Keyword Parser Report

The Grit City

Roadblocks to Scaling the Keyword Parser

1. Pricing:

- The use of APIs, particularly advanced models like llama-3.1-70b, can incur significant costs, especially when scaling up the number of requests.

2. Request Limits:

- Currently, the system is limited to processing 100 requests, which can be a significant bottleneck when dealing with larger datasets or multiple users. This limitation restricts the scalability of the system and requires optimization or increased API allowances to handle higher volumes.

3. Speed:

- The speed of keyword extraction and job search processes can be impacted by the complexity of the tasks and the responsiveness of the APIs. As the system scales, the processing time may increase, leading to slower response times for end-users.

Evaluated Alternatives for Keyword Extraction

Using NER Annotator for spaCy

Pros:

1. **Speed:** The NER model training process in spaCy can be relatively fast, especially for smaller datasets or fewer iterations.
2. **Customization:** Allows for the creation of a model tailored to specific use cases, making it more flexible in identifying custom entities.

Cons:

1. **Time-Consuming Training Process:** The requirement to manually train the model with annotated data is time-intensive, especially when dealing with a large dataset.
2. **Accuracy Issues:** The model's accuracy depends heavily on the quality and quantity of the training data. Without sufficient and well-annotated data, the model may not perform well, leading to inaccurate entity recognition.
3. **Resource-Intensive:** Training a custom NER model can be resource-intensive in terms of computational power and time, particularly when fine-tuning for high accuracy.
4. **Maintenance:** Continuous retraining may be required to maintain the model's performance, especially as new data or entity types emerge.

Current Approach

Model: Ollama llama-3.1-70b

Pros:

- **Advanced Language Understanding**: The model's large size and sophisticated architecture contribute to high accuracy and comprehension of complex text inputs.
- **High Quality Output:** It provides well-structured and detailed JSON responses, which enhances the quality of extracted key fields.
- **Flexibility:** Capable of handling various types of input text and extracting a wide range of attributes effectively.