2025 IEEE 19th International Symposium on Applied Computational Intelligence and Informatics (SACI) | 979-8-3315-1547-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/SACI66288.2025.11030141

Multimodal AI for Romanian University Support: An LLM, RAG and Voice Approach

Andrei-Răzvan Joldea*
Department of Computer and Information Technology
Universitatea Politehnica Timișoara
Timișoara, Romania
andrei.joldea@student.upt.ro

Diana Cernăzanu-Glăvan*
Department of Computer and Information Technology
Universitatea Politehnica Timișoara
Timișoara, Romania
diana.cernazanu-glavan@student.upt.ro

Vlad Sârbu*
Department of Computer and Information Technology
Universitatea Politehnica Timișoara
Timișoara, Romania
vlad.sarbu@student.upt.ro

Andrei-Ștefan Bulzan*†
Department of Computer and Information Technology
Universitatea Politehnica Timișoara
Timișoara, Romania
stefan.bulzan@student.upt.ro

***Abstract*—The era of large language models (LLMs) brings forth a new wave of automation to many fields of activity. In this work we employ the AI advancements catalyzed by these LLMs to create a smart university assistant. A chatbot that comes to assist university enrolled students and staff on administrative, legislative and public interest topics. To this end, we develop a platform that combines Large Language Models, Retrieval Augmented Generation, Speech-to-Text and Text-to-Speech technologies to automate accessibility to university-related information. We start from openly available models and resources, adapt and finetune them to our target - Romanian question answering with information retrieval - and then release our solutions publicly at https://github.com/Andrei481/RomanianChatbot.**

***Index Terms*—LLM, RAG, STT, TTS, Romanian, Chatbot, Assistant, Domain-specific fine-tuning**

I. Introduction

In academic institutions, students and staff face numerous administrative, legislative, and informational challenges. The demand for a reliable, efficient, and easily accessible source of information has never been more critical than it is today. Traditional methods of addressing these problems often result in time-consuming processes and inefficiencies. To tackle these issues, we propose the implementation of a smart university assistant — a chatbot designed to provide instant access to essential information.

Chatbots have already emerged as powerful tools in various sectors, offering automated assistance and quick access to information. These systems simulate human conversation and are widely used to enhance customer service, provide technical support, and facilitate information retrieval. Using the latest advancements in large language models (LLMs), our chatbot aims to understand the Romanian language and provide prompt and accurate responses to a wide range of queries. Integrating Retrieval Augmented Generation (RAG), our system enhances its ability to find precise answers by combining the generative power of LLMs with a retrieval mechanism from extensive datasets. Also, by integrating Speech-to-Text (STT) and Text-to-Speech (TTS) technologies, we enhance the accessibility and user experience, making the process more intuitive and effective. Our assistant leverages these AI advancements to specifically address the needs of the University Politehnica of Timișoara (UPT). In developing this chatbot, we start with openly available models and resources, which we adapt and fine-tune for Romanian language question answering and information retrieval. By making our models and resources publicly available on GitHub, we aim to support further research and development, encouraging the adoption of similar solutions in other educational institutions.

Our work builds upon existing solutions like BARKPLUG V.2 [1] at Mississippi State University and SpeakEasy [2], demonstrating the versatility and potential of AI-driven chatbots in educational settings [3]. By fine-tuning open-source models for the Romanian language, we ensure the chatbot meets the specific needs of UPT students and staff, delivering accurate and relevant information.

Key contributions of this work include:

- Development of a platform tailored for Romanian academic institutions, integrating LLMs, RAG, STT, and TTS technologies to provide enhanced information accessibility in Romanian.
- Fine-tuning foundational LLMs specifically for the Romanian language, improving performance in question answering and information retrieval.
- Implementation of voice interaction capabilities in Romanian, enhancing user accessibility.
- Public release of models and resources to support further research and development in Romanian NLP applications.

The paper is structured as follows: Section II reviews related work in AI-driven educational tools. Section III outlines our methodology. Section IV details the implementation, while

[*] Equal contribution among the authors.
[†] Corresponding author.

Section V presents the evaluation results. Finally, Section VI discusses the implications and future research directions.

II. Related Work

A. Large Language Models

Large Language Models (LLMs) have significantly advanced natural language processing (NLP), achieving near-human performance across various tasks [4]. Models like BERT and the GPT series utilize transformer architectures to understand and generate human-like text. Despite significant advancements, LLM development has predominantly focused on the English language, creating a performance gap for other languages. Prominent open models like Llama [5] and Mistral [6] have incorporated Romanian data into their training, though in relatively small proportions. Llama 2 [7] includes approximately 0.03% Romanian data, while Mistral does not focus on multilingual capabilities in their lower-parameter count model offerings.

Recent initiatives aim to enhance Romanian-specific models. *FuLG* [8], a 150-billion-token Romanian corpus, provides a robust foundation for pretraining LLMs. The *OpenLLM-Ro* [9] project further develops foundational and chat LLMs tailored for Romanian, addressing the scarcity of large-scale Romanian datasets. Additionally, *Vorbești Românește?* [10] expands this work by translating a broader range of texts and benchmarks, achieving state-of-the-art performance in Romanian language tasks. The *RoCode* [11] dataset benchmarks code intelligence in Romanian, facilitating the fine-tuning of language models for Romanian code generation.

B. Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) enhances LLMs by integrating external knowledge sources, mitigating issues like hallucination and outdated information [12]. Practical implementations demonstrate RAG's effectiveness in specialized domains. For example, *Ask-EDA* supports electronic design automation tasks using a hybrid RAG approach [13], while *ChatQA* enhances conversational AI by employing optimized retrievers and instruction tuning [14].

Frameworks like *RAFT* train models to focus on relevant documents, improving reasoning and accuracy [15]. The *FACTS* framework offers methodologies for creating secure and efficient RAG-based chatbots, highlighting the importance of fine-tuning in RAG implementations [16].

C. Speech-to-Text and Text-to-Speech

Advancements in Speech-to-Text (STT) have improved performance for low-resource languages like Romanian. Fine-tuning models with techniques such as lateral inhibition has achieved significant reductions in word error rate (WER) [17, 18]. OpenAI's Whisper [19], trained on extensive multilingual data, provides robust speech recognition capabilities. We fine-tuned Whisper for Romanian to enhance its accuracy and integration into our chatbot.

Text-to-Speech (TTS) technologies have also advanced, enabling natural and accurate speech synthesis. We utilized PiperTTS [20] to fine-tune the ”Ro Mihai” medium model, achieving high-quality Romanian speech synthesis that enhances user accessibility through auditory interactions.

Figure 1 illustrates the architecture comprising Romanian LLM Creation, RAG Integration, and STT/TTS modules.

III. Methodology

A. LLM Fine-tuning

We employed two distinct approaches for fine-tuning LLMs for Romanian, as depicted in Figure 1:

1) **Pretraining & Supervised Fine-tuning (PT & FT)**: This approach involved an initial pretraining step using a substantial Romanian corpus (Wikipedia dump and official documents) to adapt the base model to the Romanian language. This was followed by supervised fine-tuning using translated instruction datasets. Models fine-tuned with this approach are denoted as ”RO PT & FT” in our results (e.g., Llama 3 8B RO PT & FT).
2) **Direct Supervised Fine-tuning (Direct FT)**: In this approach, we bypassed the pretraining phase and directly conducted supervised fine-tuning on the base models using specific, Romanian-adapted prompts (see Listing 1) and translated instruction datasets. Models fine-tuned with this method are denoted as ”RO” in our results (e.g., Llama 2 13B RO, Llama 3 8B RO, Mistral 7B v0.2 RO Inst).

Using QLoRA [21], we optimized memory usage, enabling fine-tuning on consumer-grade hardware. Models were evaluated with Romanian-translated benchmarks: ARC [22], MMLU [23], TruthfulQA [24], and HellaSwag [25].

Listing 1: Custom prompt for Llama and Mistral models

```
<|begin_of_text|><|start_header_id|>system<|
    end_header_id|>
Sunteţi un asistent util, respectuos şi onest. Dacă
    o întrebare nu are niciun sens sau nu este
    coerentă din punct de vedere factual, explicaţi
    de ce în loc să răspundeţi la ceva incorect. Dac
    ă nu ştiţi răspunsul la o întrebare, vă rugăm să
     nu împărtăşiţi informaţii false. Trebuie sa ră
    spundeţi doar în limba română.<|eot_id|><|
    start_header_id|>user<|end_header_id|>

{instruction}<|eot_id|><|start_header_id|>assistant
    <|end_header_id|>

{response}
```

B. RAG Algorithm Integration

RAG enhances the chatbot by retrieving relevant information from a curated dataset before generating responses:

1) **Dataset Preparation**: Extracted and digitized 15 UPT-related documents, organizing them into 620 refined text chunks.
2) **Embedding Generation**: Converted text chunks into dense vectors using *all-MiniLM-L6-v2* [26].
3) **Similarity Search**: Utilized FAISS for efficient similarity searches.

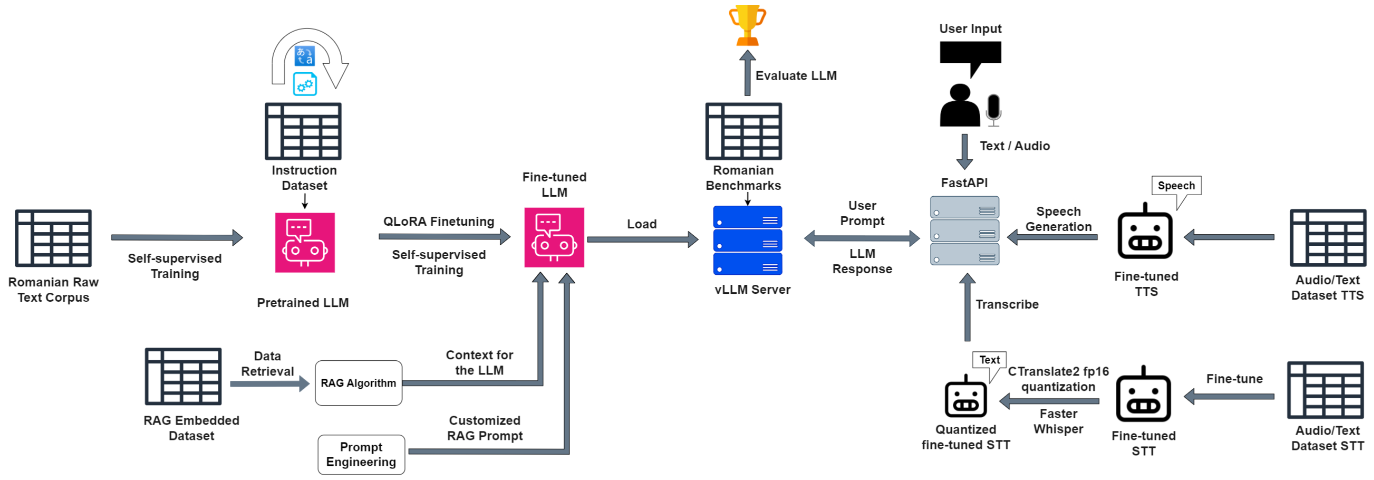

Fig. 1: Architecture Overview

4) **QA Pipeline**: Implemented a QA chain that inserts relevant documents into the LLM prompt, guiding accurate response generation.
5) **Prompt Engineering**: Designed a custom Romanian prompt to optimize RAG performance (see Listing 2).

Listing 2: Customized prompt for RAG

```
{
    <s>[INST]<<SYS>>
    Eşti un asistent util al carui scop este să
        ajute studenţii de la Universitatea
        Politehnica din Timişoara. Dacă o întrebare
        nu are niciun sens sau nu este corectă din
        punct de vedere factual, explică de ce în
        loc să răspunzi la ceva incorect.
    Dacă nu ştii răspunsul la o întrebare, te rog să
         nu împărtăşeşti informaţii false.
    Oferă răspunsuri bazate doar pe informaţiile
        primite în context.
    Dacă în context apar cuvinte în limba engleză, r
        ăspunsul le va cuprinde în limba engleză.
    {context}
    <</SYS>>

    {question}
    [/INST]
}
```

C. Voice Models Training

We developed STT and TTS models to enable voice interactions:

1) Speech-to-Text (STT): Using a fine-tuned Whisper model, we achieved improved transcription accuracy for Romanian:

1) **Dataset**: Combined Mozilla's "Common Voice 17.0" [27] and "Romanian Speech Synthesis Corpus" [28], totaling 39,730 entries.
2) **Fine-tuning**: Configured with a batch size of 10, learning rate of 1×10^{-5}, 1000 warmup steps, and 10,000 max steps.
3) **Optimization**: Applied CTranslate2 FP16 quantization [29] and integrated with FasterWhisper [30] for enhanced performance.

2) Text-to-Speech (TTS): We fine-tuned the "Ro Mihai" model from PiperTTS to generate natural Romanian speech:

1) **Dataset**: Created 241 personal recordings covering diverse phonetic structures.
2) **Fine-tuning**: Configured with a batch size of 16, checkpoint interval of 25 epochs, and up to 5000 epochs, using 32-bit floating point precision.
3) **Optimization**: Monitored training with TensorFlow to ensure high-quality synthesis.

D. Dataset Construction

We curated datasets for each component:

a) LLM Dataset: Translated Vicgalle's Alpaca-GPT4 [31] dataset using DeepL, supplemented with Hakurei's Open-Instruct-v1. The pretraining corpus includes Romanian Wikipedia and the Romanian Constitution, totaling 553MB.

b) RAG Dataset: Digitized 15 UPT-related documents, processed via OCR for scanned texts and direct extraction for PDFs. Organized into 620 text chunks with metadata in JSONL format.

c) Voice Dataset: For STT, combined "Common Voice 17.0" and "Romanian Speech Synthesis Corpus," totaling 39,730 entries. For TTS, created 241 personal audio recordings covering various phonetic structures.

IV. Implementation

We conducted training and fine-tuning on a virtual machine equipped with 128 GB RAM, 16-core processors, and three NVIDIA Tesla T4 GPUs. Despite hardware constraints, techniques like Quantized Low-Rank Adaptation (QLoRA) were used to optimize memory consumption and enable effective model fine-tuning. QLoRA was selected over alternative techniques due to its **quantization** of low-rank parameter updates, which significantly reduces the memory footprint and computational overhead; this enables effective fine-tuning of large models on consumer-grade hardware while maintaining high performance levels.

LLMs: We fine-tuned three models: Llama 2 (7/13B), Llama 3 (8B) [32], and Mistral (7B), including their instruction variants, using QLoRA to reduce memory usage. Key hyperparameters: Batch Size: 2, Gradient Accumulation Steps: 4, Optimizer: AdamW (8-bit), Learning Rate: 2×10^{-4}, Weight Decay: 0.01, Learning Rate Scheduler: Linear, Epochs: 3. Fine-tuning took 4-6 days, with Llama 3 requiring 7 days for a single epoch.

RAG: The RAG algorithm was integrated with the Llama 2 model to enhance information retrieval. This involved dataset embedding with the *all-MiniLM-L6-v2* model and similarity searches using FAISS library. The QA chain included LLM response generation based on retrieved information, using a ”stuff” chain-type for document insertion into the LLM's prompt.

STT: The Whisper model was fine-tuned using Seq2SeqTrainer. Key hyperparameters: Batch Size per Device: 10, Learning Rate: 1×10^{-5}, Warmup Steps: 1000, Max Steps: 10000, Evaluation Metric: Word Error Rate (WER).

TTS: For the VITS model, a multilingual training approach was used. Key hyperparameters: Batch Size: 16, Checkpoint Interval: 25 epochs, Maximum Epochs: 5000, Precision: 32-bit floating point.

System Integration: The application combined backend (Express.js, FastAPI) and frontend (React Native) technologies for seamless integration of models and the RAG pipeline. MongoDB Atlas was used for scalable data storage. Two FastAPI servers were configured: one for batched inference using vLLM and another for managing voice models and the RAG pipeline.

Deployment and Scalability: The Node.js server was deployed on Google Cloud Platform, using a Compute Engine VM Instance. The infrastructure supported secure user interactions with JWT tokens, efficient model inference, and scalable database management for user and conversation data.

V. Results

A. Evaluation Metrics

a) LLM Evaluation: Evaluated fine-tuned models using Romanian-translated benchmarks: ARC, MMLU, HellaSwag, and TruthfulQA. Metrics scored answers based on exact matches or similarity scores using BlackKakapo's *stsb-xlm-r-multilingual-ro* [34].

b) RAG Evaluation: GPT-4 compared RAG-generated answers to source document excerpts, assigning similarity scores from 0 to 100 to measure accuracy.

c) STT Evaluation: Used Word Error Rate (WER):

$$\text{WER} = \frac{S + D + I}{N}$$

where S = substitutions, D = deletions, I = insertions, and N = total words.

d) TTS Evaluation: Conducted qualitative experiments to gauge the similarity and naturalness of synthesized speech compared to the target voice, but no quantitative assessment. Therefore, no numerical results are available for our TTS model.

B. Experimental Results

a) LLM Performance: Table 1 shows that Romanian (RO) variants consistently outperform their base models across benchmarks, indicating enhanced reasoning, language understanding, and factual accuracy. Most notably, the Llama 3 8B RO fine-tuned model significantly surpasses its model of origin, achieving the best overall results on 3 of the tested benchmarks without having had any prior Romanian language pretraining.

Table 1: Performance of Fine-tuned LLMs on Romanian Translations of Standard Benchmarks (Accuracy %)

Model	**ARC (RO)**	**MMLU (RO)**	**HS (RO)**	**TQA (RO)**
Llama 2 13B Chat	20.09%	19.42%	19.12%	28.39%
Llama 2 13B RO	34.18%	30.82%	29.61%	65.01%
Llama 3 8B	13.75%	15.87%	11.51%	48.56%
Llama 3 8B RO	**39.05**%	**34.91**%	**33.78**%	65.42%
Llama 3 8B RO PT & FT	34.93%	29.35%	27.26%	**71.83**%
Mistral 7B v0.2 Inst	7.85%	7.12%	3.70%	70.56%
Mistral 7B v0.2 RO Inst	18.03%	18.57%	19.98%	67.07%

b) RAG Performance: Table 2 demonstrates that the custom RAG prompt achieved an average similarity score of 81, significantly higher than the default prompt's score of 50. This improvement underscores the effectiveness of tailored prompt engineering in enhancing the chatbot's ability to retrieve and utilize relevant information accurately. By customizing the prompts to better align with the specific context and language nuances of Romanian, the chatbot delivers more precise and contextually appropriate responses.

Table 2: RAG Algorithm Accuracy Evaluation for Llama 2 Models

Prompt	**Average Similarity Score**	**Wrong Answers** (< 20)	**Correct Answers** (> 70)
Custom Prompt for RAG	81	5	35
Default Prompt for Llama 2	50	10	29

c) STT Performance: Table 3 shows that fine-tuning Whisper for Romanian significantly reduced WER from 29.15% to 24.6%, demonstrating improved transcription accuracy. This enhancement ensures that spoken inputs are more accurately converted to text, something we have found in our qualitative testing to clearly increase the reliability of interactions with the chatbot.

Table 3: Whisper Models' Accuracy Evaluation

Model-size	**Raw WER**	**Transcription WER**	**Inference Time**
small-244M	41.26%	48.08%	0.89 s
Fine-tuned small-244M (ours)	**36.05%**	**43.36%**	**1.77 s**
medium-769M	29.15%	38.31%	2.34 s
Fine-tuned medium-769M (ours)	**25.02%**	**33.9%**	**4.4 s**
Quantized fine-tuned medium-769M (ours)	**24.6%**	**33.7%**	**0.83 s**
large-1550M	22.42%	32.83%	2.48 s

VI. Conclusion

This study introduces Romanian language datasets for pretraining and instruction fine-tuning, significantly contributing to future Romanian NLP research. Evaluation using established benchmarks demonstrated that Romanian variants consistently outperformed their base models, enhancing reasoning, language understanding, and factual accuracy.

Integrating Retrieval Augmented Generation (RAG) improved information retrieval capabilities, with the custom RAG model achieving a 62% higher similarity score than the default prompt model. Additionally, integrating advanced Text-to-Speech (TTS) and Speech-to-Text (STT) solutions facilitated seamless user interactions.

All development and fine-tuning were accomplished using consumer-grade hardware with limited resources. Increased computational resources would allow further scaling both across larger datasets and more complex, higher parameter count models (1 A100 80GB gpu would allow QLoRA fine-tuning of a Llama 3 70B), broadening its potential applicability to more diverse and computationally demanding tasks.

We release our models and resources publicly on GitHub to support open research.

References

[1] Neupane, S., Hossain, E., Keith, J., Tripathi, H., Ghiasi, F., Golilarz, N. A., …{&} Rahimi, S. (2024). From Questions to Insightful Answers: Building an Informed Chatbot for University Resources. arXiv preprint arXiv:2405.08120.

[2] Jeon, H., Ramachandran, R., Ploerer, V., Diekmann, Y., {&} Bagga, M. (2023). SpeakEasy: A Conversational Intelligence Chatbot for Enhancing College Students' Communication Skills. arXiv preprint arXiv:2310.14891.

[3] Hsain, A., {&} Housni, H. E. (2024). Large language model-powered chatbots for internationalizing student support in higher education. arXiv preprint arXiv:2403.14702.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... {&} Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[5] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., …{&} Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[6] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., …{&} Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

[7] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., …{&} Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[8] Bădoiu, V. A., Dumitru, M. V., Gherghescu, A. M., Agache, A., {&} Raiciu, C. (2024). FuLG: 150B Romanian Corpus for Language Model Pretraining. arXiv preprint arXiv:2407.13657.

[9] Masala, M., Ilie-Ablachim, D. C., Corlatescu, D., Zavelca, M., Leordeanu, M., Velicu, H., …{&} Rebedea, T. (2024). OpenLLM-Ro–Technical Report on Open-source Romanian LLMs trained starting from Llama 2. arXiv preprint arXiv:2405.07703.

[10] Masala, M., Ilie-Ablachim, D. C., Dima, A., Corlatescu, D., Zavelca, M., Olaru, O., …{&} Rebedea, T. (2024). "Vorbeşti Româneşte?" A Recipe to Train Powerful Romanian LLMs with English Instructions. arXiv preprint arXiv:2406.18266.

[11] Cosma, A., Iordache, B., {&} Rosso, P. (2024). RoCode: A Dataset for Measuring Code Intelligence from Problem Definitions in Romanian. arXiv preprint arXiv:2402.13222.

[12] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., . . . {&} Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

[13] Shi, L., Kazda, M., Sears, B., Shropshire, N., {&} Puri, R. (2024). Ask-EDA: A Design Assistant Empowered by LLM, Hybrid RAG and Abbreviation De-hallucination. arXiv preprint arXiv:2406.06575.

[14] Liu, Z., Ping, W., Roy, R., Xu, P., Shoeybi, M., {&} Catanzaro, B. (2024). Chatqa: Building gpt-4 level conversational qa models. arXiv preprint arXiv:2401.10225.

[15] Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., {&} Gonzalez, J. E. (2024). Raft: Adapting language model to domain specific rag. arXiv preprint arXiv:2403.10131.

[16] Akkiraju, R., Xu, A., Bora, D., Yu, T., An, L., Seth, V., . . . {&} Boitano, J. (2024). FACTS About Building Retrieval Augmented Generation-based Chatbots. arXiv preprint arXiv:2407.07858.

[17] Avram, A. M., Smădu, R. A., Păiș, V., Cercel, D. C., Ion, R., {&} Tufiș, D. (2023, July). Towards Improving the Performance of Pre-Trained Speech Models for Low-Resource Languages Through Lateral Inhibition. In 2023 46th International Conference on Telecommunications and Signal Processing (TSP) (pp. 234-237). IEEE.

[18] Avram, A. M., Vasile, P. A. I. S., {&} Tufis, D. (2020, October). Towards a romanian end-to-end automatic speech recognition based on deepspeech2. In Proc. Rom. Acad. Ser. A (Vol. 21, pp. 395-402).

[19] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., {&} Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR.

[20] rhasspy/piper: A fast, local neural text to speech system. (2023, November 14). Retrieved October 4, 2024, from GitHub website: https://github.com/rhasspy/piper

[21] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems, 36.

[22] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., . . . {&} Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457.

[23] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., . . . {&} Steinhardt, J. (2020). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

[24] Lin, S., Hilton, J., {&} Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.

[25] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., {&} Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence?. arXiv preprint arXiv:1905.07830.

[26] sentence-transformers/all-MiniLM-L6-v2 · Hugging Face. (2024, January 4). Retrieved October 4, 2024, from Huggingface.co website: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[27] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., . . . {&} Weber, G. (2019). Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670.

[28] Stan, A., Yamagishi, J., King, S., {&} Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. Speech Communication, 53(3), 442-450.

[29] OpenNMT/CTranslate2: Fast inference engine for Transformer models. (2024, September 9). Retrieved October 4, 2024, from GitHub website: https://github.com/OpenNMT/CTranslate2

[30] SYSTRAN/faster-whisper: Faster Whisper transcription with CTranslate2. (2024, July). Retrieved October 4, 2024, from GitHub website: https://github.com/SYSTRAN/faster-whisper

[31] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... {&} Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 3(6), 7.

[32] Meta, A. I. (2024). Introducing meta llama 3: The most capable openly available llm to date. Meta AI.

[33] vLLM, ”Easy, Fast, and Cheap LLM Serving with PagedAttention,” GitHub Repository, 2023. [Online]. Available: https://github.com/vllm-project/vllm.

[34] BlackKakapo/stsb-xlm-r-multilingual-ro · Hugging Face. (2024). Retrieved October 4, 2024, from Huggingface.co website: https://huggingface.co/BlackKakapo/stsb-xlm-r-multilingual-ro