GRCResponder

Brianna Steier, Liam Gass, Elijah Tavares, Cael Howard,
June Kim, Rish Sharma, Angel Li

accenture

UCI
Donald Bren
School of Information
& Computer Sciences

Table of Contents

Search Engine Updates

- Current Focus: Planning web scraper and defining high-level functionality

- Weekly Goal: Define key requirements for web scraper

Web Dev Updates

- Created Lo-Fi Wireframes
- Begin app setup
  - React w/ Typescript
  - FastAPI
- Setting up CI/CD pipeline with Github Actions

LLM Updates

- Began research and planning

Clarifications Requested

Project Scope & Target Audience

- **Key Questions:**
  - Is our solution limited to these specific cases (PG&E, SCE, SDGE) or is it intended as a general solution for similar legal proceedings?
  - Will this be a tailor-made consulting solution exclusively for these companies, or can it evolve into a repeatable offering?
  - Utilize only documents from these cases?

Clarifications Requested

Document Handling & Data Flow

- **Key Questions:**
  - In the case proceedings, there are Request documents—are these the questions the LLM is expected to answer?
  - Do users upload their own case documents directly for analysis, or is data ingestion managed via an existing system?

Clarifications Requested

Data Management, Confidentiality & User Base

- **Key Questions:**
  - Are these documents confidential, and what are the implications for data security and compliance?
  - What is the anticipated number of users, and how will that impact system scalability?

Clarifications Requested

Response Evaluation & Consistency:

- **Key Questions:**
  - How do we evaluate the quality of the LLM's responses? (Which metrics or benchmarks should be used?)
  - What does "consistency with historical responses" mean in this context, and how will we measure it?

The Data Challenge

- **Context:** PG&E's 2023 General Rate Case
  - 534 proceedings/motions
  - Multiple documents per proceeding
- **Key Point:**
  - Massive volume of legal text requiring extensive storage and processing
  - One case alone takes up significant local storage

Initial Lllama 2 Eval

- **Task:** Analyze complex legal documents
- **Model Options:** Llama 2 available in 7B, 13B, and 70B
- **Decision Factors:**
  - **Quality vs. Efficiency Trade-off**
  - 7B: Lightweight but insufficient for nuanced legal language
  - 70B: Superior quality but extremely resource-intensive
- **Optimal Choice:**
  - **13B Model:**
    - Balances strong performance with lower resource demands
    - Proven in benchmarks to handle complex, domain-specific tasks
    - Similar to Chat-GPT 4o

VRAM requirements for pure GPU inference for 4-Bit quantized Llama models

LLaMA Model	Model size	Minimum VRAM Requirement	Recommended GPU Examples
Llama 2 / Llama 3.1	8B	6GB	RTX 3060, RTX 4060, GTX 1660, 2060, AMD 5700 XT, RTX 3050
Llama 2	13B	10GB	AMD 6900 XT, RTX 2060 12GB, 3060 12GB, RTX 4070, RTX 3080, A2000
LLaMA	33B	20GB	RTX 3080 20GB, RTX 4000 Ada, A4500, A5000, 3090, 4090, 6000, Tesla V100, Tesla P40
Llama 2 / Llama 3.1	70B	40GB	A100 40GB, 2×3090, 2×4090, A40, RTX A6000, 8000
Llama 3.1	405B	232GB	10×3090, 10×4090, 6xA100 40GB, 3xH100 80GB

https://www.hardware-corner.net/guides/computer-to-run-llama-ai-model/

Limitations of Local & UCI Servers

- **Local Machines:**
  - Limited storage capacity
  - Insufficient GPU acceleration for large-scale fine-tuning
- **UCI Servers:**
  - Pricing & specifications geared toward basic virtual hosting
  - CPU-only infrastructure (no GPUs)
  - Limited RAM & disk space
- **Conclusion:**
  - Neither local nor UCI resources can meet the
  - compute and storage needs for this workload

Service	Current Rates	Rate Type
Virtual Server Hosting Service (Base system: 1 CPU core, 2 GB RAM, 50 GB Disk)/month	$13.50/base system/month	Recharge
Virtual Server Hosting Service (Large system: 2 CPU core, 4 GB RAM, 50 GB Disk)/month	$27.00/large system/month	Recharge
Additional CPU & RAM (1 additional CPU core and 2GB additional RAM increments)/month	Additional $13.50/1 additional CPU core and 2GB additional RAM/month	Recharge
Data Center Data Backup & Recovery	$0.055/1GB/month	Recharge
Data Center Storage (allocated in 100GB increments)/month	$5.70/100GB/month	Recharge

https://www.oit.uci.edu/services/infrastructure/virtual-server-hosting/#tabs%7C2||tabs|2

Proposed Strategy

Phase 1:

- Migrate data (PG&E documents) to Azure Blob Storage
- Set up high-performance GPU-enabled VM instances

Phase 2:

- Implement RAG with the Llama 2 13B model for legal analysis on Azure
- Optimize inference pipelines for rapid, scalable processing

Phase 3:

- Validate outputs and integrate with downstream analytics tools
- Monitor performance and adjust resource allocation as needed

Proposed Resources

- **Storage:**
  - We will need to download and preprocess case documents for all the cases locally.
  - Around 60-70 GB required for caching and staging of data + model storage.
- **Compute:**
  - Tuning a 13B Llama 2 model may require 1 GPU node for roughly 15 GPU-hours.
  - Eventually once we want to use the model, we will create Inference service with ~1 GPU node plus supporting CPU (~8vCPUs) and memory (~12GB per node).

Questions?