12th INTERNATIONAL SCIENTIFIC CONFERENCE ON DEFENSIVE TECHNOLOGIES
OTEH 2026
Serbia, Tara, 21 - 24 October 2026

# POSITION BIAS IN MULTILINGUAL LANGUAGE MODELS APPLIED TO SERBIAN

MILOMIR JOVIĆEVIĆ
Ministry of Defence, Belgrade, milomir.jovicevic@mod.gov.rs

SLAVIŠA ILIĆ
University of Defence, Belgrade, ilic.slavisa@gmail.com

MILAN GNJATOVIĆ
University of Criminal Investigation and Police Studies, Belgrade, milan.gnjatovic@kpu.edu.rs

KRISTINA ŽIVANOVIĆ
School of Electrical Engineering, University of Belgrade, Belgrade, zk245024p@student.etf.bg.ac.rs

***Abstract:*** *Despite rapid progress in evaluating large language models, the Lost in the Middle phenomenon — a persistent weakness of Transformer-based models whereby information near the middle of the context window exerts less influence on generation — remains unresolved. This paper examines the phenomenon in Serbian, among the first such studies outside English, evaluating three recent open-weight small language models on a filtered multi-document QA dataset. Each query contains a question and ten documents, only one of which holds the answer; its position is shifted from first to tenth place to measure the resulting change in accuracy. A dedicated evaluation procedure was developed to account for the morphological complexity of Serbian, combining string-based matching with a final LLM-based judgment stage (Claude Haiku) for cases obscured by grammatical case variation. Within this dataset and context length, a Cochran's Q test revealed no statistically significant accuracy degradation for any of the three models as the gold document approached the middle of the context. Larger differences were instead observed between the models: Gemma-4-26B-A4B (92.9%) and Qwen3.6-35B-A3B (91.9%) could not be distinguished statistically ($p = 0.151$), while GPT-OSS-20B (87.3%) was reliably weaker than both ($p < 0.0001$), suggesting that model choice may play a more important role than gold-document position.*

***Keywords:*** *position bias, lost in the middle, information retrieval, language-specific bias, open-weight LLMs*

## 1. INTRODUCTION

Large language models largely represent the most advanced technology in the world of natural language processing. Although very widespread in everyday use, the Transformer architecture, as the basis of most contemporary language models, faces many problems that have not been completely overcome to date. Most of these shortcomings are related to the context window, a component that in many ways negatively affects the accuracy of the model. One of the most well-known disadvantages of language models is referred to as the Lost in the Middle phenomenon. It assumes that the performance of the model is higher when important information is located at the beginning or at the end of the input context, while on the other hand, the ability of the model to find relevant information drops significantly when position of relevant information tends to the middle of the context window [1].

The research on this phenomenon is of great importance primarily due to a significant number of systems based on language models and designed for tasks such as Retrieval-Augmented Generation, information retrieval and multi-document QA. These systems have in common that finding an answer is directly related to the described phenomenon, which significantly affects their adoption in production environments. So far, most of the research dealing with this problem is focused on the English language, while only a few experiments considered other languages. Accordingly, this research considers the Lost in the Middle phenomenon in the context of the Serbian language.

Since there is no narrowly specialized language model for the Serbian, in this research we consider three open-weight general-purpose multilingual language models: Qwen3.6-35B-A3B, Gemma-4-26B-A4B and GPT-OSS-20B. For none of these models has already been examined whether the Lost in the Middle phenomenon has the same correlation with all the languages or whether the position bias reflects differently for different languages. Since a multilingual model shares all of its parameters across languages, language-specific behavior arises from the composition of the training data and from the tokenizer, and can therefore induce specific phenomena that manifest differently for different models and languages. The Lost in the Middle phenomenon is not exclusively related to open-weight models, but is also present in large language models such as GPT-4 and Claude 3 Opus [2,3].

**Corresponding author:** M. Jovićević;
E-mail: milomir.jovicevic@mod.gov.rs
DOI: 10.5937/OTEH26093J

However, this research is aimed at examining the Lost in the Middle phenomenon in language models applied to Serbian, which is a highly inflective and morphologically rich language.

The main contributions of this work are:

(1) This research is one of the first studies on the Lost in the Middle phenomenon in language models applied on Serbian.

(2) An evaluation procedure was developed that accounts for the morphological complexity of the Serbian language.

(3) A comparative evaluation of three current small open-weight models was conducted, showing that Gemma-4-26B-A4B and Qwen3.6-35B-A3B perform comparably in extractive question answering in Serbian, while GPT-OSS-20B is reliably weaker.

## 2. RELATED WORKS

The Lost in the Middle phenomenon was first described in [1], where the performance of language models, plotted against the position of the relevant information, forms a characteristic "U-curve". Later work proposed ways to mitigate it without fine-tuning, such as Multi-scale Positional Encoding [4] and rescaling a single dimension of the hidden states at inference time, which improves results by up to 15% [5]. Other work examined its causes, showing that residual connections significantly affect cumulative attention and thereby contribute to the U-curve [6]. Menschikov et al. showed that position effects differ across languages and pointed out the lack of evaluation for languages with fewer speakers [3]. Among broader long-context benchmarks, RULER showed that models which solve vanilla "Needle In A Haystack" tests still lose accuracy as the context grows [7], and the bilingual LongBench showed that Chinese is similarly affected when the answer must be found among many documents [8].

The single-gold-document design used in this work goes back to [1], where each question is accompanied by several documents, only one of which contains the answer. Later benchmarks extended multi-document question answering in other directions: the "Leave No Document Behind" benchmark, for example, is built so that every document is relevant to the final answer, and only one of its four task types (spotlight locating) places the evidence in a single document [9]. In this work, the distractors are taken from the same topic as the gold document, which is expected to make the task more demanding than when the documents come from different topics.

## 3. DATASET

At the moment of writing this paper, there was no publicly available dataset aimed at the research and evaluation of the Lost in the Middle phenomenon for Serbian language. Generally speaking, the Serbian language lacks a sufficient number of data corpora that will contribute to the development of language models for this language. Therefore a significantly modified version of the SQuAD-sr-md part of the dataset [10] was used. Originally, it was defined as a QA dataset in the Serbian language, suitable for training or fine-tuning models for extracting data from text. The dataset was created by translating texts from foreign languages using the adapted Translate-Align-Retrieve method, while also manual revision of the entire translation process was performed. The dataset originally consists of 7484 questions and 1739 contexts (documents) in which the answer to them is found. The documents vary in length, but for each question it has been determined which document contains the necessary information for the answer.

**Table 1.** Dataset preview for first topic

| Doc. | Question | Answer |
|---|---|---|
| 1 | Koliko akcija postoji u imovini Arsenala? | 62.217 |
| 2 | Koliki procenat akcija u Arsenalu je Kroenke posedovao do 2011. godine? | 62.89 |
| 3 | U kom filmu je prikazan FK Arsenal? | misterija Arsenalovog stadiona |
| 4 | Ko je pobedio Arsenal 2011. godine u Liga kupu? | Birmingem Sitija |
| 5 | U kom takmičenju je Arsenal prvi ušao u finale? | UEFA Lige šampiona |
| 6 | Koja je razlika u plavoj boji gostujuće sportske opreme Arsenala posle 1983.godine? | tamnijom nijansom |
| 7 | Koji brend je bio dobavljač sportske opreme pre 1994. godine? | Adidasa |
| 8 | Iz kojih zajednica dolaze pristalice Arsenala? | različitih društvenih klasa |
| 9 | Koji trener je krajem devedesetih godina prošlog veka doneo uspeh Arsenalu? | Arsena Vengera |
| 10 | Koji bivši igrač Arsenala je postao trener 1986. godine? | Džordža Grejema |

To adopt this dataset for the purpose of this work, the filtering procedure was designed to ensure that the answer to each question appears in exactly one document. It was applied in four steps. First, questions were removed if their answer appeared in any other document belonging to the same topic (case sensitivity was not used), which reduced the number of questions from 7484 to 5960. Second, to account for the rich morphology of Serbian, questions were additionally removed if the stem of any word from the answer occurred in another document of the same topic; this step can only make the filtering stricter. Third, duplicate questions were eliminated, i.e., a small number of repeated questions whose answers differed only in minor variations; after the second and third steps, 5381 questions in 1650 documents remained. Finally, the documents were balanced in length:

documents whose word counts fell outside predefined lower (121) and upper (188) bounds were discarded, so that all documents in a query are of comparable length. Of the 48 topics, only the fifteen that retained at least ten documents within these bounds were kept; from each of them, ten documents were selected at random, and for each document one question was selected at random among the questions that survived filtering. This yields the final set of 150 questions and documents (fifteen topics with ten documents each).

This design ensures a consistent context length across queries: within a topic, all queries contain the same ten documents and differ only in their order, so their length is identical, while differences between topics are negligible. Furthermore, the topics span a wide range of domains (including sports, notable individuals, emotional wellbeing, music, the internet, history, and politics) so the evaluation does not depend on any narrowly specialized subject matter.

Table 1 shows the first topic of the dataset: each of its ten documents is accompanied by one question and the reference answer, which is a verbatim span of that document.

## 4. EXPERIMENTAL SETUP

The main concepts of the experiment are based on testing the Lost in the Middle phenomenon for open weight models that are deployed in a local infrastructure physically separated from the Internet. The goal is to investigate and evaluate the model's ability to find the answer to a given question and investigate whether the location of the document containing the answer directly affects the model's performance, while using modest models in the controlled environment.

### 4.1. Language Models

For the purposes of this work, we used one of the most current open-weight models of large companies whose architecture is based on Mixture-of-Experts technology and which can be deployed on consumer hardware that has 32 GB of VRAM. Three models were selected:

(1) Qwen3.6-35B-A3B, developed by Alibaba Cloud, released on April 15, 2026.

(2) Gemma-4-26B-A4B, developed by Google, released on April 2, 2026.

(3) GPT-OSS-20B, developed by OpenAI, released on August 5, 2025.

All three models were run using the llama.cpp inference framework, version 9859 (commit 4fc4ec554). Gemma-4-26B-A4B and Qwen3.6-35B-A3B were used with Q8_0 quantization; since the Q8_0 weights do not fit entirely within 32 GB of VRAM, the portion of the weights that exceeded the available VRAM was offloaded to system memory, which affects inference speed but not the model's output. GPT-OSS-20B was used as the full model in the MXFP4 format in which its weights were released, without additional quantization. The context window allocated on the server was 100,000 tokens. For Qwen3.6-35B-A3B, the temperature was set to 1.0, top-p to 0.95, and top-k to 20, while thinking mode was enabled. For Gemma-4-26B-A4B, the temperature was set to 1.0, top-p to 0.95, and top-k to 64, while thinking mode was enabled. For GPT-OSS-20B, the temperature was set to 1.0, top-p to 1.0, and top-k was disabled, while the reasoning effort was set to medium. These decoding parameters were selected according to the recommendations of the respective model developers for general-purpose tasks. The maximum output length was not limited, so generation ended at the model's end-of-sequence token. A fixed seed (42) was sent with every request. If a response did not contain the "<odgovor>" tag, or the request failed at the network level, the same request was repeated, up to three attempts in total, and the response of the last attempt was kept. All system and user prompts were written in Serbian using the Latin script.

These three models represent the latest generation models of the world's largest companies that contribute to the development of multilingual open-weight models, in the absence of language models specialized for the Serbian language.

### 4.2. Methodology

Each query contains all ten documents of one topic followed by the question, and exactly one of the documents, the gold document, contains the answer. Every document of a topic takes the role of the gold document in turn, and in ten consecutive queries it is placed at each position from the first to the tenth, while the nine distractors keep the same relative order. With fifteen topics of ten documents each, this gives 15 × 10 × 10 = 1500 queries per model, 150 for each gold position.

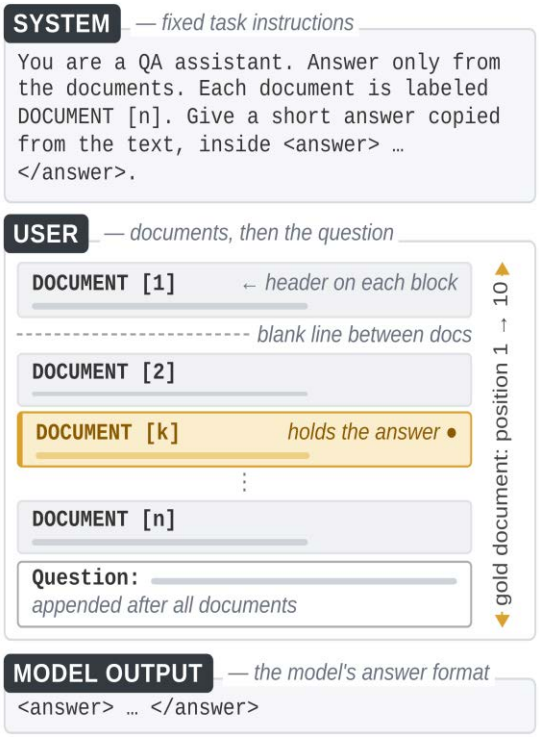



**Figure 1.** Structure of the prompt, with the gold document changing position across queries.

Unlike [1], which used contexts of ten, twenty, and thirty

documents and three gold positions (the first, fifth, and tenth) in its ten-document setting, only ten documents are used here, but the gold document is placed at all ten positions. In [1], accuracy was highest when the relevant document was at the beginning or the end of the context and lowest in the middle, giving the characteristic "U-curve"; comparing the edge positions with the middle ones is therefore the main aim of the design. Figure 1 shows the structure of a query. The system part instructs the model not to answer in its own words, but to return the part of the document that contains the answer inside the <odgovor> tag. This reflects the extractive nature of the dataset, in which every reference answer is a verbatim span of its document, and also makes it possible to check whether the model follows the instruction. For readability, the markers in Figure 1 are shown in English; the actual prompts use the Serbian equivalents (DOKUMENT [n], Pitanje:, <odgovor>).

## 5. RESULTS AND DISCUSSION

As already mentioned, the original dataset gives an answer to each question that represents a text sequence extracted from the document to which it belongs and without any change of word form. If we take into account the nature of the Serbian language, it is clear that it is not possible to use only the "exact match" method to evaluate the results of the model because the language models gave answers of different lengths, even though they all had the same instructions.

Because Serbian is a morphologically different language from English, a model can give a completely correct answer that still differs from the ground truth answer only in grammatical form — for example, a name may appear in a different case ending depending on its role in the sentence. Judging answers with a plain, character-by-character comparison would therefore unfairly mark many correct answers as wrong. To address this, each model's answers were evaluated using three checks, applied in sequence, where each approach's resulting curve is plotted and shown on figures 2, 3 and 4. First, we checked for an exact match between the model's answer and the reference answer. Second, for answers that did not match exactly, we normalized both texts (lowercasing them, removing quotation marks and trailing punctuation, and collapsing extra spaces) and compared them again, which accounts for purely cosmetic differences. Third, for answers still unmatched after the first two checks, we evaluated the remaining pairs using an LLM-as-a-judge approach: the judge model (Claude Haiku 4.5) was given the question, the reference answer, and the model's answer, together with a fixed, explicit set of criteria - grammatical and case-ending differences, synonyms, and minor differences in phrasing were to be treated as correct, while a different, incorrect, or unrelated answer was to be treated as incorrect - and asked to judge whether the two carry the same information. This step was introduced because even a permissive character-level comparison cannot recognize an answer that is phrased in an entirely different way while still conveying the correct information, which becomes increasingly relevant for longer or more paraphrased model answers. The same explicit criteria were applied uniformly to every answer, regardless of which model produced it or where the gold document had been placed in the context, and the judge was given no information about the experimental setup itself; this step was also only applied to the subset of answers left unresolved by the first two checks, rather than being used to re-evaluate every answer. Taken together, this keeps the evaluation consistent across all models and positions and confines the more permissive judgment to genuinely ambiguous cases. The judge was used on 26 September 2026 with its default decoding settings. In total, 779 answers of Gemma-4-26B-A4B, 819 of Qwen3.6-35B-A3B, and 983 of GPT-OSS-20B (2,581 overall) reached this stage, of which 673, 698, and 792 were accepted. To validate this stage, a native speaker of Serbian independently judged a random sample of the judge's decisions, stratified by model (50 per model, 150 in total), including both accepted and rejected answers. The human judgments agreed with the judge in all 150 cases (100%; 95% Wilson confidence interval 97.5–100%), and in all 50 cases for each model (95% Wilson interval 92.9–100% per model).

Because the judge never learns where the gold document was placed, the position analysis was also repeated on the outcomes of the first two stages alone, which the judge does not affect. The position effect was again not significant for any model (Cochran's Q: Gemma-4-26B-A4B $Q = 2.88$, $p = 0.969$; Qwen3.6-35B-A3B $Q = 4.44$, $p = 0.880$; GPT-OSS-20B $Q = 10.86$, $p = 0.286$), so the conclusion about position does not depend on the judge.

In the next three subsections the results per used LLM are provided. To evaluate results, three matching criteria explained earlier in this section, are shown for each evaluated LLM.

### 5.1. Gemma-4-26B-A4B results

Gemma-4-26B-A4B reaches the highest final accuracy of the three models (92.9% after all three evaluation stages) and shows a largely flat curve across all ten gold-document positions, ranging only from about 91% to 95% (though not significantly higher than Qwen3.6-35B-A3B, see Section 6). There is no visible dip when the gold document sits in the middle of the context, which is the pattern the original Lost in the Middle study reported for many models. As shown in Figure 2, Gemma-4-26B-A4B maintains a consistently high and stable accuracy across all gold document positions, with the 95% Wilson confidence intervals overlapping substantially at every position and no pronounced drop near the middle of the context window. A Cochran's Q test across the ten positions finds no statistically significant position effect ($Q = 5.18$, $p = 0.818$), and McNemar tests contrasting the first and the tenth position against the fifth position (the design used by Liu et al. [1]) likewise find no significant differences (92.7% and 93.3% vs. 91.3%; Holm-corrected $p = 0.750$ in both cases). The differences between the edge positions and the fifth position are +1.3 percentage points (95% CI −2.3 to +5.3) and +2.0 points (95% CI −1.3 to +5.8), so a drop toward the middle larger than about 6 points is unlikely, whereas a smaller one cannot be excluded; the position-to-position fluctuations visible in the curve therefore cannot be distinguished from sampling noise at this sample size (n=150 per position).

In 1499 of 1500 queries (99.9%), the model returned a correctly formatted <odgovor> tag on the first attempt, and only a single query remained without it. In 84.5% of cases, the extracted answer appeared verbatim in the gold document. The answers were generally short (median of three words, comparable to the reference answers), with only 7.3% exceeding ten words despite the instruction to return only a few words.

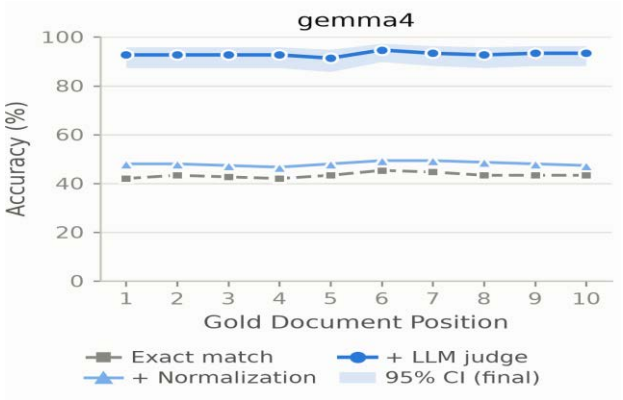



**Figure 2.** Gemma-4-26B-A4B accuracy by gold document position after exact matching, normalization, and LLM judge (final); shaded: 95% Wilson CI of final accuracy (n=150 per position).

## 5.2. Qwen3.6-35B-A3B results

Qwen3.6-35B-A3B performs close to Gemma-4-26B-A4B in absolute terms (91.9% at the final stage) and, like Gemma-4-26B-A4B, shows a largely flat curve across positions, staying within a narrow 89%–94% band with no clear dip in the middle. Its raw exact-match accuracy (42.6%) is noticeably lower than its final accuracy, and the largest jump in the cascade comes from the LLM-as-a-judge stage rather than from normalization alone, indicating that most of its apparent errors are surface-level or paraphrastic mismatches rather than genuine failures to locate the correct information. As shown in Figure 3, Qwen3.6-35B-A3B exhibits accuracy comparable to Gemma-4-26B-A4B, remaining largely stable across gold document positions and likewise showing no strong Lost in the Middle effect. A Cochran's Q test across the ten positions comes closer to conventional significance than for the other two models but still does not reach it ($Q = 15.48$, $p = 0.079$), and McNemar tests contrasting the first and the tenth position against the fifth position likewise find no significant differences (93.3% and 91.3% vs. 91.3%; Holm-corrected $p = 0.906$ and $p = 1.000$; differences +2.0 points, 95% CI −1.8 to +6.2, and 0.0 points, 95% CI −3.3 to +3.3), so the dips visible in the curve cannot be distinguished from sampling noise at this sample size. The model returned a well-formed tag on the first attempt in all 1500 queries (100%). In 79.5% of cases, the extracted answer appeared verbatim in the gold document. It also produced the shortest answers of the three models with only 5.6% exceeding ten words.

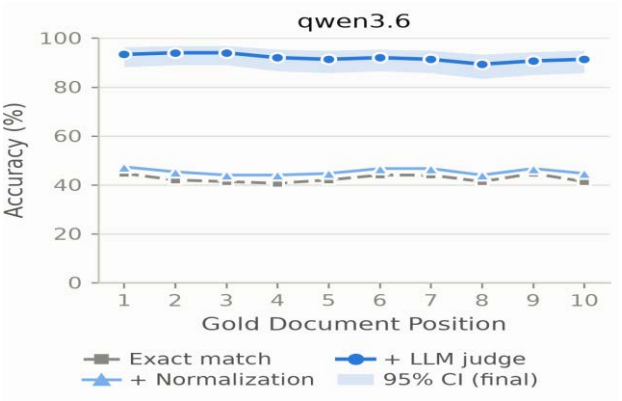



**Figure 3.** Qwen3.6-35B-A3B accuracy by gold document position after exact matching, normalization, and LLM judge (final); shaded: 95% Wilson CI of final accuracy (n=150 per position).

## 5.3. GPT-OSS-20B results

GPT-OSS-20B produces the lowest overall accuracy of the three models (87.3% at the final stage), swinging between roughly 85% and 89% across positions. The lowest point falls at position 5 and the highest at position 1, but this difference is within the range of sampling noise.

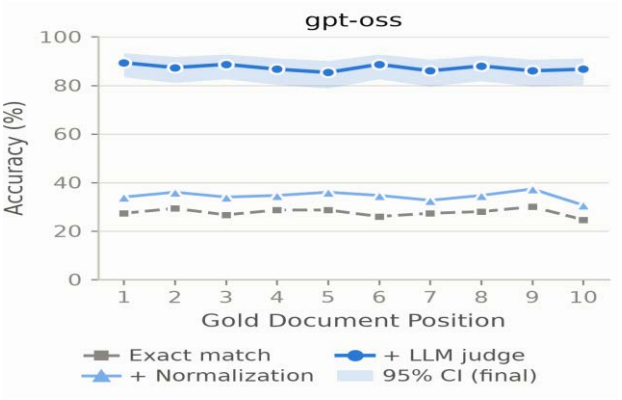



**Figure 4.** GPT-OSS-20B accuracy by gold document position after exact matching, normalization, and LLM judge (final); shaded: 95% Wilson CI of final accuracy (n=150 per position).

As shown in Figure 4, GPT-OSS-20B achieves the lowest overall accuracy among the tested models; however, its position-to-position swing (about 4 percentage points) is no larger than Qwen3.6-35B-A3B's (about 5 percentage points), and a Cochran's Q test across the ten positions finds no statistically significant position effect ($Q = 6.45$, $p = 0.694$) — if anything, a weaker signal than for Qwen3.6-35B-A3B. McNemar tests contrasting the first and the tenth position against the fifth position likewise show no significant differences (89.3% and 86.7% vs. 85.3%; Holm-corrected $p = 0.420$ and $p = 0.727$; differences +4.0 points, 95% CI −1.3 to +9.6, and +1.3 points, 95% CI −2.7 to +5.5). Taken together, these results suggest that GPT-OSS-20B's lower overall accuracy reflects a general weakness at extracting the correct answer rather than a heightened sensitivity to where that answer appears in the context.

A correctly formatted tag was obtained on the first attempt in 1498 of 1500 queries (99.9%), while the remaining two required a second attempt. The extracted answer appeared verbatim in the gold document in 64.8% of cases, noticeably less often than for the other two models. It also produced the longest answers (12.0% exceeded ten words), suggesting a weaker adherence to the instruction to return a short excerpt taken directly from the text.

## 6. CONCLUSION

This paper examined the Lost in the Middle phenomenon in the context of the Serbian language, using a filtered Multi-document QA dataset in which the position of the gold document was varied across ten positions. With awareness of the rich morphology of the Serbian language, a dedicated three-stage evaluation procedure was introduced, in which a large language model (Claude Haiku) acted as an automatic judge in the final stage. The obtained results, based on three current open-weight small language models, did not reveal a statistically significant accuracy degradation when the gold document was positioned near the middle of the context (Cochran's Q, all $p > 0.05$). The 95% confidence intervals of the edge-to-fifth-position differences reach at most about 6 percentage points for Gemma-4-26B-A4B and Qwen3.6-35B-A3B and 10 for GPT-OSS-20B, so a drop of the size Liu et al. [1] reported for GPT-3.5-Turbo in a comparable ten-document setting (about 16 points) is well outside what the data allow, whereas a drop of a few points could have gone unnoticed. This finding is specific to this dataset and these three models and suggests that the Lost in the Middle effect may not be equally manifested across all models and languages; whether this reflects the models, the length and composition of the context, or the Serbian language remains an open question for future work. By contrast, larger differences were observed between the models: GPT-OSS-20B performed reliably worse than both Gemma-4-26B-A4B (Holm-corrected $p = 1.3 \times 10^{-11}$) and Qwen3.6-35B-A3B ($p = 6.8 \times 10^{-9}$), whereas the difference between the latter two did not reach significance ($p = 0.151$), suggesting that model choice may have had a greater impact on retrieval accuracy than the position of the relevant information. The proposed evaluation methodology proved effective in correctly identifying answers despite grammatical case variation confirming that language-specific characteristics must be taken into account when evaluating large language models. For extractive question answering over ten documents in Serbian, in the configurations tested here, Gemma-4-26B-A4B and Qwen3.6-35B-A3B are equally suitable choices, whereas GPT-OSS-20B is less suitable.

### References

[1] LIU, N.F., LIN, K., HEWITT, J., PARANJAPE, A., BEVILACQUA, M., PETRONI, F., LIANG, P.: Lost in the Middle: How Language Models Use Long Contexts, Transactions of the Association for Computational Linguistics 12, pp. 157-173, 2024.

[2] DSOUZA, A., GLAZE, C., SHIN, C., SALA, F.: Evaluating Language Model Context Windows: A 'Working Memory' Test and Inference-time Correction, arXiv:2407.03651, 2024.

[3] MENSCHIKOV, M., KHARITONOV, A., KOTYGA, M., PORVATOV, V., ZHUKOVSKAYA, A., KAGRAMANYAN, D., SHVETSOV, E., BURNAEV, E.: Beyond Early-Token Bias: Model-Specific and Language-Specific Position Effects in Multilingual LLMs, arXiv:2505.16134, 2025.

[4] ZHANG, Z., CHEN, R., LIU, S., YAO, Z., RUWASE, O., CHEN, B., WU, X., WANG, Z.: Found in the Middle: How Language Models Use Long Contexts Better via Plug-and-Play Positional Encoding, Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, Canada, pp. 60755–60775, 2024.

[5] YU, Y., JIANG, H., LUO, X., WU, Q., LIN, C.Y., LI, D., YANG, Y., HUANG, Y., QIU, L.: Mitigate Position Bias in Large Language Models via Scaling a Single Dimension, Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, pp. 6092-6111, 2025.

[6] HERASIMCHYK, H., LABRYGA, R., PRUSINA, T., LAUE, S.: A Structural Theory of Position Bias in Transformers, arXiv:2602.16837, 2026.

[7] HSIEH, C.-P., SUN, S., KRIMAN, S., ACHARYA, S., REKESH, D., JIA, F., GINSBURG, B.: RULER: What's the Real Context Size of Your Long-Context Language Models?, First Conference on Language Modeling (COLM 2024), 2024.

[8] BAI, Y., LV, X., ZHANG, J., LYU, H., TANG, J., HUANG, Z., DU, Z., LIU, X., ZENG, A., HOU, L., DONG, Y., TANG, J., LI, J.: LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, pp. 3119–3137, 2024.

[9] WANG, M., CHEN, L., CHENG, F., LIAO, S., ZHANG, X., WU, B., YU, H., XU, N., ZHANG, L., LUO, R., LI, Y., YANG, M., HUANG, F., LI, Y.: Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, Florida, USA, pp. 5627–5646, 2024.

[10] RAĐENOVIĆ, J., KITANOVIĆ, O., STANKOVIĆ, R., ŠKORIĆ, M.: Development of Serbian QA Datasets through Prompt-Based Generation and Human Validation (te-sla/QuestionAnswering), Hugging Face, https://huggingface.co/datasets/te-sla/QuestionAnswering, 2026.