NeuralMind is offering 4 positions for the summer internship 2023-2024 with the following subjects:

Program 1: Artificial Intelligence Chatbot Enhancement
Program 2: Deployment of a Neural Search Engine

These projects will be developed in a home-office environment with a mentor to supervise his/her work and the student will work with a team of about five developers.

The minimum period of the internship is three months, expected to start December 1st, but can be flexible depending on the academic schedule.

The stipend for working 30 hours per week is R$ 2.500,00 per month plus other benefits.

The application can be done by sending an email to talentos@neuralmind.ai with the subject "Summer Internship 2023"

Please enclose your CV and Academic Records.

Join us in shaping the future of AI-driven products and projects. We eagerly anticipate the unique insights and expertise you'll bring to our team.

Roberto Lotufo
CTO, NeuralMind

Summer Internship 2023-2024 at NeuralMind

Title: Artificial Intelligence Chatbot Enhancement
Duration: December 1st 2023 - February 29th 2024
Objective: Refinement of the AI Chatbot Assistant through Contextual Learning
Modality: home-office or at NeuralMind's office at Unicamp Tech Park
Areas of Interest: Natural Language Processing, MLOps

About the Project:
Our endeavor is directed towards enhancing the capabilities of NeuralMind's NeuralSearchX Chatbot. As the corporate world becomes more interconnected, the essence of "Query-based Summarization" has found its niche, especially with the emergence of Large Language Models like OpenAI's GPT series. The potential now lies in harnessing these models to construct chatbots that provide concise, accurate responses to intricate questions regarding formal documents such as regulations, legislations, guidelines, and other multifaceted texts.

NeuralMind is on a mission to optimize its AI chatbot which is built upon a semantic search engine. We aim to improve the fidelity of our chatbot's responses, enhance our document transformation protocols for indexing by NeuralSearchX, fine-tune prompt engineering, and devise advanced metrics for chatbot performance assessment. This internship is a golden opportunity for individuals to immerse themselves in state-of-the-art natural language processing systems.

Duration Details: 12 weeks in total (Christmas and New Year holidays excluded)

Internship Tasks & Activities:
1. Large Language Models (LLMs) and Retrieval Augmented Generation (RAG).
2. MultiDoc Question-Answer Systems.
3. NeuralMind's NeuralSearchX architecture.
4. Deployment of APIs in cloud environments.
5. Monitoring and optimizing chatbot performance.
6. Final report.

activ/week	1	2	3	4	5	6	7	8	9	10	11	12
1 LLM/RAG	X	X										
2 MultiDocQA		X	X									
3 NeuralSearchX			X	X	X	X						
4 Deployment						X	X	X	X			
5 Monitoring								X	X	X	X	
6 Final Report												X

References:

- Tom Brown, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

- T. Almeida et al. "NeuralSearchX: Serving a Multi-billion-parameter Reranker for Multilingual Metasearch at a Low Cost", DesIRes 2022 Conference, 2022: https://desires.dei.unipd.it/2022/papers/paper-05.pdf

- Lewis, P., Liu, Y., Goyal, N., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.

Summer Internship 2023-2024 at NeuralMind

Title: Deployment of a Neural Search Engine
Period: December 1st 2023 - February 29th 2024
Objective: Optimize the deployment of a deep learning-based search engine
Modality: home-office
Areas of Interest: Natural Language Processing, Information Retrieval, Deployment of Deep Neural Models and Optimization

Summary: Pretrained language models such as BERT are widely used by the top performing academic and industrial search engines (including Google Search and MS Bing). This project aims at deploying and optimizing a state-of-the-art search engine for various types of document search. The project also comprises an exploration of modern compression techniques such as distillation to reduce query latency, which will eventually allow us to run the system on CPUs rather than GPUs. To learn more about our system and the expected activities, check out our paper: "NeuralSearchX: Serving a Multi-billion-parameter Reranker for Multilingual Metasearch at a Low Cost", presented at DesIRes 2022 Conference in August 2022:
https://desires.dei.unipd.it/2022/papers/paper-05.pdf

Schedule: 12 weeks (excluding Christmas and New Year)

List of Activities:

1. Search engines based on deep learning.
2. Writing document indexing scripts.
3. Deploy an open source search engine based on high-performance pretrained models.
4. Learn about modern quantization and model compression techniques.
5. Model optimization methods for low latency inference.
6. Report writing

activ/week	1	2	3	4	5	6	7	8	9	10	11	12
1	X											
2		X	X	X								
3				X	X	X						
4						X	X	X				
5								X	X	X	X	
6											X	X

Resources:

Text Ranking with Transformers:

- Document Ranking with a Pretrained Sequence-to-Sequence Model: https://arxiv.org/pdf/2003.06713.pdf

Model Compression:

- Distillation: https://arxiv.org/pdf/1503.02531.pdf

Quantization:

- Q8BERT: Quantized 8Bit BERT

Serverless architectures:

- AWS: https://aws.amazon.com/serverless/
- Google Cloud https://cloud.google.com/serverless
- Serverless BERT: https://towardsdatascience.com/serverless-bert-with-huggingface-and-aws-lambda-625193c6cc04