, since Serbian has a richer morphological system than English and grammatical errors are qualitatively distinct from fluency issues.

NQsr dimension analysis. Descriptive statistics across NED quartiles shown in Table 2 reveal that both grammatical correctness and fluency degrade as NED increases. Gram-

[4] We use Claude Sonnet 4.6 with the adaptive thinking technique at a high effort level.

mar score decreases monotonically from Q1 to Q4, while fluency shows no consistent trend in mean scores across quartiles. However, standard deviation increases monotonically across quartiles for both dimensions, indicating that higher NED introduces growing inconsistency in translation quality for both fluency and grammar. Overall means remain relatively high (grammar: 4.362, fluency: 4.258), suggesting that translation artefacts are limited in absolute terms. For model fine-tuning, these findings imply that the majority of training samples provide reliable supervision. However, samples with higher NED introduce unevenly distributed noise in both grammatical accuracy and phrasing naturalness. This may impair the fine-tuned model's ability to generate grammatically correct and naturally phrased text, particularly in outputs with high entity density. Adequacy scores show no consistent dependence on NED, as neither the mean nor the standard deviation changes monotonically across quartiles. This suggests that NED does not systematically affect the semantics of the translated texts. Overall, adequacy is the highest-rated and most consistent dimension, with an overall mean of 4.563 and the lowest standard deviation of 0.313 across all three dimensions, indicating that the translated texts reliably preserve source semantics.

3.3. SrQAA Adapter Evaluation Datasets

The evaluation of the question-answering adapter was conducted using two complementary datasets designed to assess both in-domain and out-of-domain generalization.

The first evaluation dataset represents an in-domain benchmark and consists of 5,000 examples, corresponding to a 5% held-out subset of the NQsr dataset described in Section 3.2. This subset preserves the distribution of question and answer formats present in the training corpus. It provides a reliable estimate of model performance under conditions that closely match the SFT regime.

The second evaluation dataset, referred to as QAsr, contains 100 question–answer examples and serves as an out-of-domain benchmark. QAsr is a custom Serbian question-answering dataset constructed from heterogeneous public sources and curated to ensure diversity in topics, styles, and difficulty levels. Gold-standard reference answers were created using the GPT-5.1 model. This dataset enables assessment of the model's robustness when confronted with previously unseen domains and lexical variability.

Together, these two datasets provide a comprehensive assessment of the question-answering adapter. The held-out portion of the NQsr dataset measures performance under conditions that closely mirror the supervised fine-tuning distribution, allowing evaluation of learning efficiency and consistency. In contrast, the QAsr dataset contains questions and linguistic patterns not present in the training corpus, enabling an examination of robustness, domain transfer, and the model's ability to generalize beyond memorized structures. The combination of these two datasets offers a balanced and rigorous evaluation framework that captures both task fidelity and real-world adaptability.

4. Methodology

This study proposes an efficient two-stage fine-tuning pipeline for adapting the L3.2-11B-V model to Serbian and downstream tasks. The approach leverages PEFT through QLoRA to minimize computational cost while preserving the expressive power of the base model. Figure 1 provides an overview of the workflow, which consists of:

– fine-tuning a Serbian language adapter for CLM (SrLangA),
– fine-tuning a task-specific adapter for Serbian question answering (SrQAA) on top of SrLangA.

Each stage is followed by an independent evaluation to assess generalization and robustness.

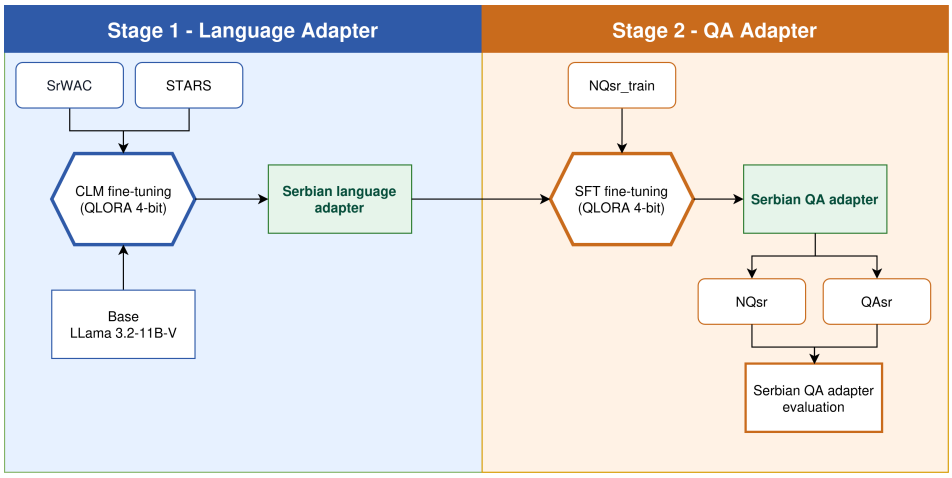


Fig. 1. Two-Stage Adapter Fine-Tuning and Evaluation Pipeline (Language + Task Adapters)

4.1. Language Adapter (SrLangA) Fine-Tuning

The first stage focuses on adapting the model to Serbian linguistic structure and stylistic variability. We fine-tuned the base L3.2-11B-V model on the SrWaC and STARS datasets using the CLM objective. The model predicts the next token given the preceding context, enabling it to internalize Serbian morphology, syntax, and stylistic patterns.

Adapter modules were inserted into the transformer's attention layers, updating only a small subset of parameters while keeping the remaining weights frozen. Specifically, LoRA modules were applied to the query and value projection matrices. This configuration enables efficient specialization without retraining the full model, which would be infeasible for an 11-billion-parameter architecture under limited hardware resources.

After fine-tuning, the SrLangA was merged with the base model and used as the Serbian-language foundation for the subsequent QA adaptation stage. Additional implementation details and hyperparameters are provided in Section 5.

4.2. Question-answering Adapter (SrQAA) Fine-Tuning

The second stage extends the SrLangA with a task-specific adapter for QA, leveraging the pretrained language adapter as a foundation. The NQsr dataset served as the training resource, containing search queries paired with long-form answers [17].

The SrQAA adapter was trained using supervised fine-tuning (SFT) to generate an answer given a question. Each example was converted into a structured input–output pair, where the model learns to generate the answer sequence conditioned on the question as the instruction. This formatting encourages the model to follow instructions explicitly and improves alignment with downstream QA behavior. LoRA modules were applied to the attention projection matrices (queries and values) to selectively adapt attention mechanisms relevant to QA while preserving the underlying Serbian representations from the language adapter.

Evaluation was performed using two datasets described in Section 3.3: a 5% subset of NQsr for in-domain assessment and the QAsr dataset for out-of-domain evaluation. Performance was measured using F1-score, ROUGE-L to capture lexical accuracy, and BERTScore to measure semantic similarity.

4.3. Summary of the Two-Stage Pipeline

The methodology follows a modular, extensible design:

- The SrLangA adapter provides a Serbian monolingual foundation trained under a CLM objective.
- The SrQAA adapter builds on this foundation to specialize the model for extractive and abstractive question answering.

This design highlights that meaningful adaptation to a low-resource language can be achieved with limited computational resources by leveraging quantized adapter-based fine-tuning. The modularity of the pipeline also enables future extensions, including multimodal Serbian-language tasks using the visual capabilities of the underlying L3.2-11B-V architecture.

5. Experimental Setup

Before launching the full training run, a series of controlled pilot experiments was conducted to tune the main fine-tuning hyperparameters, balancing GPU utilization, training stability, model quality, and overall training time. Through iterative experimentation, we tuned the effective batch size and gradient accumulation strategy to fully saturate GPU memory without triggering out-of-memory errors. The final configuration, with a per-device batch size of 4 and gradient accumulation steps of 8, resulted in nearly 100% utilization on our hardware while preserving training stability. We employed a warm-up phase of 1000 steps to stabilize early optimization, a learning rate of 2e-5, identified as a stable and effective setting for LoRA-based fine-tuning of a highly quantized large model, and bf16 precision to maximize throughput while maintaining numerical robustness.

5.1. Hardware, Software, and Fine-Tuning Configuration

All experiments were conducted on a single NVIDIA A100 GPU with 40 GB of VRAM and a host system equipped with 512 GB of RAM. The experiments were executed on the fastMLe platform of the National AI Platform of Serbia, utilizing computational resources provided by the State Data Center. The software environment was based on the

nvidia-pytorch:23.08 container image. The base model used was L3.2-11B-V, loaded in 4-bit quantization using the QLoRA framework [8]. We employed the Hugging Face Transformers and PEFT libraries for adapter training, together with BitsAndBytes for quantized parameter handling. Training and evaluation pipelines were implemented in Python with PyTorch, and the following libraries: Transformers, PEFT, BitsAndBytes, and Accelerate.

Fine-tuning followed a two-stage procedure, consisting of a language adapter stage and a subsequent task adapter stage. In the first stage, the model was adapted to Serbian using a causal language modeling objective on the SrWaC and STARS corpora, where the model predicts the next token from left to right. LoRA adapter modules were inserted only into the query and value attention projections, and all remaining parameters of the base model were kept frozen. The effective training configuration for this stage was:

- Quantization: 4-bit NF4 with double quantization and bfloat16 compute.
- LoRA settings: rank $r = 16$, scaling factor $\alpha = 8$, dropout $= 0.1$.
- Attention projections updated: query and value only.
- Learning rate: 2e-5 with a linear scheduler and 1,000 warm-up steps.
- Per-device batch size: 4.
- Global batch size: 32 (gradient accumulation steps $= 8$).
- Maximum sequence length: 2,048 tokens.
- Number of epochs: 2.
- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and zero weight decay.

The choice of LoRA rank depends on task complexity and the degree of distribution shift. While prior work shows that low-rank adaptation can be effective for many downstream tasks, its behavior is task-dependent and may require higher capacity under substantial shifts. Given that our setting involves cross-lingual continued pretraining for Serbian, we adopt a moderate rank ($r = 16$) as a practical trade-off between adaptation capacity and computational efficiency. A more detailed discussion of rank sensitivity and related considerations is provided in Appendix A.

In the second stage, a task-specific QA adapter was trained with an SFT objective on the NQsr dataset. The objective used a masked loss that backpropagates only through the answer span, so that the model learns to produce an answer conditioned on the question while ignoring the loss on non-answer tokens. The QA adapter was stacked on top of the Serbian-language adapter and used the same query and value attention projections as the adaptation points. This design selectively modifies attention components relevant to question answering, while the underlying Serbian linguistic knowledge encoded by the language adapter remains intact, thereby keeping the overall computational footprint low.

Efficiency considerations. To further clarify the efficiency of the proposed approach, we compare different fine-tuning strategies for the LLaMA 3.2 11B model under a fixed hardware constraint (single NVIDIA A100 40GB GPU), as summarized in Table 3. While LoRA significantly reduces the number of trainable parameters, it does not reduce the base model's memory footprint, which must still be stored in full precision during training [14]. For an 11B-parameter model, this corresponds to approximately 22 GB of VRAM in bf16/fp16 precision, leaving limited memory for activations and other training components.

This limitation becomes particularly critical in our setup, where fine-tuning is performed with a maximum sequence length of 2048 tokens. At this context size, activation

memory dominates the total training footprint, making non-quantized LoRA fine-tuning impractical on a single 40 GB GPU, even with gradient checkpointing and minimal batch sizes. In practice, for transformer models of this scale, activation memory at sequence lengths of 2048 tokens can reach on the order of 10–20 GB depending on implementation details, which, when combined with full-precision model weights, exceeds the available memory budget of a 40 GB GPU.

QLoRA addresses this limitation by quantizing the base model to 4-bit precision while preserving performance [8]. This enables efficient fine-tuning of large models within constrained hardware environments. Notably, prior work shows that even 65B-parameter models can be fine-tuned on a single 48 GB GPU using QLoRA [8]. This makes QLoRA not only an efficiency optimization, but a necessary condition for feasibility in our experimental setting.

Table 3. Comparison of fine-tuning strategies for LLaMA 3.2 11B under a single A100 40GB GPU constraint

Approach	Precision	Trainable Params	Estimated VRAM	Feasible
Full fine-tuning	bf16/fp16	All	$>$ 40 GB	No
LoRA (no quantization)	bf16/fp16	Few (adapters)	$\sim$ 40 GB+	No
QLoRA	4-bit + LoRA	Few (adapters)	$<$ 40 GB	Yes

Note: The estimates are based on established memory characteristics of transformer training. Full fine-tuning requires storing model weights, gradients, and optimizer states, while LoRA reduces trainable parameters but not base model memory [14]. QLoRA reduces memory via 4-bit quantization, enabling fine-tuning of large models on a single GPU [8].

5.2. Evaluation Metrics

We evaluate the Serbian QA adapter with a mix of lexical and semantic metrics (F1, ROUGE-L, and BERTScore), reflecting complementary dimensions of answer quality.

F1, ROUGE-L, and BERTScore (QA Adapter). For QA, we report:

– F1 word-overlap, computed from precision P and recall R over normalized token sets of prediction and reference, with

$$F1 \;=\; \frac{2PR}{P+R}.$$

– ROUGE-L F1, derived from the longest common subsequence (LCS) between prediction y and reference y^*. With $\mathrm{LCS}(y, y^*)$ their LCS length,

$$P_{\mathrm{LCS}} = \frac{\mathrm{LCS}(y,y^*)}{|y|}, \quad R_{\mathrm{LCS}} = \frac{\mathrm{LCS}(y,y^*)}{|y^*|}, \quad F1_{\mathrm{ROUGE\text{-}L}} = \frac{2P_{\mathrm{LCS}}R_{\mathrm{LCS}}}{P_{\mathrm{LCS}}+R_{\mathrm{LCS}}}.$$

ROUGE-L F1 rewards partially correct answers when word order differs, which is particularly useful for free-form QA.

– BERTScore, which measures semantic similarity via contextual embeddings. Let $e(\cdot)$ denote token embeddings, and $\cos(\cdot,\cdot)$ cosine similarity. Precision and recall aggregate best-match similarities:

$$P_{\text{bert}} = \frac{1}{|y|}\sum_{y_i \in y} \max_{x_j \in y^*} \cos\big(e(y_i), e(x_j)\big), \quad R_{\text{bert}} = \frac{1}{|y^*|}\sum_{x_j \in y^*} \max_{y_i \in y} \cos\big(e(x_j), e(y_i)\big),$$

and $F1_{\text{bert}}$ is the harmonic mean of P_{bert} and R_{bert}. Following multilingual best practices, XLM-RoBERTa-large [7] is used as the underlying encoder to obtain robust scores on Serbian.

6. Results and Discussion

6.1. Question Answering Adapter

The effectiveness of the Serbian QA adapter (SrQAA) was evaluated on two complementary datasets described in Section 3.3: a held-out portion of NQsr representing in-domain performance (Table 6), and the QAsr dataset representing out-of-domain generalization (Table 7).

For clarity, we use the following notation for model variants throughout this section:

– **BASE** denotes the original LLaMA 3.2 11B Vision model without adapters.
– **SrLangA** is a Serbian language adapter trained on top of the base model using the CLM objective.
– **SrQAA** denotes a Serbian QA adapter trained on the base model merged with SrLangA.
– **SrQAA*** denotes a Serbian QA adapter trained directly on the base model without prior language adaptation.
– **Full stack** (**BASE+SrLangA+SrQAA**) refers to the base model merged with the Serbian language adapter and subsequently merged with the QA adapter.

The comparison with the unadapted base model serves as a controlled reference, allowing us to isolate the effects of Serbian-language adaptation, task-specific QA adaptation, and their interaction within the proposed two-stage framework.

Impact of NQsr translation artefacts on performance of QA models. Since the QA adapters are trained on NQsr distribution, with NQsr dataset demonstrating degradation of fluency and grammar with NED increasing, we test whether this translation noise has influenced the QA adapter by examining F1, ROUGE-L, and BERTScore across NED quartiles, as reported in Table 4. If the model has incorporated noise as a regular data pattern, no systematic decrease in automated metrics should be visible with increasing NED. Conversely, if the model managed to learn proper grammar and fluency from non-noisy samples, its outputs would diverge from the noisy Q4 references, and a consistent decline in automated metrics from Q1 to Q4 would be expected. For both models, automated metrics show a slight upward trend in F1 and ROUGE-L across quartiles, while BERTScore remains flat. Spearman correlations between NED and each metric, computed within each quartile at the individual sample level, are consistently near zero and largely

non-significant with inconsistent signs across quartiles and models. This confirms that entity density has no systematic influence on model performance even within individual NED ranges, further supporting the interpretation that translation noise has been absorbed into the learned data distribution, with adapter models generating outputs that mimic the stylistic and linguistic patterns of the noisy references.

Table 4. Performance overview of QA variants fine-tuned on NQsr train dataset. Spearman r denotes correlation between NED and the respective metric within each quartile. Statistically significant results are denoted with * ($p < 0.05$).

	F1				**ROUGE-L**				**BERTScore**			
	Mean	**Std**	**Median**	r	**Mean**	**Std**	**Median**	r	**Mean**	**Std**	**Median**	r
QA Adapter only (BASE+SrQAA)*												
Q1	0.140	0.061	0.140	-0.035	0.101	0.039	0.100	-0.083*	0.842	0.021	0.843	-0.137*
Q2	0.145	0.060	0.144	+0.005	0.102	0.038	0.101	+0.026	0.841	0.019	0.842	+0.022
Q3	0.150	0.063	0.145	+0.007	0.106	0.042	0.102	+0.019	0.842	0.019	0.842	+0.061*
Q4	0.154	0.067	0.152	-0.013	0.114	0.048	0.108	+0.050	0.841	0.021	0.842	-0.037
All	0.143	0.061	0.134	+0.070*	0.103	0.043	0.100	+0.088*	0.838	0.018	0.840	-0.016
Full Stack (BASE+SrLangA+SrQAA)												
Q1	0.147	0.060	0.145	+0.053	0.106	0.038	0.104	+0.005	0.847	0.017	0.846	-0.081*
Q2	0.155	0.061	0.154	+0.073*	0.108	0.040	0.105	+0.076*	0.846	0.017	0.845	+0.049
Q3	0.164	0.064	0.157	+0.026	0.116	0.041	0.110	+0.032	0.848	0.016	0.847	+0.044
Q4	0.166	0.063	0.161	-0.024	0.121	0.043	0.117	+0.012	0.847	0.017	0.846	-0.066*
All	0.159	0.067	0.155	+0.118*	0.113	0.040	0.108	+0.145*	0.846	0.017	0.845	+0.012

Table 5. Spearman correlations between automated metrics and manual annotation dimensions on the 100-sample evaluation subset. Statistically significant results are denoted with * ($p < 0.05$).

	QA Adapter (BASE+SrQAA*)			**Full Stack (BASE+SrLangA+SrQAA)**		
	Adequacy	**Fluency**	**Grammar**	**Adequacy**	**Fluency**	**Grammar**
F1	+0.228*	+0.169	+0.092	-0.004	+0.092	+0.067
ROUGE-L	+0.199*	+0.154	+0.076	+0.030	+0.089	+0.040
BERTScore	+0.201*	+0.192	+0.181	+0.224*	+0.230*	+0.172

Alignment of automated metrics with human judgment. To assess whether automated metrics reliably reflect human-perceived quality, we compute Spearman correlations between F1, ROUGE-L, and BERTScore and manual annotations of adequacy, fluency, and

grammar on the 100 instances of NQsr evaluation dataset. Results are reported in Table 5. BERTScore shows consistent positive correlations with adequacy across both model configurations and additionally correlates significantly with fluency in the full stack model, making it the most reliable proxy for human-perceived output quality among the three automated metrics. F1 and ROUGE-L show significant correlations with adequacy only for the QA adapter-only model, which we attribute to that model's tendency to mimic the phrasing of the noisy translated references rather than reflecting true lexical quality assessment. Correlations between fluency or grammatical correctness and any lexical metric are consistently small and non-significant across both configurations, indicating that F1 and ROUGE-L do not capture grammatical or phrasing quality as perceived by human annotators. These findings suggest that BERTScore is the most meaningful indicator of true model improvement in this setting, while F1 and ROUGE-L gains should be interpreted cautiously as they may partly reflect reference artefacts rather than genuine output quality. Reliable evaluation of fluency and grammar, therefore, requires human annotation or metrics specifically designed for grammatical assessment. Grammar evaluation of generated texts is partially addressed through a cloze-style grammatical evaluation on a separate set, reported in Section 6.2, while manual fluency analysis remains for future work.

In-domain evaluation (NQsr). Table 6 reports performance on the held-out NQsr evaluation set, which closely matches the distribution of the supervised fine-tuning data. The results show that progressively adding adapters consistently improves all QA metrics. The Serbian language adapter alone (BASE+SrLangA) yields only marginal gains over the base model, indicating that language adaptation by itself is insufficient for effective question answering. In contrast, the QA adapter trained directly on the base model (BASE+SrQAA*) substantially improves all metrics, demonstrating the impact of supervised task-specific fine-tuning. This configuration increases F1 and ROUGE-L while also yielding a notable gain in BERTScore, reflecting better semantic alignment with reference answers. The full stacked model (BASE+SrLangA+SrQAA) further improves performance over both the base model and the task-only adapter. Relative to the base model, it achieves improvements of 15.2% in F1, 21.3% in ROUGE-L, and 3.1% in BERTScore. Compared with the task-only adapter (BASE+SrQAA*), the gains are smaller but consistent across metrics, indicating that prior Serbian-language adaptation provides complementary benefits beyond QA supervision alone. While BERTScore provides evidence of improved semantic quality for the full stack model, the additional gains in F1 and ROUGE-L should be interpreted cautiously, as these metrics do not correlate significantly with human-perceived quality for the full stack model (Table 5) and may partly reflect closer alignment with NQsr reference phrasing.

The relatively small improvements in BERTScore indicate that most gains in semantic similarity are already achieved through task-specific QA fine-tuning (BASE+SrQAA*), whereas the addition of the language adapter yields smaller but consistent gains. As BERTScore is the most reliable proxy for human-perceived quality in this setting, the 3.1% improvement it captures over the base model is the most meaningful indicator of true quality gain. Overall, these results show that while supervised QA fine-tuning accounts for the majority of in-domain performance gains, incorporating a Serbian language

adapter before task adaptation yields additional improvements, leading to more stable and structurally coherent answer generation.

Table 6. Performance of the BASE (L3.2-11B-V) and adapter variants on the in-domain **NQsr_eval** Serbian QA dataset.

#	Model / comparison	F1	ROUGE-L F1	BERTScore F1
1	BASE	0.137	0.093	0.821
2	BASE+SrLangA	0.142	0.097	0.817
3	BASE+SrQAA*	0.147	0.106	0.841
4	Full_stack (BASE+SrLangA+SrQAA)	0.158	0.113	0.847
Relative improvements of Full_stack (%)				
	vs. BASE	15.2%	21.3%	3.1%
	vs. BASE+SrLangA	11.3%	15.6%	3.6%
	vs. BASE+SrQAA*	7.5%	6.4%	0.6%

Out-of-domain evaluation (QAsr). Table 7 presents results on the QAsr dataset, an out-of-domain benchmark designed to assess robustness and generalization beyond the supervised training distribution. In contrast to the in-domain setting, this dataset better reflects native Serbian stylistic and structural characteristics and introduces additional variability in topics and answer forms.

The Serbian-language adapter alone (BASE+SrLangA) performs worse than the base model on all metrics, confirming that language-modeling adaptation without task-specific supervision is insufficient for effective question answering. In contrast, the QA adapter trained directly on the base model (BASE+SrQAA*) yields clear improvements in F1 and ROUGE-L. These gains should be interpreted with caution, given that F1 and ROUGE-L do not correlate significantly with human-perceived quality for the full stack model in Table 5 and may partly reflect reference artefact mimicry rather than genuine answer quality improvement.

However, when the Serbian language adapter is combined with the QA adapter in the full stacked configuration (BASE+SrLangA+SrQAA), it achieves the best overall performance. Relative to the base model, it improves F1 by 19.4% and ROUGE-L by 21.9%, while also yielding a modest gain in BERTScore. Human evaluation correlations shown in Table 5 indicate that, for the full stacked model, BERTScore is the metric most consistently aligned with adequacy and fluency judgments, whereas F1 and ROUGE-L primarily reflect lexical and structural overlap. Compared to the language-only configuration, the full stack yields substantial gains of 39.1% in F1 and 29.3% in ROUGE-L, highlighting the complementary roles of language and task adapters in this setting. These results suggest that Serbian-language adaptation improves lexical alignment and structural coherence. At the same time, the smaller BERTScore gains, together with its stronger correlation with human adequacy and fluency judgments, indicate modest improvements in semantic consistency during generalization to previously unseen topics and linguistic patterns.

Compared to the task-only adapter (BASE+SrQAA*), the full stack still provides consistent, while smaller, improvements across all metrics, with BERTScore gains suggesting modest but genuine improvement in semantic output quality.

Table 7. Performance of the BASE (L3.2-11B-V) and adapter variants on the out-of-domain **QAsr** Serbian QA dataset.

#	Model / comparison	F1	ROUGE-L F1	BERTScore F1
1	BASE	0.152	0.105	0.848
2	BASE+SrLangA	0.131	0.099	0.827
3	BASE+SrQAA*	0.169	0.120	0.862
4	Full_stack (BASE+SrLangA+SrQAA)	0.182	0.129	0.862
Relative improvements of Full_stack (%)				
	vs. BASE	19.4%	21.9%	1.8%
	vs. BASE+SrLangA	39.1%	29.3%	4.3%
	vs. BASE+SrQAA*	7.5%	7.3%	0.1%

An interesting observation is that the language-adapted configuration (BASE+SrLangA) performs slightly worse than the base model on the QAsr benchmark. This behavior is expected because the SrLangA adapter is trained with a causal language modeling objective, thereby improving linguistic modeling rather than task-specific reasoning. Without additional QA supervision, the model may generate more fluent Serbian text while still lacking the mechanisms required to accurately identify answer spans or produce precise answers. Once the QA adapter is introduced (BASE+SrLangA+SrQAA), the improved linguistic representations provided by the language adapter become beneficial and lead to stronger generalization on out-of-domain questions.

Although the QAsr evaluation set contains only 100 examples, we additionally assess the statistical significance of the reported improvements using paired statistical tests on per-example metric values. Since all model variants are evaluated on the same questions, we apply both paired t-tests and Wilcoxon signed-rank tests for F1, ROUGE-L, and BERTScore. The Full_stack model achieves statistically significant improvements over the BASE model across all metrics. For F1, the improvements are significant under both the paired t-test ($p = 0.0011$) and Wilcoxon signed-rank test ($p = 0.0005$). Similarly, ROUGE-L improvements are statistically significant ($p = 0.0003$ under the paired t-test), while BERTScore improvements are also highly significant ($p < 10^{-6}$). These findings indicate that the observed gains are consistent across evaluation samples and unlikely to arise from random variation despite the relatively small dataset size.

Ablation of adapter composition and training order. Although a full hyperparameter ablation over LoRA configurations was not feasible due to the high computational cost of training the 11B model, the results in Tables 6 and 7 provide a targeted ablation of the proposed two-stage pipeline. Specifically, they isolate the effects of language adaptation alone (BASE+SrLangA), task adaptation alone (BASE+SrQAA*), and their combination in the full stacked model.

Across both in-domain and out-of-domain settings, the results show that language adaptation alone is insufficient for QA and can even degrade performance in the out-of-domain case, while task-specific adaptation provides the primary performance gains. At the same time, the full stacked model consistently achieves the best overall results, particularly on the QAsr dataset, indicating that prior language adaptation provides complementary benefits once followed by QA supervision. This supports the proposed training order.

Taken together, these findings demonstrate that the proposed two-stage adapter architecture is particularly effective for low-resource QA. While task-specific fine-tuning accounts for most of the gains in in-domain evaluation, Serbian-language adaptation plays a critical role in improving robustness and transferability in out-of-domain scenarios. The substantial improvements in QAsr indicate that the language adapter mitigates overfitting to translation artifacts in NQsr and enables the QA adapter to operate on more natural and heterogeneous Serbian input. This confirms the importance of separating language adaptation from task adaptation when building QA systems for morphologically rich, low-resource languages.

6.2. Error Analysis – Cloze-style Grammatical Completion Test

This section provides a detailed examination of the strengths and remaining weaknesses of the Serbian language adapter. The goal is to complement the quantitative results presented earlier with analyses that reveal specific linguistic behaviors that are not captured by aggregate automatic metrics.

To assess grammatical sensitivity more directly, we evaluated the base model and the Serbian language adapter on a Serbian Cloze-style grammatical-completion benchmark comprising 110 items. Each sentence contains one masked token. For each cloze item, we enumerated all candidate completions and computed the sentence-level perplexity of the resulting full sentence for each candidate; the model prediction was taken to be the candidate with the lowest perplexity. The task requires the correct selection of case endings, function words, prepositions, pronouns, and other elements that are central to Serbian.

Table 8. Cloze test accuracy comparison between the L3.2-11B-V model and the Serbian language adapter

Dataset	**Accuracy**		**Improvement**
	Base	**BASE+SrLangA**	**Rel Δ**
Serbian Cloze Eval	37.27%	47.27%	26.80%

The results shown in Table 8 indicate that the Serbian adapter increases accuracy from 37.27 percent to 47.27 percent. This represents an absolute improvement of 10 percentage points. The gain confirms that the adapter strengthens internal representations of Serbian morphology and sentence structure. The examples in Table 9 illustrate representative error types; however, they are not exhaustive. More broadly, the model exhibits recurring challenges typical for morphologically rich languages, including agreement errors (case, gender, number), verb tense formation (especially future constructions),

and context-dependent lexical selection. These findings suggest that, while the language adapter improves fluency and grammaticality, certain fine-grained morphological phenomena remain difficult to model.

Although the Cloze evaluation set is relatively small ($n = 110$), we assess the statistical significance of the observed improvement using McNemar's exact test on paired predictions. This test is appropriate for binary outcomes (correct/incorrect) evaluated on the same examples. The results show that the improvement is highly statistically significant ($p = 9.76 \times 10^{-4}$). Notably, all disagreement cases correspond to improvements of the adapted model over the base model, with no observed degradations. This indicates that the gains are consistent across the evaluation set and are unlikely to arise from random variation.

Table 9. Representative examples from the Serbian cloze evaluation with corresponding error types.

Prompt	**Base**	**BASE+SrLangA**	**Correct form**	**Error type**
Sutra ćemo *Tomorrow we will*	Sutra ćemo putujemo *Tomorrow we will traveling*	Sutra ćemo putovati *Tomorrow we will travel*	putovati *travel*	Incorrect future construction
U rerni se *In the oven, it was*	U rerni se jeo *In the oven, it was eaten*	U rerni se pekao *In the oven, it was being baked*	pekao *being baked*	Incorrect lexical verb choice
Srpska pravoslavna crkva slavi Božić *The Serbian Orthodox Church celebrates Christmas*	Srpska pravoslavna crkva slavi Božić 1. januara. *The Serbian Orthodox Church celebrates Christmas on January 1*	Srpska pravoslavna crkva slavi Božić 7. januara. *The Serbian Orthodox Church celebrates Christmas on January 7*	7. januara *January 7*	Factual error
U restoranu je naručio *In the restaurant, he ordered*	U restoranu je naručio čas *In the restaurant, he ordered an hour*	U restoranu je naručio večeru *In the restaurant, he ordered dinner*	večeru *dinner*	Incorrect lexical choice
On trči vrlo *He runs very*	On trči vrlo brz *He runs very fas**	On trči vrlo brzo *He runs very fast*	brzo *fast*	Adjective–adverb confusion

The examples in Table 9 illustrate representative error types and are not exhaustive. More broadly, the analysis across the dataset shows that most errors fall into a small number of dominant categories. The most frequent are agreement-related errors (case,

gender, and number), followed by incorrect case selection in prepositional constructions and errors in verb agreement and person marking.

A smaller but still notable portion of errors is related to comparative and superlative formation and pronoun selection. In addition, some errors arise from lexical and semantic competition, as well as from idiomatic expressions and gaps in factual knowledge. These patterns reflect the inherent complexity of Serbian as a morphologically rich language and indicate that, while the language adapter improves fluency and grammaticality, certain fine-grained morphological phenomena remain challenging.

6.3. Qualitative Analysis of Generated Answers

To complement the quantitative results presented in Sections 6.1–6.2, we conducted a qualitative analysis of representative answers from the language-only adapter (BASE+ SrLangA) and the full stacked model (BASE+SrLangA+SrQAA). This analysis helps clarify whether the remaining errors are primarily grammatical, semantic, or factual, and provides a more fine-grained interpretation of the automatic metrics.

The examples reveal a clear difference in failure modes between the two configurations. The language-only adapter frequently produces off-topic continuations, semantic drift, and degraded answer structure. Although it often generates more fluent Serbian than the base model, it is not sufficiently aligned with the question-answering task and may fail to remain within the correct semantic space. This observation is consistent with the quantitative findings from Section 6.1, where BASE+SrLangA performs slightly worse than the base model on the out-of-domain QAsr benchmark. In contrast, the Cloze results in Section 6.2 indicate that the same adapter improves morphological and grammatical competence, suggesting that better Serbian language modeling alone is not sufficient for robust question answering.

The full stacked model is substantially better aligned with the question intent and usually produces answers within the correct topical range. However, its remaining errors are mainly factual rather than grammatical. In particular, the model often identifies the correct topic or central fact, but introduces incorrect supporting details, inaccurate causal explanations, or hallucinated claims. This suggests that task-specific supervision improves answer relevance and structural alignment, while factual grounding remains the main unresolved challenge.

Overall, the qualitative evidence indicates that the dominant errors are not primarily related to Serbian morphology or agreement, which are examined separately in Section 6.2, but rather to semantic drift in the language-only configuration and factual hallucinations in the full stacked model. More broadly, this pattern suggests that in morphologically rich low-resource languages, language adaptation and task adaptation play complementary roles: the first primarily improves fluency and linguistic well-formedness, while the second is necessary to align generation with task intent. Representative examples are provided in Appendix B, Table B1.

7. Conclusion

In this paper, we presented a modular and resource-efficient pipeline for adapting the large multimodal L3.2-11B-V model to Serbian text generation and the QA task. The approach

uses a staged 4-bit QLoRA-based PEFT, with SrWaC and STARS for Serbian-language adaptation and NQsr for supervised QA fine-tuning. Training is designed to remain feasible on a single NVIDIA A100 40GB without full-model fine-tuning. We observe improved QA scores on NQsr and QAsr across F1, ROUGE-L, and BERTScore. The language adaptation also improves performance on a Serbian grammatical cloze benchmark that targets morphosyntax and lexical selection. This evidence supports the claim that the method strengthens Serbian linguistic competence and improves downstream generation quality.

Our results support three broader takeaways. First, the modular two-stage design supports composable deployment, since language and task adapters can be stacked and extended to new tasks with low overhead. Second, the fully quantized 4-bit QLoRA setup demonstrates that adapter-based fine-tuning can remain effective for large multimodal architectures under constrained compute resources. Third, the dataset construction pipeline, including entity-aware correction for Serbian NQ translation, provides a reproducible strategy for building supervised datasets in other low-resource languages. Overall, these findings suggest that adapter-based language specialization is a practical approach to improving NLP for low-resource languages without extensive computational infrastructure.

8. Limitations

Despite the promising results, this study has several limitations that should be considered when interpreting the findings.

Evaluation methodology limitations. The evaluation datasets contain hundreds to a few thousand samples, which is sufficient to observe consistent trends but may not fully reflect the variability of real-world Serbian usage. Results are primarily based on automatic metrics, which provide reproducible comparisons but do not capture all qualitative aspects of generation quality. This is evidenced by our correlation analysis of automatic metrics with human annotations, which shows that F1 and ROUGE-L do not align with human judgments of fluency or grammatical correctness, and that even BERTScore only partially reflects human-perceived quality. Furthermore, qualitative analysis of generated outputs reveals error types such as semantic drift and factual hallucinations that are not captured by any of the three metrics. Human evaluation was partially conducted to evaluate the translation quality of the NQsr dataset as a semi-manual annotation of adequacy, fluency, and grammatical correctness on a 100-sample subset. Grammatical competence of the language adapter was evaluated through a cloze-style benchmark of 110 items. However, human evaluation of the fluency of generated model outputs was not conducted and remains for future work.

Training data diversification. While the SrWaC and STARS corpora cover diverse functional styles, they do not include all Serbian domains, such as informal conversational text, social media, or dialectal variation. Performance in such domains may therefore differ from the reported results. Although this study focuses on Serbian, the proposed staged adapter-based framework is not language-specific and may apply to other morphologically rich low-resource languages under constrained computational resources. However, such generalization was not explicitly evaluated in this work.

NQsr dataset translation quality. Since the NQsr dataset is an automated translation of the NQ dataset, it may contain translation artifacts and unnatural phrasing introduced

by NMT and LLM models. Despite a pipeline designed to improve name handling, errors still arise from NER misdetections or incorrect LLM translations, especially in texts with many named entities, where placeholder replacement reduces context for NMT. Our translation quality analysis found that artefacts are limited in absolute terms, with overall adequacy scores remaining high (mean 4.56/5), but fluency and grammatical correctness degrade with increasing NED, introducing unevenly distributed noise into the training data. These issues limit the maximum achievable performance during task adapter fine-tuning.

QAsr reference answer bias. The QAsr evaluation dataset uses GPT-5–generated references, which may reflect stylistic preferences that differ from native Serbian usage. This can introduce bias into semantic metrics such as BERTScore, favoring stylistic similarity to the reference over true answer correctness.

Fine-tuning context length constraints. Fine-tuning was limited to a maximum context length of 2,048 tokens due to computational and stability constraints associated with 4-bit QLoRA. As a result, performance on long-form or document-level inputs was not examined.

Model evaluation scope. In this work, we intentionally evaluate the proposed parameter-efficient adaptation method within a single base model to isolate its effect without introducing confounding factors related to architectural differences, tokenizer design, or pretraining data. As such, the results should be interpreted as demonstrating the effectiveness of the adaptation strategy rather than as a direct comparison with alternative model families. Our approach is therefore complementary to existing encoder-based and multilingual methods, extending Serbian NLP capabilities toward generative reasoning and answer generation. Future work will explore systematic cross-model comparisons under controlled conditions.

9. Future Work

Future work will explore extending the pipeline to tasks such as summarization, reasoning, and multimodal instruction following. Given that the L3.2-11B-V backbone is inherently vision-enabled, combining the Serbian language adapter with image-grounded training opens opportunities for multimodal applications in Serbian. In parallel, we plan to investigate adapter composition and routing strategies to enable simultaneous specialization across multiple Serbian NLP tasks with minimal overhead. Manual evaluation of generated output fluency and factual correctness will also be conducted on larger validation subsets to complement the automatic metrics reported in this work.

References

1. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023)
2. Anthropic: The claude 3 model family: Opus, sonnet, haiku. Tech. rep., Anthropic (2024), https://www.semanticscholar.org/paper/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627
3. Anthropic: Introducing claude 4 (2025), [Online]. Available: https://www.anthropic.com/news/claude-4 (current October 2025)

4. BigScience Workshop, Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al.: BLOOM: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
5. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media (2009), https://books.google.rs/books?id=KGIbfiiP1i4C
6. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-) evaluation of machine translation. In: Callison-Burch, C., Koehn, P., Fordyce, C.S., Monz, C. (eds.) Proceedings of the Second Workshop on Statistical Machine Translation. pp. 136–158. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://aclanthology.org/W07-0718/
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.747/
8. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems 36, 10088–10115 (2023)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018), http://arxiv.org/abs/1810.04805
10. Dodović, M., Tufegdžić, J., Hrvačević, L., Drašković, D.: From documents to answers: A university-focused dbqa system with llms. In: Conference on Information Technology and its Applications. pp. 333–343. Springer (2025)
11. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 10 (2022)
12. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python. Zenodo (2020)
13. Hrvačević, L., Dodović, M., Tufegdžić, J., Drašković, D.: A comprehensive review of large language models for document-based question answering. In: Conference on Information Technology and its Applications. pp. 289–301. Springer (2025)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR 1(2), 3 (2022)
15. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the NLP world. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6282–6293. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.560/
16. Koehn, P., Monz, C.: Manual and automatic evaluation of machine translation between European languages. In: Koehn, P., Monz, C. (eds.) Proceedings on the Workshop on Statistical Machine Translation. pp. 102–121. Association for Computational Linguistics, New York City (Jun 2006), https://aclanthology.org/W06-3114/
17. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics 7, 452–466 (2019), https://aclanthology.org/Q19-1026/
18. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
19. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)

20. Ljubešić, N., Lauc, D.: Bertić–the transformer language model for bosnian, croatian, montenegrin and serbian. arXiv preprint arXiv:2104.09243 (2021)
21. Ljubešić, N., Klubička, F.: bs, hr, sr wac-web corpora of bosnian, croatian and serbian. In: Proceedings of the 9th web as corpus workshop (WaC-9). pp. 29–35 (2014)
22. Marovac, U.A., Avdić, A.R., Milošević, N.L.: A survey of resources and methods for natural language processing of serbian language (2023), https://arxiv.org/abs/2304.05468
23. OpenAI: GPT-4 technical report (2023), https://arxiv.org/abs/2303.08774
24. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: Adapterfusion: Non-destructive task composition for transfer learning. arXiv preprint arXiv:2005.00247 (2020)
25. Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I.: Adapterhub: A framework for adapting transformers. arXiv preprint arXiv:2007.07779 (2020)
26. Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7654–7673 (2020)
27. Samardžić, T., Ljubešić, N., Miličević, M.: Regional linguistic data initiative (ReLDI). In: Piskorski, J., Pivovarova, L., Šnajder, J., Tanev, H., Yangarber, R. (eds.) The 5th Workshop on Balto-Slavic Natural Language Processing. pp. 40–42. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (Sep 2015), https://aclanthology.org/W15-5306/
28. Shuttleworth, R., Andreas, J., Torralba, A., Sharma, P.: Lora vs full fine-tuning: An illusion of equivalence. arxiv 2024. arXiv preprint arXiv:2410.21228 (2024)
29. Singh, S., Romanou, A., Fourrier, C., Adelani, D.I., Ngui, J.G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W.Q., Susanto, Y., et al.: Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 18761–18799 (2025)
30. Snegha, A., Sen, S., Pasi, P.S., Singhania, A., Jyothi, P.: Zero-shot cross-lingual transfer using prefix-based adaptation. In: Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025). pp. 385–396 (2025)
31. Tong, S., Fan, D., Nguyen, J., Brown, E., Zhou, G., Qian, S., Zheng, B., Vallaeys, T., Han, J., Fergus, R., et al.: Beyond language modeling: An exploration of multimodal pretraining. arXiv preprint arXiv:2603.03276 (2026)
32. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
33. Ulčar, M., Robnik-Šikonja, M.: Finest bert and crosloengual bert: less is more in multilingual models. In: International Conference on Text, Speech, and Dialogue. pp. 104–111. Springer (2020)
34. Vīksna, R., Skadina, I.: Multilingual Slavic named entity recognition. In: Babych, B., Kanishcheva, O., Nakov, P., Piskorski, J., Pivovarova, L., Starko, V., Steinberger, J., Yangarber, R., Marcińczuk, M., Pollak, S., Přibáň, P., Robnik-Šikonja, M. (eds.) Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. pp. 93–97. Association for Computational Linguistics, Kiyv, Ukraine (Apr 2021), https://aclanthology.org/2021.bsnlp-1.11/
35. Xuan, W., Yang, R., Qi, H., Zeng, Q., Xiao, Y., Feng, A., Liu, D., Xing, Y., Wang, J., Gao, F., et al.: Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 1513–1532 (2025)
36. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies. pp. 483–498 (2021)
37. Škorić, M., Janković, N.: New textual corpora for serbian language modeling. Infotheca 24 (2024), https://arxiv.org/abs/2405.09250

Appendix

A. LoRA Hyperparameter Selection and Rank Sensitivity

Due to the high computational cost of training the 11B model (individual runs exceeding 20 days), we did not perform an exhaustive hyperparameter search over LoRA configurations. Instead, we initialized the setup based on prior work and conducted a limited number of preliminary experiments to assess sensitivity to key parameters, including the LoRA rank r and scaling factor α. Candidate configurations were evaluated using early training dynamics, specifically the rate of loss decrease and optimization stability during the initial training phase. The final configuration ($r = 16$, $\alpha = 8$) provided the most favorable trade-off between convergence speed, stability, and computational efficiency. The choice of rank in low-rank adaptation is known to be task-dependent. The original LoRA work [14] shows that relatively small ranks can be sufficient for many downstream tasks, but also emphasizes that this behavior should not be expected universally, particularly under significant distribution shift. In particular, when the target task differs substantially from the pretraining distribution (e.g., cross-lingual settings), higher-capacity updates may be required to capture the necessary transformation. Further evidence suggests that LoRA and full fine-tuning produce qualitatively different parameter updates. Recent analysis shows that low-rank adaptation introduces task-specific directions in weight space (referred to as "intruder dimensions"), which are associated with forgetting and can impact generalization [28]. The number and effect of such directions depend on both the rank and scaling parameters, indicating that insufficient rank may limit the expressiveness of the adaptation. This further supports the view that rank selection controls a trade-off between expressivity and stability in parameter-efficient adaptation. At the same time, large-scale empirical studies such as QLoRA [8] show that when low rank is applied broadly across model layers, performance can become relatively insensitive to the exact choice of rank in standard fine-tuning settings. However, these observations are primarily derived from instruction tuning and same-language tasks, where the required adaptation is relatively limited. In contrast, our setting involves continued pretraining for Serbian, which constitutes a substantial cross-lingual distribution shift. In such scenarios, the required transformation of representations is expected to be more complex than in standard downstream tasks. We therefore adopt a moderate-rank configuration ($r = 16$), which provides sufficient adaptation capacity while maintaining the efficiency advantages of parameter-efficient fine-tuning. This choice should be interpreted as a practical compromise rather than an optimal setting, and a systematic investigation of rank sensitivity in cross-lingual adaptation remains an important direction for future work.

B. Additional Qualitative Analysis of Generated Answers

Table B1. Representative examples of generated answers illustrating typical failure modes of the language-only adapter (BASE+SrLangA) and the full stacked model (BASE+SrLangA+SrQAA). Each cell first reports the Serbian text, followed by its English translation. Long outputs are shortened to representative excerpts.

Question	Reference	BASE+SrLangA Output	Full_stack Output	Error pattern by model
Kada je prestonica Rimskog carstva premeštena u Konstantinopolj? *When was the capital of the Roman Empire moved to Constantinople?*	**330. godine; grad je proglašen prestonicom 11. maja 330.** *330 AD; Constantinople was formally inaugurated on May 11, 330.*	**1055. godine, kralj Svetogorskih zemalja, Stefan Nemanja...** *In 1055, the king of the Holy Mountain lands, Stefan Nemanja...* Produces a fully off-topic answer about medieval Serbian history.	**330. godine, Konstancije II je izgradio... Novi Rim...** *In 330, Constantius II built... New Rome...* It identifies the correct year, but introduces incorrect historical details and wrong entities.	***SrLangA:*** *Semantic drift.* ***Full stack:*** *Correct answer with imprecise historical attribution.*
Ko je pobedio u Drugom svetskom ratu, Japan ili Sjedinjene Američke Države? *Who won World War II, Japan or the United States?*	**Sjedinjene Američke Države i saveznici.** *The United States and the Allies.*	**1945. godine, a koji je ujedno i bio prvi predsednik SAD...** *In 1945, who was also the first president of the USA...* The answer quickly turns into an incoherent and off-topic narrative without answering the question.	**1945. godine, SAD i Sovjetski Savez su se dogovorili da Japan kapituliše...** *In 1945, the USA and the Soviet Union agreed that Japan would surrender...* It stays on topic and points to Japan's surrender, but the explanation is historically imprecise.	***SrLangA:*** *Severe semantic drift.* ***Full stack:*** *Largely correct answer with minor historical imprecision.*
Kolika je dužina britanske obale? *What is the length of the British coastline?*	**Približna dužina zavisi od definicije i načina merenja; npr. oko 11 073 milje za Veliku Britaniju ili 7 723 milje za UK.** *Approximate length depends on definition and measurement method; e.g. about 11,073 miles for Great Britain or 7,723 miles for the UK.*	**1. Koju duzinu britanska obala ima? 2. Kako je britanka obalna duzina?...** *1. What length does the British coast have? 2. What is the British coastal length?...* The output degenerates into a sequence of malformed paraphrases of the question without a real answer.	**12.000 milja (19.312 km) britanska obala je najduža obalna linija u svetu.** *12,000 miles (19,312 km), the British coastline is the longest coastline in the world.* The answer is fluent and largely plausible, but includes a hallucinated claim that the British coastline is the longest in the world.	***SrLangA:*** *Fluency degradation / generation collapse.* ***Full stack:*** *Plausible answer with minor factual inaccuracy.*
Ko je prvi započeo izgradnju Panamskog kanala? *Who first started the construction of the Panama Canal?*	**Ferdinand de Leseps i francuski pokušaj iz 1881. godine.** *Ferdinand de Lesseps and the French attempt in 1881.*	**1914. godine, a završen je 1926... Štaviše, Srbija je bilo jedino zemljište u Evropi...** *In 1914, and it was completed in 1926... Moreover, Serbia was the only land in Europe...* The answer contains an incorrect chronology and then drifts into completely unrelated geopolitical content.	**1881. godine, Francuska je počela da izgradi kanal, ali je napustila projekt 1894. nakon što je izgubila rat sa Španijom...** *In 1881, France started building the canal, but abandoned the project in 1894 after losing the war with Spain...* It correctly identifies the French start of construction, but hallucinates the reason for the project's failure.	***SrLangA:*** *Semantic drift.* ***Full stack:*** *Correct core answer with unsupported causal explanation.*

C. NQsr annotation rules

Annotation rules for adequacy (Table C1), fluency (Table C2), and grammar correctness dimensions (Table C3).

Table C1. Adequacy annotation rules. Responds to the question: *What proportion of the source meaning is lost in the translation?*

Score	Label	Description
1	All lost	Almost nothing of the source meaning survives. The translation is unrelated, fabricated, or so distorted that the original meaning is unrecoverable.
2	Majority lost	More than half the meaning is missing or distorted. A reader would have a substantially incorrect or incomplete understanding of the source.
3	Noticeable loss	A meaningful portion of the content is missing or wrong. A reader would have a partial or partially incorrect understanding. The reader gets the general topic but misses or misunderstands something notable.
4	Negligible loss	Almost all meaning is preserved. A reader would have a near-complete understanding, with only a trivial detail imprecise or a named entity in a non-standard form.
5	Nothing lost	The translation preserves all meaning. A reader would gain a complete and correct understanding of the source. No facts, entities, numbers, or claims are missing or wrong.

Table C2. Fluency annotation rules. Responds to the question: *What proportion of the translation sounds unnatural to a native Serbian speaker?*

Score	Label	Description
1	Completely unnatural	Almost nothing sounds like natural Serbian. Pervasive awkward phrasing, unnatural word order, or literal translation throughout - a native speaker would immediately identify this as machine-translated.
2	Mostly unnatural	The sentence feels foreign or stilted throughout. While parseable, it consistently reads like a literal translation rather than natural Serbian.
3	Noticeable unnaturalness	Some parts of the sentence read naturally, others feel stilted or machine-translated. A native speaker would notice issues but could follow the meaning without difficulty.
4	Negligible unnaturalness	One or two isolated phrases feel slightly awkward or non-idiomatic, but the sentence reads naturally overall and only a careful reader would notice.
5	Completely natural	The sentence reads naturally. A native speaker would not notice anything unusual - phrasing, word choice, and register are all idiomatic.

Table C3. Grammar Correctness annotation rules. Responds to the question: *What proportion of the translation contains grammatical errors?*

Score	Label	Description
1	Pervasive errors	The translation is grammatically broken throughout. Case, conjugation, and agreement are systematically wrong and the text is barely parsable as Serbian.
2	Majority erroneous	Grammatical errors appear in most clauses. Reading is significantly disrupted. The reader must actively work to recover meaning.
3	Noticeable errors	Several grammatical errors are present but do not prevent understanding. A reader notices problems but can still parse the meaning without difficulty.
4	Negligible errors	One or two isolated grammatical slips (e.g. a single wrong case ending or minor agreement failure in an otherwise correct translation).
5	No errors	The translation is grammatically correct throughout. Case endings, verb conjugations, and agreement all conform to standard Serbian with no detectable errors.

Adela Ljajić is a Research Associate at the Institute for Artificial Intelligence Research and Development of Serbia. Her research spans machine learning, natural language processing, generative AI, retrieval-augmented generation, and AI evaluation. Much of her work focuses on adapting AI methods to low-resource, multilingual, and data-constrained settings, with an emphasis on reliable outputs and practical evaluation. Her broader interests include trustworthy AI, transfer learning, and reliable AI applications.

Marko Zubac is a junior research assistant at the Institute for Artificial Intelligence Research and Development of Serbia. His research spans machine learning, large language models, knowledge graphs, natural language processing, generative AI and retrieval-augmented generation. His work focuses on improving knowledge extraction and graph-based question answering, as well as adapting and evaluating language models for Serbian and other low-resource language settings. His broader research interests include context engineering, explainable AI, and the development of reliable AI systems.

Gorana Gojić is a Research Associate at the Institute for Artificial Intelligence Research and Development of Serbia, working in the field of natural language processing. Her work spans large language models, topic modeling, and semantic text similarity, with a focus on the Serbian language.

Miloš Košprdić was a Researcher at the Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, where he worked on natural language processing (NLP) and large language models (LLMs), including semantic search, question answering, and Serbian-language NLP. He is also a Senior Associate at the Linguistics Department of Petnica Science Center, where he previously served as Head of the Department. His research interests include claim verification, natural language inference, sentiment analysis, and the development and evaluation of NLP resources and models for Serbian.

Nikola Milošević is currently a Data Science Manager and Science Fellow at Bayer Pharmaceuticals in Berlin, leading the development of scalable NLP platforms, knowledge graphs, information retrieval, and generative AI systems. He is also a Research Fellow at the Institute for Artificial Intelligence of Serbia. Nikola is deeply involved in developing GenAI workflows, agentic systems, knowledge graphs, and information retrieval engines. His work includes applying machine learning to biomedical text mining and combating hallucinations in large language models.

Dragiša Mišković (Senior Member, IEEE) received the Ph.D. degree in telecommunications and signal processing from the Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia, in 2017. He is a Senior Research Associate with the Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, where he is the Head of the Human-Computer Interaction Group. His research interests are robotics, speech recognition and NLP, human-robot interaction, distributed machine learning, and applied artificial intelligence.

Received: December 15, 2025; Accepted: September 29, 2026.