



# Slovenian Instruction-based Corpus Generation

Gašper Spagnolo, Žiga Klun, Žiga Črv

## Abstract

This project explores the utilization of Large Language Models (LLMs) for conversational agent development in the Slovene language. We investigate various state-of-the-art LLMs suitable for fine-tuning, considering factors such as compatibility with Slovene, computational infrastructure requirements, and model capabilities. Emphasis is placed on understanding the creation process of LLMs and the construction of high-quality conversational datasets. Our methodology involves reviewing datasets and categorizing instructions for training Instruct-based LLMs. We devise a comprehensive plan for data gathering, identifying sources such as med-over.net and slo-tech forums. Crawlers are developed to efficiently collect conversational data, which is organized systematically to facilitate fine-tuning of LLMs. Additionally, we examine pertinent literature, including research on models like MixTal and LLaMa 2, to ascertain key considerations in data preparation. By synthesizing these insights, we prepare a corpus of conversational data tailored for fine-tuning LLMs, ensuring its relevance and quality. Furthermore, we discuss the potential adaptation of an existing LLM using the gathered data, offering insights into the practical application of our methodology. Our findings are consolidated in a final report, providing a comprehensive overview of the process and its implications for developing conversational agents in Slovene using LLMs.

## Keywords

Large Language Models (LLMs), Slovene language, fine-tuning, dataset construction, data gathering, data organization, LLaMa 2, MixTal, Med.Over.Net, Viva.bhc.si, Vizita.si

Advisor: Slavko Žitnik

## Introduction

Introduction se doda čisto na koncu, ko je projekt dokončan.

## Methods

V razvoju konverzijskih agentov, ki delujejo v slovenskem jeziku, se osredotočamo na napredne pristope procesiranja naravnega jezika (NLP). Pomemben del našega raziskovalnega dela predstavlja analiza in prilagoditev dveh sodobnih modelov: LLaMA2 [1] in MixTal [2]; predstavljata temelj za razvoj naših konverzijskih sistemov. Oba modela bomo prilagodili za delovanje v slovenskem jeziku, s posebnim poudarkom na razumevanju in generiranju naravnega, tekočega dialoga.

Za uspešno prilagajanje modelov na slovenski jezik in specifične slovenskega komunikacijskega prostora bomo izvajali "scraping" (pobiranje podatkov) s slovenskih forumov, kot so MedOverNet [3], Vizita [4] in Viva [5]. Ti forumi predstavljajo bogat vir realnih konverzijskih vzorcev in specifičnih izrazov, ki so ključni za razumevanje in ustrezno odzivanje na povpraševanja uporabnikov. Posebno pozornost bomo

namenili selekciji besedil, saj želimo zagotoviti, da bodo odgovori generirani s strani naših konverzijskih agentov temeljili izključno na verodostojnih informacijah, podanih s strani zdravnikov ali specializantov na posameznih področjih. Tak pristop bo zagotovil večjo zanesljivost in kredibilnost odgovorov, kar je še posebej pomembno pri temah, ki so zdravstvene ali strokovne narave.

Prilagajanje modelov, kot sta LLaMA2 in MixTal, za specifične jezikovne in kulturne kontekste predstavlja izziv, a hkrati ponuja obetavne možnosti za razvoj visoko funkcionalnih in prilagodljivih konverzijskih sistemov. Naš cilj je ustvariti modele, ki ne bodo samo razumeli slovenskega jezika, ampak bodo sposobni voditi smiselne in kontekstualno relevantne dialoge, kar bo izboljšalo interakcijo med uporabniki in tehnologijo ter povečalo uporabnost konverzijskih agentov v slovenskem jezikovnem prostoru.

Za nadaljnje izboljšave in prilagoditve naših modelov, bo naša raziskovalna akcija temeljila na zbirki orodij in virov, dostopnih v repozitoriju Brevdev Notebooks [6]. Ta bogat vir vsebuje obsežno zbirko zvezkov Jupyter, ki ponujajo raznolike

metode, tehnike in primere kode, ki so ključni za razumevanje naših modelov.

### Priprava učne množice

V procesu prilagajanja pogovrnih modelov smo izvedli temeljito filtriranje podatkov, pridobljenih s forumov MedOverNet, Viva in Vizita. Filtriranje je bilo izvedeno v več korakih, da bi zagotovili kakovost in relevantnost podatkov za nadaljnje treniranje naših modelov.

1. **Preverjanje obstoja vprašanj in odgovorov:** Najprej smo preverili, ali podatki vključujejo tako vprašanja kot odgovore.
2. **Čiščenje neželenih znakov:** Odstranili smo neželene znake iz vprašanj in odgovorov, da bi izboljšali kakovost teksta za procesiranje.
3. **Preverjanje obiska strani:** Izločili smo vsebine, ki niso prejele vsaj enega ogleda, saj to kaže na nizko interakcijo ali pomembnost vsebine.
4. **Dolžina vsebine:** Zagotovili smo, da vprašanja vsebujejo več kot 19 znakov, odgovori pa več kot 50 znakov, kar prispeva k večji informativnosti dialogov.
5. **Preverjanje anonimnosti odgovora:** Izključili smo odgovore anonimnih uporabnikov, saj dajemo prednost verodostojnejšim in zanesljivejšim informacijam od identificiranih strokovnjakov ali uporabnikov.
6. **Gramatična pravilnost:** Opravili smo preverjanje gramatičnih napak v vprašanjih in odgovorih, da bi zagotovili jezikovno pravilnost. Besedilo brez gramatičnih napak kaže na avtorjevo sposobnost obvladovanja jezika, kar lahko prispeva k večji verodostojnosti odgovora.
7. **Preverjanje prisotnosti neprimernih besed:** Uporabili smo funkcijo za prepoznavanje neprimernih izrazov v vprašanjih in odgovorih, prevedenih v angleščino, da bi zagotovili primernost vsebine za uporabo.

Ta natančna metodologija filtriranja nam omogoča, da ohranjamo visoko kakovost podatkovnega nabora, ki se bo uporabljala za treniranje in prilagajanje našega pogovrnega modela za slovenski jezik.

## Results

Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

## Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

## Acknowledgments

Here you can thank other persons (advisors, colleagues ...) that contributed to the successful completion of your project.

## References

- [1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [2] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Deven-dra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [3] Medovernet - svetovanje. <https://med.over.net/>, 2024. Dostopano: 17.3.2024.
- [4] Vizita.si - za vaše zdravje. <https://vizita.si/>, 2024. Dostopano: 17.3.2024.
- [5] Viva - portal za zdravo življenje. <https://viva.bhc.si/>, 2024. Dostopano: 17.3.2024.
- [6] Brevdev notebooks. <https://github.com/brevdev/notebooks>, 2024. Dostopano: 17.3.2024.