Development of Serbian QA Datasets through Prompt-Based Generation and Human Validation

Jovana Rađenović, Olivera Kitanović, Ranka Stanković, Mihailo Škorić
Faculty of Mining and Geology, University of Belgrade, Serbia
{jovana.radjenovic, olivera.kitanovic, ranka.stankovic, mihailo.skoric}@rgf.bg.ac.rs

Abstract

LLMs capable of answering questions, fulfilling diverse user requests, and functioning as chatbots rely heavily on extensive datasets. However, for the Serbian language, there is a significant lack of high-quality datasets structured in a question-and-answer (QA) format. To address this, we extracted a portion of the *SQuAD-sr* dataset, which, to the best of our knowledge, is the largest QA dataset in Serbian and contains over 87k samples. While this dataset is an incredibly valuable resource, it was translated using an adapted Translate-Align-Retrieve method and contains errors and terminological inaccuracies. In this work, we systematically reviewed and corrected more than 7k samples from the *SQuAD-sr* dataset, significantly improving the dataset's reliability and quality. We call this modified subset of the *SQuAD-sr* dataset, the *SQuAD-sr-md* dataset. The corrections that were made are crucial for training accurate and robust QA models in Serbian, ensuring that AI systems can leverage the full potential of this dataset. We also introduce an additional QA dataset generated from encyclopedia articles, Wikipedia pages, and scientific paper abstracts using LLMs, which contains ~74k samples. We name this dataset the *SerbianQA-Gen*.

Keywords: Question-Answering, Dataset, Large language models

1. Introduction

Recent advancements in Natural Language Processing (NLP) and the evolution of Large Language Models (LLMs) have substantially improved the performance of Question Answering (QA) systems. Despite remarkable progress, current QA systems and LLMs still face notable challenges, especially when dealing with linguistic diversity and low-resource languages. While English remains the dominant and most widely represented language in NLP resources, datasets for languages such as Serbian remain scarce, leaving little to no room for model improvement or reliable evaluation for this language. This scarcity limits both model training (data coverage, domain breadth) and reliable evaluation (task variety, robust metrics). Moreover, linguistic properties of Serbian such as rich inflectional morphology, freer word order, clitic placement, and two alphabets (Cyrillic and Latin) with variant transliteration of named entities, complicate span alignment and answer normalization. Finally, domain and register diversity (news, encyclopedic, conversational, technical) is weakly represented with Serbian history and culture. These factors jointly motivate the creation of carefully corrected extractive data and new, high-quality generative QA resources for Serbian.

In order to fulfill this task, we first conducted a comprehensive review of current Question Answering approaches, models, and datasets to position our work within the broader research landscape. QA datasets span various types, such as:

- extractive, e.g., SQuAD (Rajpurkar et al., 2018), Natural Questions (Kwiatkowski et al., 2019);
- abstractive, e.g., NarrativeQA (Kočiský et al., 2018), MS MARCO (Bajaj et al., 2018);
- knowledge base (KBQA), e.g., WebQuestions (Berant et al., 2013), MetaQA (Zhang et al., 2018);
- conversational (CQA), e.g., CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018);
- open-domain, e.g., TriviaQA (Joshi et al., 2017);
- multi-hop, e.g., HotpotQA (Yang et al., 2018), ComplexWebQuestions (Talmor and Berant, 2018);
- commonsense, e.g., CommonsenseQA (Talmor et al., 2019), Social IQA (Sap et al., 2019);
- multimodal, e.g., TextVQA (Singh et al., 2019);
- domain-specific, e.g., MedQA (Jin et al., 2020);
- Code-Related, e.g., CoNaLa (Yin et al., 2018), CodeSearchNet (Husain et al., 2019);
- Multiple-choice QA (MCQ), e.g., RACE (Lai et al., 2017).

Some datasets may be in multiple categories, but the diversity of these datasets is crucial for advancing AI capabilities.

In extractive Question Answering, the goal is to identify a specific span of text, that provides the best answer for a given question. Using extractive QA provides a clear and objective evaluation of a model's reading comprehension (RC) abilities. The Stanford Question Answering Dataset

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3369–3378
11–16 May 2026.

(SQuAD) v1.1 (Rajpurkar et al., 2016) and its extended version, that introduces unanswerable questions, the *SQuAD v2.0* (Rajpurkar et al., 2018), are widely used datasets for extractive question answering. The datasets consists of questions and answers created by crowdworkers based on English Wikipedia articles. To the best of our knowledge the *SQuAD-sr* (Cvetanović and Tadić, 2024) dataset is the largest Serbian QA resource, with over 87k samples in both Cyrillic and Latin scripts. While this dataset is an incredibly valuable resource for training QA models, it was created by translating the *SQuAD v1.1* dataset using an adapted Translate-Align-Retrieve method and contains errors and terminological inaccuracies. As the original authors concluded, *SQuAD-sr* is of sufficient quality for fine-tuning a QA model in Serbian, in the absence of a manually created and annotated dataset.

Unlike extractive approaches, generative or abstractive QA systems aim to produce natural language answers that paraphrase or summarize the relevant information. This enables broader reasoning, summarization, and handling of questions that cannot be answered by simple span extraction, and better resembles human behavior. Although generative QA datasets have proven highly useful for training models in many languages, such resources are largely unavailable for Serbian, limiting the development of generative QA systems for this low-resource language.

This work introduces an initiative to broaden and enhance Serbian QA resources. We present two key contributions. First, we perform a detailed manual linguistic and semantic correction of a subset of the *SQuAD-sr* dataset to improve its grammatical accuracy and naturalness. Second, using prompt-based generation with LLMs, we automatically create question–answer pairs and construct a new QA dataset. On this dataset we also have started to perform a detailed linguistic and semantic correction. The datasets introduced in this work are publicly available (Rađenović and Stanković, 2026).

The remainder of this paper is organized as follows. Section 2 describes the methodology for refining the *SQuAD-sr* dataset. Section 3 presents the process of generating QA pairs using LLM prompting. Section 4 concludes with future directions.

2. Refinement of the *SQuAD-sr* Dataset

This section focuses on correcting a subset of the original *SQuAD-sr* dataset, which is a translation of the English *SQuAD v1.1*. Since *SQuAD-sr* was generated through an adapted Translate-Align-Retrieve method, the dataset contains numerous inconsistencies and unnatural expressions that require manual revision.

The task of correcting the dataset was done by authors of this paper, philology student volunteers and domain experts. The task was split into two stages. The first stage consisted of manually correcting 7,484 contexts-question pairs, chosen by topic, from the *SQuAD-sr* dataset (Cvetanović and Tadić, 2024). The second stage was aimed at adjusting the answers according to the modified contexts, by using an open source data labeling tool Label Studio (Tkachenko et al., 2020).

2.1. Manual Linguistic Revision

In the first stage, 7,484 contexts-question pairs were chosen by topic from the Latin version of the *SQuAD-sr* dataset. The original JSON dataset was transformed and organized in a spreadsheet format for manual editing. The final selection covered a diverse set of topics, such as computing, football, linguistics, history, and many others.

The main part of the first stage was a manual linguistic revision of the contexts and questions shown in the spreadsheet. During this phase, grammatical errors were corrected, and sentence structure was improved to ensure fluency, grammatical correctness, and overall naturalness in the Serbian language. Special attention was given to maintaining consistency with the norms of the Serbian language and to addressing fine-grained issues, such as replacing English style quotation marks, which were originally present throughout the dataset, with their Serbian equivalents. Whenever such changes did not alter the original meaning or conceptual integrity of the text, foreign words that occasionally appeared in the contexts were removed. In the spreadsheet, the corresponding entries from the English *SQuAD v1.1* dataset (Rajpurkar et al., 2016) were also displayed. This served as a reference to guide the correction process and stay as true to the original as possible. On average, the annotators revised approximately 7 contexts-question pairs per hour.

2.2. Data Alignment and Annotation

The second stage focused on annotating answers in the corrected contexts. First, the revised dataset was adapted into a format compatible with Label Studio and imported into the platform. In addition, the original answers from the *SQuAD v1.1* and the *SQuAD-sr* dataset were also integrated as answer predictions. Even though most of the answers from *SQuAD-sr* were incorrect due to modifications in the contexts, they served as very useful guidance for the annotation process.

The positions of previous answers were highlighted in gray within the contexts. This immediately directed annotators' attention to the approximate

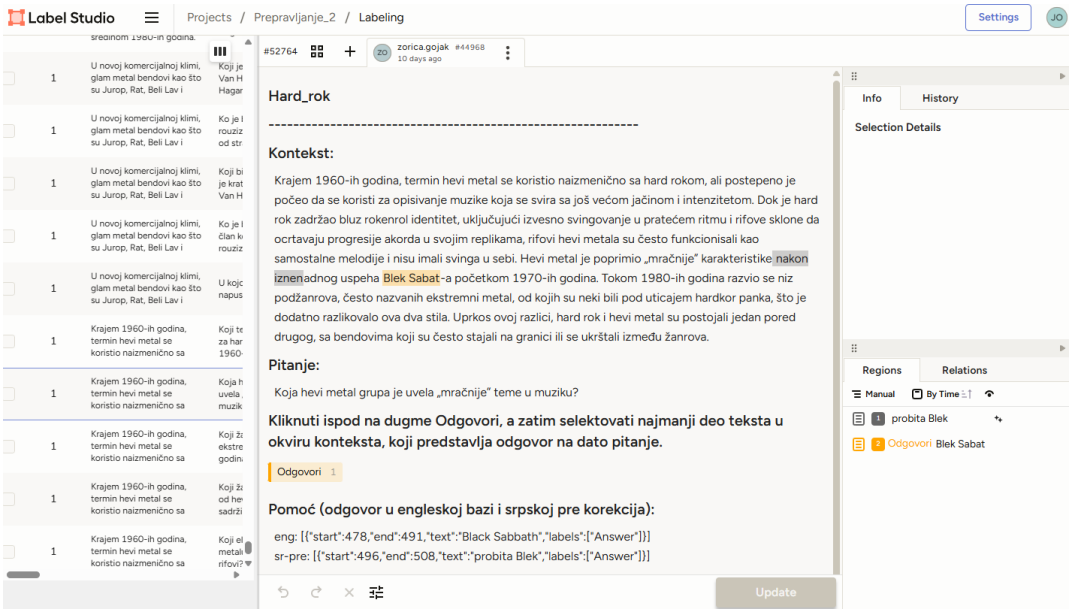


Figure 1: Label Studio interface with the preloaded English and Serbian answers, highlighted previous answer positions in gray, and new answers in orange.

location of the expected answer, significantly reducing the time needed to find the correct answers. Moreover, the parallel display of English and Serbian answers provided contextual cues that helped annotators confirm meaning and precisely locate the correct span. Each question was displayed together with its corresponding context, allowing annotators to efficiently connect the question with the relevant part of the text. An example of this interface setup is shown in Figure 1.

As the result, this stage was considerably faster than the first one, since annotators only needed to select the correct answer span for each question. On average, they were able to complete around 70 answers per hour.

To illustrate the outcome of the corrections made in the process described in subsections 2.1 and 2.2, an example of corrections applied to the title, context, question, and answer, for an article related to phonology, is presented in Table 1. In the example, red highlights indicate parts of the translations that are incorrect or terminologically inaccurate, while blue marks show minor grammatical or morphological adjustments. For the context, only the sentence containing the answer is shown, even though corrections were made in other parts of the context as well. The refinement process resulted in a dataset that we name the *SQuAD-sr-md*.

2.3. Statistical Overview

To better understand the types of modifications made during the revision and annotation process, dataset statistics are presented using boxplots in Figure 2. The plots show the distributions of character length and word count across the Latin version of the *SQuAD-sr* dataset and the *SQuAD-sr-md* dataset. Due to the difference in context size compared to questions and answers, the y-axis is logarithmic to provide a clearer visualization.

It can be observed that the distributions of character lengths and word counts across the two datasets are quite similar, which indicates that effort was made to preserve the structure and content of the original text entries. On average, the main challenges were related to word choice, as some words in the *SQuAD-sr* were unusual or appeared in incorrect case form and verb conjugations. This is to be expected due to the complexity of Serbian grammar. However, Figure 2 shows that there were individual cases where significant modifications were made, reflecting careful adjustments in specific instances. For example, the question that has the maximal character length in *SQuAD-sr* contains one word repeated consecutively many times, although other words are also present. Notably, minimum character lengths and word count for answers correspond to single-digit values, which are the same across all datasets.

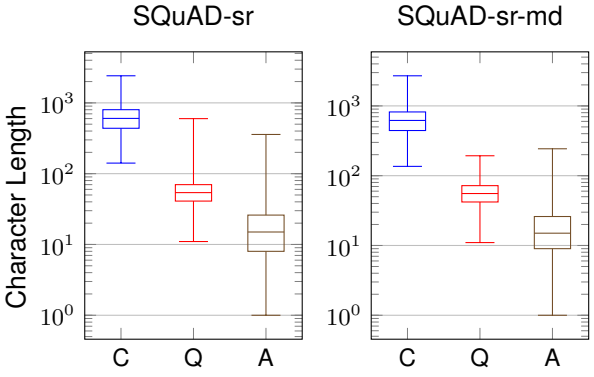
The lengths of article titles from which the contexts were taken are not provided in the figure. However, some entries in the *SQuAD-sr* dataset were in English, while some had their English and Serbian titles combined into one.

2.4. Evaluation

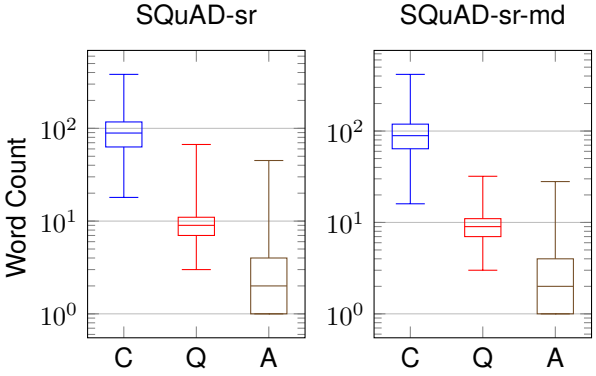
According to the *SQuAD-sr* study (Cvetanović and Tadić, 2024), the highest-performing model was obtained by fine-tuning BERTić (Ljubešić and Lauc,

Element	SQuAD-sr	SQuAD-sr-md
Title	Fonelogija	Fonologija
Context	Za mnoge lingviste, fonetika pripada opisni lingvistici, a fonologija teorijskoj lingvistici, iako je uspostavljanje fonološki sistem jezika neophodno primena teorijskih principa za analizu fonetičkih dokaza.	Za mnoge lingviste fonetika pripada deskriptivnoj lingvistici, a fonologija teorijskoj lingvistici, iako je uspostavljanje fonološkog sistema jezika nužno primena teorijskih principa za analizu fonetskih iskaza.
Question	Kogo se smatra da je fonetika deo koje lingvistike?	Kojem ogranku lingvistike se smatra da pripada fonetika?
Answer	opisni	deskriptivnoj

Table 1: Example of a title, context, question and answer correction.



(a) Character length



(b) Word count

Figure 2: Character length and word count distributions for contexts, questions, and answers (labeled as C, Q, A, respectively, on the horizontal axis) in the *SQuAD-sr* and *SQuAD-sr-md* datasets. The y-axis is shown on a logarithmic scale.

2021). Although the pre-trained model is publicly available (Cvetanović, 2023), any type of direct comparison would not be fair due to the difference in size between our dataset and theirs. Instead, we fine-tune a new set of models on (i) our *SQuAD-sr-md* dataset and (ii) a subset of *SQuAD-sr* that corresponds exactly to the articles and QA pairs that were refined to create the *SQuAD-sr-md* dataset, to assess the impact of our dataset refinement on the learning process. Thus, both training sets are based on identical content in terms of topics, titles, questions, and example count. The only difference lies in the quality of the dataset: our version incorporates structural corrections, improved answer span consistency, and linguistic normalization, while the *SQuAD-sr* subset preserves the original content.

For the fine-tuning process, we adopted the standard Hugging Face procedure for fine-tuning BERT-like models on *SQuAD1.0* (Hugging Face Team, 2023), using parameters defined in Table 2 and used it to fine tune six different Serbian encoder models: BERTić model from the original study, Jerteh-81 and Jerteh-355 models (Škorić, 2024) based on RoBERTa, XLM-R-BERTić (Ljubešić et al., 2024) based on XLM-RoBERTa, as well as XLMali and TeslaXLM models (Škorić and Petalinkar, 2025) based on the same architecture.

Hyperparameter	Value
Learning rate	3×10^{-5}
Batch size	16
Number of epochs	6
Max sequence length	512
Document stride	128

Table 2: Hyperparameters used for fine-tuning the models.

Each fine-tuned model was evaluated on a 100 randomly selected QA pairs from the *SQuAD v1.1* development set translated into Serbian.

2.4.1. Results

For the evaluation, we used Exact Match (EM) and F1 metrics. The procured results are shown in Table 3 as the best scores for each model-dataset pair, considering up to 6 epochs for each model. Although the average improvement in EM is modest (+0.84%), the average increase in F1 (+2.06%) suggests the model produces answers that better match the expected responses when trained on

SQuAD-sr subset

Base model	EM (%)	F1 (%)
BERTić	71	81.29
Jerteh-81	48	57.33
Jerteh-355	64	71.64
XLM-r-BERTić	79	87.94
XLMali	68	78.74
TeslaXLM	**80**	**89.10**
Average	68.83	77.75

SQuAD-sr-md

Base model	EM (%)	F1 (%)
BERTić	71	80.61↓
Jerteh-81	52↑	61.55↑
Jerteh-355	60↓	73.01↑
XLM-r-BERTić	77↓	88.49↑
XLMali	75↑	83.00↑
TeslaXLM	**83**↑	**92.23**↑
Average	69.67↑	79.81↑

Table 3: Results of evaluation done on a 100 translated *SQuAD v1.1* QA pairs, for six base models and their average, for the *SQuAD-sr* subset (top) and the new *SQuAD-sr-md* dataset (bottom).

our refined dataset. In addition to the average improvement it should be noted that the F1 scores also improve for 5 out of 6 tested models with the improvements of up to 4.22% and 3.13% for the best performing model, TeslaXLM. Moreover the average improvement of EM and F1 for the top 3 models is +2.66% and +2.50% respectively.

The qualitative comparison, illustrated with selected examples in Table 4, reveals several recurring improvement patterns:

- **More accurate answers:** Some original answers are factually or contextually inaccurate, e.g., “Džonom Foksom” instead of “Garija Kubiaka”, or “Biblijskih nauka” instead of the precise reference “„Sentenci“ Petra Lombarda”.
- **Complete and unambiguous spans:** Answers missing key details can change interpretation, e.g., “2,50” vs. “2,50 dolara” (currency is essential) or “27.814” vs. “27.814 € godišnje” (time frame is ambiguous without “yearly”).
- **Concise and relevant spans:** Some spans include unnecessary content that does not answer the question, e.g., “juženom Velsu” shortened correctly to “južnom” with our dataset.

Since the models are trained under identical hyperparameters and evaluated on the same test set, the observed improvements can be attributed primarily to differences in the quality of the training data. Even on a relatively small evaluation sample (100 QA pairs), training on our refined dataset yields consistent gains in EM and F1. The qualitative analysis further confirms the importance of careful dataset refinement, where improvements in grammatical accuracy and linguistic coherence can directly translate into measurable performance gains.

3. Generating QA Dataset using LLMs

To speed up the process of creating a new dataset, prompt-based generation with LLMs was used. In contrast to the *SQuAD-sr* dataset, which follows the extractive QA paradigm in which answers are spans from the given context, the new dataset generated through LLM prompting adopts generative, or abstractive approach, producing answers that are paraphrased or synthesized representations of the contextual information.

When deciding which LLM to employ, we considered factors such as speed and output quality. We conducted initial tests using both *GPT-4.1* (OpenAI, 2024) and *GPT-5*. Although *GPT-5* produced slightly higher-quality results, it was significantly slower (at the time of the experiment), so we decided to go with the *GPT-4.1* model. The results reported in the Serbian LLM Benchmark (datatab, 2024) show that GPT-4–series models achieved the highest reported performance (e.g., GPT-4-0125-preview 0.9199, GPT-4o-2024-05-13 0.9196), outperforming other models such as Qwen2-72B-Instruct (0.8425), GPT-3.5-turbo-0125 (0.8245), and Llama3.1-70B-Instruct (0.8185). Although these results were reported approximately a year ago, they indicate that GPT-4–level models provide strong performance for Serbian language tasks, which motivated their use in our experiments. Further experiments with this newer model are planned for the next stage of our work.

For contexts in the dataset, several sources were used including the *PaSaz* corpora (Škorić et al., 2025), the Serbian Wikipedia articles (Wikimedia Foundation, 2025) and the *Sveznanje* encyclopedia[1].

PaSaz (Škorić and Janković, 2024) represents a parallel Serbian-English corpus of doctoral dissertation abstracts. It contains a total of 10,492 parallel units, abstracts, where the vast majority, in addition to that, has parallel titles. For this research, we selected 3,254 Serbian abstracts, with an average of 1,410 characters, or 187 words. We summarized those abstracts using *GPT 4.1*, and the output had approximately 270 characters or 33 words. In the first phase of QA generation, we used the abstracts for context, but 34% of calls were not successful. Than, we used summarized versions as context, and only one out of 3,254 was unsuccessful. On a very small selection of 10 random samples, we

[1] https://sr.wikisource.org/wiki/Sveznanje

Question	SQuAD-sr-based model	Our model
Ko je glavni trener Bronkosa?	Džonom Foksom	Garija Kubiaka
Koliko je iznosila previše velikodušna tantijema koju je Tesla primao?	2,50	2,50 dolara
Kolika je osnovna plata nastavnika, u evrima?	27.814	27.814 € godišnje
Šta je Pol Rouz rekao da je Luter uneo u nemačko mišljenje?	histerični i demonizujući mentalitet	izazvao histerični i demonizujući mentalitet prema Jevrejima
Na šta se odnosila Luterova diploma iz 1509. godine?	Biblijskih nauka	„Sentenci“ Petra Lombarda
Koji je drugi naziv za skladište za ugalj?	lančani ili pužni mehanizam za loženje	bunkera
U kom geografskom delu Velsa se nalazi Aberkinon?	juženom Velsu	južnom
Kada je Luter putovao u Mansfeld dva puta?	1545.	krajem 1545. godine

Table 4: Representative qualitative differences in answers from the evaluation of models trained on the two datasets, with questions listed first.

have verified that summarization produced context that is appropriate for the task.

From Wikipedia pages for selected topics, we extracted 7,998 articles, with an average of 523 characters (76 words) that were summarized to 214 characters and 30 words. So, like in the case of *PaSaz*, we have continued with using summarized versions of articles, for this experiment.

In *Sveznanje*, we have 5,459 articles with an average length of 517 characters (82 words), as well as summarized versions with an average length of 227 characters (32 words). In this case, summarization mainly involved simplification and resolving abbreviations. We are aware that our contexts are shorter than the ones in *SQuAD*, but for the first experiments we wanted to have simpler versions and later enlarge the dataset with larger contexts.

Biblisha (Stanković et al., 2016) contains collection of texts from journals and projects, comprising documents generated from TMX, stored in a Mongo database. For this research we selected 392 abstracts, with average of 824 characters (112 words) and summaries with average of 254 characters (32 words). For summarization and question–answer pair generation, we used the DSPy library, a framework for programming language model pipelines through declarative modules and optimization strategies. DSPy enables the structured composition of LLM-based tasks and facilitates the development of reproducible pipelines for complex NLP workflows. In our setup, DSPy was employed to orchestrate the summarization of source texts and the subsequent generation of QA pairs, allowing systematic experimentation with prompt structures and model configurations. This approach supports modular design and easier adaptation of the pipeline to different datasets and tasks (Khattab et al., 2023).

To automatically generate question–answer pairs from the contexts, we used a carefully designed prompt in Serbian. An example of a prompt that was used is shown in Figure 3. In the continuation of the listing, we provide the prompt in English as well.

```
prompt_sr = f"""
  Iz sledećeg enciklopedijskog članka izdvoji {
     n_pitanja} pitanja
  koji pokrivaju ključne informacije. Za svako
     pitanje daj bar jedan
  odgovor, a mogu i dva. Koristi jasan i
     informativan stil.

  Vrati rezultat u JSON formatu sa poljima:
  [
    {{"question": "...", "answers": [{{"text": "..."
     }}, ... ] }},
    ...
  ]

  Članak:
  '''{clanak}'''
"""

prompt_en = f"""
  From the following encyclopedia article, extract {
     num_questions}
  questions that cover the key information.
  For each question, provide at least one answer (
     two if appropriate).
  Use a clear and informative style.

  Return the result in JSON format with the
     following structure:

  [
    {"question": "...", "answers": [{"text": "..."},
      ...]},
    ...
  ]
"""
```

Figure 3: Prompt for generating the QA dataset.

The prompt instructs the model to generate up to three questions based on key information from the input context and to provide at least one, and optionally two, answers for each question. It should be noted that the prompt formulation did not impose a strict limit on the number of generated questions and answers, but rather served as a guideline, resulting in a small number of outliers with up to eight questions per context, primarily in longer passages.

```
original:
VETERINA, veterinarstvo je  (lat.), nauka koja
     obuhvata odgajivanje stoke, čuvanje, lečenje i
     iskorišćavanje životinja. Veterinar, stručnjak
     za v.; lice koje je svršilo vet. fak. ili vet.
     visoku šk. V. vojni, lice sa svršenim vet. fak.
      i položenim ispitom za rez. oficira, posle slu
     žbe u kadru, ili pitomac mst. voj. koji po svrš
     enom školovanju otslužuje svoj rok; u voj. se
     prima sa činom vet. por. Veterinarski arhiv,
     stručni v. časopis, osn. 1931.; izdaje ga v.
     fak. u Zagrebu. V. konvencija, međunar. [...]
     Jugosl. ih zaključila s Austr., It., Bug., Mađ
     ., Rum., Grč. i Nem. V. nastava, počela
     osnivanjem v. šk. u Lionu (1762.) u Frc.; danas
      se vrši na v. fak. ili v. visokim šk. koje
     imaju skoro sve evr. zemlje; u Jsl. postoji v.
     fak.  [...] v. upotrebu. Gl. zadatak drž. v.
     službe je suzbijanje stočnih zaraza.

summarized:
Veterinarstvo je nauka o odgajivanju, lečenju i
     iskorišćavanju životinja. U Jugoslaviji postoje
      veterinarski fakulteti u Beogradu i Zagrebu.
     Državna veterinarska služba i zakonske mere
     usmerene su na suzbijanje stočnih bolesti.
     Postoje različite veterinarske ustanove, arhivi
     , međunarodne konvencije i proizvodnja cepiva.
```

Figure 4: Original and summarized article from *Sveznanje* enciclopedia.

```
"id": "sveznanje-10026",
"context": "VERN Žil (1828--1905) bio je francuski
     pisac, poznat po avanturističkim romanima sa
     naučnom osnovom, u kojima su predviđeni mnogi
     savremeni pronalasci. Najpoznatija dela su Put
     oko zemlje za 80 dana i Putovanje sa Zemlje na
     Mesec.",
"n_questions": 3,
"qas": [
   {"question_no": 50,
    "question": "Ko je bio Žil Vern?",
    "answers": [
      {"text": "Žil Vern je bio francuski pisac."},
      {"text": "Poznat je po avanturističkim
     romanima sa naučnom osnovom." }],
   },
   {"question_no": 51,
    "question": "Koja su najpoznatija dela Žila
     Verna?",
    "answers": [
      {"text": "Put oko zemlje za 80 dana"},
      {"text": "Putovanje sa Zemlje na Mesec"}],
   },
   {"question_no": 52,
    "question": "Po čemu su romani Žila Verna znač
     ajni?",
    "answers": [
        {"text": "U njegovim romanima su predviđeni
     mnogi savremeni pronalasci."}],
   } ]
```

Figure 5: A dataset entry with generated QA.

In contrast to *SQuAD-sr*, where each question in the training dataset is linked to only one answer, the proposed approach supports multiple answers per question, following the paradigm introduced in (Kočiský et al., 2018). The model is further instructed to produce the output in JSON format, following the *SQuAD* data structure. This approach ensured that the generated QA pairs were contextually relevant and as grammatically correct as possible.

Overall, the resulting dataset comprises of 15,935 contexts, 51,644 questions, and 73,825 answers. We name this dataset the *SerbianQA-Gen*. Figure 5 shows an example from *Sveznanje*, consisting of a context and three generated questions, two with two answers each and one with a single answer.

Generated data has been partially refined and filtered through manual revision, with additional filtering and postediting planned for the future stages. Question–answer pairs were reviewed by annotators to ensure factual correctness, fluency, and naturalness. Any incorrect, ambiguous, or incomplete answers were corrected or replaced, while the dataset structure was preserved for consistency.

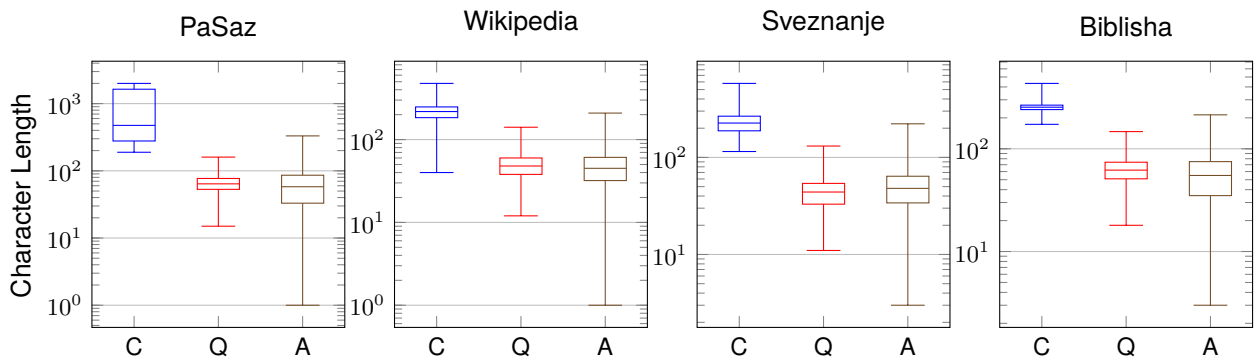
3.1. Statistical Overview

Similar to Section 2.3, where the core statistics were presented for the *SQuAD-sr* and *SQuAD-sr-md* datasets, this section shows the same set of statistics for the newly developed *SerbianQA-Gen* dataset. Since the overall dataset consists of four individual sources, we present the statistics for each of them separately. This comparison provides a clearer understanding of the structural and linguistic characteristics of the data across different sources.

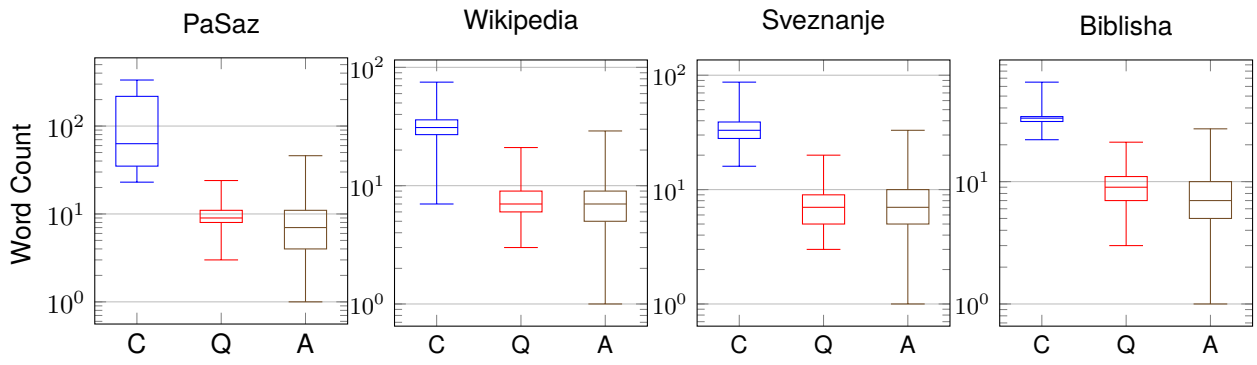
From Figure 6 it can be observed that the datasets differ notably in the distribution of context lengths. The *PaSaz* dataset contains considerably longer contexts, with both the median and the upper quartile significantly higher than in the other datasets. This is also reflected in the average context length in *PaSaz*, which is around 885 characters (103 words). Longer contexts in *PaSaz* are expected, given that the dataset consists of rich, paragraph-level texts. In contrast, the contexts of the *Wikipedia*, *Sveznanje*, and *Biblisha* datasets are considerably shorter and more tightly distributed. Their boxplots show smaller interquartile ranges and lower medians, suggesting that these contexts are generally concise and focused on single facts or concepts. The average context length for these datasets is around 229–254 characters (33 words).

Despite the differences in context length, the distributions of question lengths remain relatively consistent across all datasets. The boxplots indicate similar medians and relatively small variability, suggesting a uniform style of question formulation. This is also reflected in the average lengths of the questions, which range between 44 and 65 characters (7–10 words).

Answer lengths also show comparable distributions across the datasets and typically remain short, with average lengths ranging between 48 and 63



(a) Character length



(b) Word count

Figure 6: Character length and word count distributions for contexts, questions, and answers (labeled as C, Q, A, respectively, on the horizontal axis) across the *PaSaz*, *Wikipedia*, *Sveznanje*, and *Biblisha* QA datasets. The y-axis uses a logarithmic scale.

characters (3–8 words).

Overall, the analysis shows that the datasets display different context length characteristics while maintaining relatively consistent question and answer lengths. The variation in context length across the datasets introduces diversity in the training data. Longer passages in *PaSaz* provide richer contextual information, while shorter contexts from *Wikipedia*, *Sveznanje*, and *Biblisha* emphasize conciseness and factual focus. This diversity can be beneficial for training QA models, as it exposes them to both extended and compact contexts, potentially improving their adaptability across different text types.

4. Conclusion

In this work, we presented the development of two question–answering (QA) datasets for the Serbian language, the *SQuAD-sr-md* and the *SerbianQA-Gen*. The first, *SQuAD-sr-md* dataset, contains around 7k examples and represents a manually refined version of a subset of the original *SQuAD-sr*, which, despite being an exceptional contribution to Serbian QA research, contains numerous inconsistencies resulting from its translation-based generation process using an adapted Translate-Align-Retrieve method. Although, the manual correction process was time-consuming, it was essential for improving the linguistic quality and reliability of the data, ensuring its suitability for training extractive QA models as showcased on 6 base encoder models for Serbian. Performed evaluation also pinpointed TeslaXLM base model as a potential favorite for future trainings of extractive QA models for Serbian.

The second, *SerbianQA-Gen* dataset, consists of approximately 74k automatically generated QA pairs created with the assistance of large language models (LLMs). It is currently undergoing manual revision to enhance grammatical correctness, fluency, and naturalness of the data. In the absence of large-scale generative QA datasets for Serbian, we believe that this resource represents a valuable contribution.

Together, these datasets enable the training and evaluation of language models capable of a deeper understanding of the Serbian language, with its complex morphology and syntax. In future work, we aim to further broaden QA resources for Serbian and use the datasets to fine-tune and improve models, as well as, to make all datasets publicly

available and encourage other researchers to contribute to developing better models for the Serbian language. Future research should investigate the impact of context summarization on the quality of generated QA pairs, particularly whether shortened summaries retain sufficient information to support reliable question generation.

5. Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia: Text Embeddings – Serbian Language Applications–TESLA #7276, Ministry of Science, Technological Development and Innovation #451-03-34/2026-03/, and COST Action GOBLIN (CA23147) "Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs".

6. Bibliographical References

Payal Bajaj, Daniel Campos, Nick Craswell, et al. 2018. Ms marco: A human generated machine reading comprehension dataset. In *30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. ACL.

Eunsol Choi, He He, Mohit Iyyer, et al. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. ACL.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. CodeSearchNet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*.

Di Jin, Eileen Pan, Nassim Oufattole, et al. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the ACL*, pages 1601–1611, Vancouver, Canada. ACL.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, et al. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the ACL*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the ACL*, 7:453–466. Open-domain extractive QA dataset. Accessed: 2025-10-15.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. ACL.

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, Kiyv, Ukraine. ACL.

Nikola Ljubešić, Vít Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024. 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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the ACL*, pages 784–789, Melbourne, Australia. ACL.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. ACL.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the ACL*, 7:249–266.

Maarten Sap, Hannah Rashkin, Derek Chen, et al. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. ACL.

Amanpreet Singh, Vivek Natarjan, Meet Shah, et al. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2016. Keyword-based search on bilingual digital libraries. In *International KEYSTONE Conference on Semantic Keyword-Based Search on Structured Data Sources*, pages 112–123. Springer.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1*, pages 641–651.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1*, pages 4149–4158, Minneapolis, Minnesota. ACL.

Mihailo Škorić and Nikola Janković. 2024. New textual corpora for serbian language modeling. *Infotheca*, 24:71–96.

Zhilin Yang, Peng Qi, Saizheng Zhang, et al. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. ACL.

Pengcheng Yin, Bowen Deng, Edgar Chen, et al. 2018. Learning to mine aligned code and natural language pairs from stack overflow. In *International Conference on Mining Software Repositories*, MSR, pages 476–486. ACM.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, et al. 2018. Variational reasoning for question answering with knowledge graph. In *AAAI*.

Mihailo Škorić. 2024. New language models for serbian. *Infotheca*, 24(1):7–28.

Mihailo Škorić and Saša Petalinkar. 2025. Quality textual corpora and new south slavic language models. In Jasmina Moskovljević Popović and Ranka Stanković, editors, *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig : Thematic Collection of Papers*, volume 1 of *South Slavic Languages in the Digital Environment JuDig*, chapter 19, pages 337–348. University of Belgrade — Faculty of Philology, Belgrade. 19.

7. Language Resource References

Aleksa Cvetanović. 2023. *BERTić-squad-sr-lat*. Accessed: 2026-03-01.

Aleksa Cvetanović and Predrag Tadić. 2024. *Serbian SQuAD Dataset*. Hugging Face. Accessed: 2025-10-15.

datatab. 2024. *Serbian LLM Evaluation Benchmark Dataset*.

Hugging Face Team. 2023. *Question answering: Fine-tuning BERT on SQuAD1.0*. Accessed: 2026-03-01.

OpenAI. 2024. *ChatGPT-4.1: OpenAI Large Language Model*. OpenAI. OpenAI. Accessed via OpenAI API on 2025-10-15. Model version: gpt-4.1.

Pranav Rajpurkar and Robin Jia and Percy Liang. 2018. *SQuAD2.0 The Stanford Question Answering Dataset*. Stanford University. Accessed: 2025-10-15.

Jovana Rađenović and Ranka Stanković. 2026. *Serbian Question-Answering Datasets: TeslaQA*. TESLA Project, Science Fund of the Republic of Serbia #7276. HugginFace. Hugging Face model card; base-model: XLM-R; license: CC-BY-SA-4.0.

Maxim Tkachenko and Mikhail Malyuk and Andrey Holmanyuk and Nikolai Liubimov. 2020. *Label Studio: Data Labeling Software*. Heartex / HumanSignal. Open source software; accessed 2025-10-15.

Wikimedia Foundation. 2025. *Serbian Wikipedia*. Wikimedia Foundation. Wikipedia, the Free Encyclopedia.

Mihailo Škorić and Nikola Janković and jerteh. 2025. *PaSaz: Serbian Text Classification and Summarization Dataset*. Hugging Face. Dataset on Hugging Face, license CC-BY-4.0. Accessed: 2025-10-12.