

Hewlett Packard Enterprise



From Chatbots to Interfaces: Diversifying the Application of Large Language Models for Enhanced Usability pap117

Jonathan Sparks, Pierre Carrier, and Gallig Renaud May 8, 2024

Motivations

- Workshop assistant
 - Create a chatbot to help with background and off-topic questions.
- Engineering Q&A assistant
 - Create a chatbot to help internal engineers to search performance documentation.
- Environment assistant
 - Assistant for HPC schedulers and configurations.

Training & development assistant

Environment assistant Performance Q&A assistant

Potential of LLMs

Overview

Large language models have revolutionized artificial intelligence, particularly natural language processing. Trained on extensive corpora, they can generate coherent, contextually relevant outputs that mimic human text, opening new avenues for human-computer interaction.

Versatility

The versatility of LLMs enables their application across diverse domains, from powering customer service chatbots to facilitating advanced analytics in data science. LLMs contribute to automating tasks requiring greater understanding, thus improving problem-solving and decision-making processes.

User Experience Enhancements

By democratizing access to complex systems, LLMs foster ease of use and enhance user experience. Their applications in specialized environments help provide real-time assistance to developers, improve information retrieval from technical documentation, and simplify the interaction with complete systems and batch schedulers.

Retrieval-Augmented Generation – RAG

- Benefits of RAG
 - The computational and financial costs of retraining foundational models are high. RAG is a more cost-effective approach to introducing new data to the LLM. It makes generative artificial intelligence (generative AI) technology more broadly accessible and usable.
- RAG
 - Documents of many types can be passed into the LLM context window , enabling interactive chat or Q&A assistants.

RAG : Illustration



The ftn command is a driver that invokes the Cray Fortran Compiler when the Cray Compiling Environment is loaded,

```
ftn [-A module_name[,module_name]...]',
metadata={'doc_id': 'ad1cbfdf-f5bc-4a1e-bbf9-e165a15c85b0',
'page': 9,
'source': '../tutorial/corpus/PE/HPE_a00119423en_us_HPE Cray Fortran Reference Manual (13.0) (S-3901).pdf'})]
```

5

Commercial vs. Open Source Models

- Providers: OpenAI, Aleph Alpha
 - GPT-4, GPT-3.5, luminous
 - Tokens == \$\$\$
 - API access-tokens
- Providers: Ollama, Llama.cpp
 - LLama2, Mistral, Mixtral MoE, Phi-2
 - No API access-token required





Model	Input tokens	Output tokens
gpt-4-turbo-2024-04-09	\$10.00 / 1M tokens	\$30.00 / 1M tokens
Luminous-extended	\$9.59 / 1M tokens ¹	

LLM Observability

LLM Observability Challenge

• GSM8K¹ dataset

- Challenge: How do we compare different LLMs using this type of dataset?
- Methods
 - Embedding Distance: measure semantic similarity (or dissimilarity) between a prediction and a reference label string
 - LLM comparator

LLM Correctness Validation Framework



GSM8K Results



Foundational model LLMs

Model	Score
GPT4	72%
GPT3.5	58%
Luminous-extended	0%
Llama2	38%
Mixtral MoE	74%
Phi-2 ²	50%

HPE Cray EX 2000 platform, Nvidia A100 HPE benchmark system, Nvidia H100, A100



LLM Observability w/o RAG

Private models





HPE Cray EX 2000 platform, Nvidia A100 HPE benchmark system, Nvidia H100, A100

12

HPE Slurm w/o RAG



HPE Cray EX 2000 platform, Nvidia A100 HPE benchmark system, Nvidia H100, A100

LLM Score Summary

HPE DE Scores

Model	Score	RAG Score	Delta
Phi-2	68	68	0%
LLama2	74	74	0%
Mistral	79	90	+11%
Mixtral MoE	84	95	0%

100 80 60 40 20 0 Phi-1 Llama2 Mistral Moe Score Rag Score Trend

HE DE Scores

HPE Slurm Scores

Model	Score	RAG Score	Delta
Phi-2	11	11	0%
LLama2	44	56	+12%
Mistral	77	89	+12%
Mixtral MoE	100	100	0%



HE Slurm Scores

Use Cases

Use Cases: Three archetypes

Programming Environment Assistance

LLMs provide developers with real-time assistance, streamlining application building and execution by offering information about programming environments and HPC coding practices.

Internal Document Search

Using LLMs, the paper investigates the improvements in retrieving relevant information from extensive technical documents, enhancing operation efficiency.

Interface to complex systems

LLMs are explored as interfaces for system batch schedulers like Slurm or PBS. By accepting natural test reports, they replace the need for domain-specific languages, making it easier for users to interact with and access these systems.

Use Case Infrastructure

- We are running in local mode. No external services are running.
- The popularity of projects like llama and/or llama.cpp underscores the demand to run LLMs locally, on your infrastructure
- This has at least two significant benefits:
 - Privacy: Your data is not sent to a third party, and it is not subject to the terms of service of a commercial service
 - Cost: There is no inference fee, which is essential for token-intensive applications (e.g., long-running simulations, summarization)

Note

• As previously described, we must be careful about choosing the LLM and services used as we will deal with proprietary information. In some cases, we can use publicly available materials, but when we use an HPE data set, we can conditionally include source materials.

UC₁ – HPE DE assistant

- Motivation: Create a chatbot to help users understand the complexity of Cray PE. This "bot" will be focused on the Cray PE materials and should be able to answer questions on Cray compilers and general HPC programming techniques.
- Environment
 - Use an open-source LLM engine
 - Use private LLM models
 - Use Retrieval-Augmented Generation (RAG) model
 - HPE Development Environment corpus
 - public sources (web)
 - internal training materials (static documents)
 - Standalone chatbot
- Investigation
 - Use of RAG in augmenting results
 - Methods to reduce LLM hallucinations
 - The effect on LLM results with different corpus data and split algorithms
 - Methods to test results

Different Model Selection

Foundation Models



Different RAG methods

RAG methods



Q&A Chat Session



21

UC₂ – HPE Performance assistant

- Motivation: Create a chatbot to help users navigate the internal performance team's corpus of documentation. This chatbot (like UC₁) will be focused on the performance team's web documentation and should be able to answer questions about various posted documentation.
- Environment
 - Use an open-source LLM engine
 - Use private LLM models
 - Use Retrieval-Augmented Generation (RAG) model
 - Performance ingest corpus
 - Internal training materials are hosted on various platforms (Jira/email/confluence).
 - Internal and external data sources
 - Standalone chatbot
- Investigation
 - Different LLM models and the effect on answers
 - Methods to reduce LLM hallucinations
 - The effect on LLM results with different data

Performance Engineer Chat

Enter your Confluence information

https://rndwiki-pro.its.hpecorp.net confluence username jonathan.sparks@hpe.com jonathan.sparks@hpe.com confluence space @ HPCAppsPerfEngr confluence api key @ location http://16.16.184.81:11434 wodel mixtral:8x7b Submit	paste the confluence URL	
confluence username jonathan.sparks@hpe.com confluence space IPCAppsPerfEngr confluence api key confluence api key location http://16.16.184.81:11434 LLM model mixtral:8x7b Submit	https://rndwiki-pro.its.hpecor	rp.net
jonathan.sparks@hpe.com ② Confluence space ③ HPCAppsPerfEngr ③ Confluence api key ④ Confluence api key ③ Confluence api key ④ Confluence api key ⑥ Confluence api key ⑥ Confluence api key ⑥ Confluence api key ⑧ Confluenc	confluence username	0
confluence space ⑦ HPCAppsPerfEngr ⑦ confluence api key ⑦ Location http://16.16.184.81:11434 LLM model mixtral:8x7b Submit	jonathan.sparks@hpe.com	
HPCAppsPerfEngr confluence api key Confluence api key Conflue	confluence space	(?)
confluence api key ⑦ Location LLM model mixtral:8x7b Submit	HPCAppsPerfEngr	
Iocation LLM model mixtral:8x7b Submit	confluence api key	(?)
location http://16.16.184.81:11434 LLM model mixtral:8x7b Submit		o
http://16.16.184.81:11434 LLM model mixtral:8x7b Submit	location	
LLM model mixtral:8x7b ~ Submit	http://16.16.184.81:11434	~
mixtral:8x7b 🗸		
Submit	LLM model	
	LLM model mixtral:8x7b	~

HPE Performance Engineering Confluence Q&A

Ask me a question

what is flux?

Get Answer

Flux is a new workload manager that Lawrence Livermore National Laboratory (LLNL) is developing. It is similar to SLURM, another workload manager, but has some unique features such as "bulksubmit" which allows for the submission of millions of jobs in truly parallel fashion. This is different from job arrays in SLURM, where jobs are iteratively executed one after the other. Flux is currently being used on the machines "tioga" and "tuolumne", and will be used for El Capitan acceptance runs.

UC₃ – HPE Expert System

- Motivation: Create an expert system to help users navigate the system using natural text prompts to query, submit, and create jobs and workflows. This environment will be focused on HPC workflows and should be able to answer questions and interact in real-time with the system.
- Environment
 - Use an open-source LLM engine
 - Use private LLM models
 - Use Retrieval-Augmented Generation (RAG) model
 - System knowledge corpus
 - Internal training materials are hosted on various internal platforms.
 - Public sources
 - Standalone local agent
- Investigation
 - Different LLM models and engines may use commercial/public offerings as a comparison.
 - Methods to reduce LLM hallucinations, and accuracy of LLM responses
 - The effect on LLM results with different models
 - Investigate security challenges and engagement Guardrails
 - Usability



Agent Examples

> list all nodes in the cluster The nodes in the cluster are: o184i[027,038-039,060-061,073,077-078,080,083,086-087], o185i[086-087,084,164-167], o186i[180-187,190-197,225,233-234], o184i[110-125], o186i[130-137].

> show all slurm running jobs

There are currently two jobs running on the cluster. The first job, with ID 140610, is named 'det-ai_c' and is running in the 'champollion' partition on node o186i225. The second job, with ID 140613, is named 'bash' and is running in the 'misc_gpus' partition on node o185i087.

> What is the node configuration of o184i082? Node o184i082 has 2 sockets, each with 64 cores. It has a total of 128 CPUs and 1TB of memory. The node is currently idle and has 8 AMD MI210 GPUs available.

> show all nodes which have a GPU
The nodes with a GPU are o185i086, o185i087, o184i082, o184i084, and o184i085.

Conclusion

- This study explores the utilization of Large Language Modes to enhance user experiences and system efficiency.
- Demonstrated that RAG can be an effective method to augment LLM training data without the need to retrain models
- Harness LLMs to explore large corpora of internal and external HPC data
- Demonstrated the capabilities of an expert system in assisting workload manager interactions

Future

- Looking into productizing Chat interfaces and knowledge databases
- Investigating expert agents for onboarding users
- SC24 demos

Thank you

Jonathan Sparks, jonathan.sparks@hpe.com Pierre Carrier, pierre.carrier@hpe.com Gallig Renaud, gallig.renaud@hpe.com



Confidential Disclosure Agreement

- The information contained in this presentation is considered Confidential Information of Hewlett Packard Enterprise (HPE) and is subject to the Confidentiality terms and conditions in your HPE vendor or partner agreement
- This presentation may be shared with HPE Authorized Service Providers and Channel Partners in hard-copy or electronic format for their internal training purposes only
- This presentation is NOT to be used as a 'leave behind' for HPE external customers
- The information in the presentation may only be shared verbally with HPE external customers who are under CDA and should NOT be used for commercial purposes
- Do not remove any classification labels, warnings or disclaimers on any slide or modify this presentation to change the classification level
- Do not remove this slide from the presentation
- HPE does not warrant or represent that it will introduce any product to which the information relates
- The information contained herein is subject to change without notice
- HPE makes no warranties regarding the accuracy of this information
- The only warranties for HPE products and services are set forth in the express warranty statements accompanying such products and services
- Nothing herein should be construed as constituting an additional warranty
- HPE shall not be liable for technical or editorial errors or omissions contained herein
- Strict adherence to the HPE Standards of Business Conduct regarding this classification level is critical