**LLM developer**

Your proposal will be declined if you use AI to create your cover letter, profile or resume. You can use it to clean up your original responses. Please respond in your own words.

Experienced developer to design, develop and deploy LLM solutions.

**Preferred tech stack:**

1. Perplexity.ai as the SaaS provider of LLM’s. We don’t how they implement their LLM's. It is quite good with high accuracy, very low hallucination, very fast and with real time data from internet access. We want to copy them.
2. LLAMA3.1 70B and 405B are quite good and the performance of LLAMA3.1 7B may be improved by Distillation of Knowledge or Mixture of Experts
3. Anthropic excels in safe and ethical AI deployment, which is paramount in healthcare applications where data sensitivity is high.
4. Mistral 7B is robust in handling nuanced language processing tasks for more precise interpretations of complex medical terminology and patient interactions.

**Tasks:**

1. Iterative rounds of Supervised Fine Tuning (SFT) of specific tasks and datasets.
2. Direct Preference Optimization (DPO) to refine model responses from human evaluation/responses.
3. Distillation of knowledge to smaller models
4. Set up Mixture of Experts architecture
5. Long running conversations
6. Real time automatic labelling of health care data
7. Real time vectorization, indexing and updates of datasets
8. Prompt engineering
9. Complement responses with real time data from the internet

**Applications:**

1. Voice models to do speech to text (ASR) and text to speech (TTS) tasks. 2-way voice i.e. voice 🡨> text. Automatic translation English 🡨> language among those supported (user selects).
   1. We will also use diarization to identify multiple speakers. LLAMA3.1 has native support for these features.
   2. Another application: user speaks/types --> AI converts to formal style --> speech using native US speaker's voice
2. Summarization 🡪 speech (TTS)
3. Prompts applied to circumscribed data
4. API calls
5. Convert technical terminology to simple language
6. AI agents and chatbots

Detailed requirements and process flows will be provided

You must deploy as soon as a feature is developed so that we can test and provide feedback as early as possible.

1. You are certifying that you have the experience and skills to implement the full features requested. Please quote your price to complete the task.
2. You agree that the timeline agreed to will be a hard date except for delays caused by us. What is your delivery time?

Please make sure you understand and agree that this is an all or nothing request. Failure to meet development timelines will result in cancellation of the contract without any payment.

1. Submit resume with details of experience.
2. What is your availability - days, times (UTC or PST) of the day.
3. What is your proficiency level in spoken and written US English?

Dataset information for the messaging application:

1. Background information:
   1. We are developing an application that uses Zulip messaging server ([www.zulip.com](http://www.zulip.com) ) – similar to Slack – to facilitate communication between medical services providers and patients and other interested parties.
   2. The data is created during the interaction between the doctor and the patient and populated with other inputs as they arrive.
   3. The data is either technical (input by the doctor and their colleagues) or normal conversational (input by the patient and other interested parties).
   4. Technical data can be converted to normal conversational data on demand. This can be a programmatic setting based on the parties viewing the data or on demand whe required.
2. There is no previous data available. The data must be vectorized, indexed and labelled as it is added to Zulip.
3. John Snow Labs has tools with better performance in healthcare than many of those available. Can be used as Mixture of Experts.
   1. NLP open-source text processing library <https://www.johnsnowlabs.com/spark-nlp/>
   2. Healthcare LLM <https://www.johnsnowlabs.com/healthcare-llm/>
   3. Model tuning and validation <https://www.johnsnowlabs.com/generative-ai-lab/>
4. Look at the leaderboard - <https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard>
5. <https://huggingface.co/jiviai/medX_v1> (but link is broken) seems to be better than John Snow Labs (<https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v7.0> ) as of the last update of the leaderboard.
6. Another strong performer is <https://huggingface.co/bongbongs/NewMes-v10.2> although the performance seems to get worse with the newer versions compared to v8.3

Other datasets for chatbot application – FAQ’s, knowledge base articles, blogs, website, presentations. Mostly static data that will be edited and added to over time.

Traditional performance metrics - accuracy, precision, recall, and f1-score. Natural language requires - robustness, bias, accuracy, toxicity, fairness, safety, efficiency, clinical relevance, security, representation, disinformation, political orientation, sensitivity, factuality, legal concerns, and vulnerabilities.

Integration between the LLM solutions and the Zulip messaging server will be via webhooks. Text data (and speech to text data) in the Zulip messages may have to be simultaneously written to a vector database for real-time vectorization, indexing and labelling. We are not processing image data currently.

Healthcare terminologies and data formats - Spark NLP open-source text processing library <https://www.johnsnowlabs.com/spark-nlp/>