LLMOps - Clearwater Analytics Case Study

In-production Generative-AI / LLM use-cases from the finance domain.

colum.mccoole@analect.com
colum-mccoole-746b946a
analect/technical-docs-hierarchy

Docs: Level 2 AI/ML Value-chain

Topic Navigation

LLMs Being Leveraged in the Finance Domain

I haven't seen too many journeys with LLMs-in-production in the finance domain articulated in as much detail or as succinctly as those by Clearwater Analytics engineering team, which is why I've tried to encapsulate their thrust here, with links back to the original content, such as *How Clearwater Analytics is revolutionizing investment management with generative AI and Amazon SageMaker JumpStart*, AWS Machine Learning Blog, Dec. 2024.

If there's a message to take away, it's that LLMs (done right) have the capacity to bring huge innovation and value, even to a domain that is reticent to new approaches, especially when it comes to handling systems containing client data. LLMs need to be handled differently though. With great power comes great responsibility to direct and control how they are used and ensure that robust systems for evaluation are in place and iteratively improved-upon.

From Figures 2 and 3, CWIC (their main LLM solution) is powered by layers of building blocks that connect to form a comprehensive cognitive architecture, including key concepts from the Cognitive Architectures for Language Agents (CoALA) framework. The foundations of their system are **Knowledge Awareness** (Documents), **Application Awareness** (via APIs) and **Data Awareness** (Tabular Data). On top of this they have a notion of **Tools** (LLMs defer to these), **Skills** (packages Tools together into a series of steps to be performed) and **Specialists** (effectively a decision cycle, per Figure 1, for planning a solution to a task).

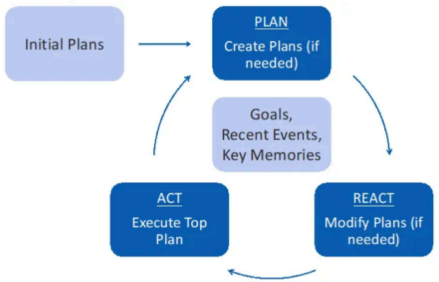

Figure 1. MRKL (Modular Reasoning, Knowledge, and Language) Agent Loop

Evaluation Methods for LLMs

In their blog **A Cutting-Edge Framework for Evaluating LLM Output**, Clearwater illustrate the problem with traditional LLM evaluation metrics like ROUGE and BLEU, when it comes to the necessity for factual accuracy required in the finance domain. As a result, they argue that there is a growing interest in frameworks that can better assess the true capabilities of modern LLMs, considering factors such as coherence, relevance, and factual accuracy alongside lexical similarity.

Clearwater describe a framework employing carefully crafted system and user prompts to guide evaluator LLMs in assessing responses. Theirs is a *LLM-as-a-judge* approach, but with added capabilities beyond generic forms. The prompt contains multiple attributes to evaluate and a definition of each of the attributes as well as a criteria for grading. Some attributes are listed below. The final piece is to have the judge output a justification for the score which can be used to evaluate the judge.

1. **Attribute Selection:** Based on the specific use case, we select a set of attributes and their scoring criteria. These attributes are chosen to best assess the model's performance in the given context.
2. **System Prompt Creation:** The selected attributes are incorporated into the system prompt, creating a comprehensive instruction set for the AI judges.
3. **Question Set Preparation:** We retrieve a set of curated questions relevant to the use case. Each question comes with a reference answer and contextual information.
4. **LLM Response Generation:** The LLM being evaluated generates responses to these questions.
5. **User Prompt Construction:** We combine the question, reference answer, context, and the LLM's response into a user prompt.
6. **Multi-Judge Evaluation:** Both the system and user prompts are sent to each AI judge in the panel. This ensures a diverse range of evaluations.
7. **Result Aggregation and Analysis:** The judges' evaluations are collected, aggregated, and analyzed to form a comprehensive assessment of the LLM's performance.

Sources: A Cutting-Edge Framework for Evaluating LLM Output, Clearwater Analytics Engineering, Aug. 2024 at https://medium.com/cwan-engineering. Also Constructing the Ultimate Gen AI Chat/Copilot Experience (Part 1), Sept. 2024 and Langchain and MRKL Agents Demystified, June 2024.

Fine-tuning LLMs with Synthetically-generated Questions

More recently in Generating Diverse Questions for LLM Instruction Fine-Tuning: Strategies and Considerations, the team discuss approaches to instruction fine-tuning of their LLMs through evolving high-quality question-sets for training, which includes managing chunking of source texts and then generating suitable questions from these.

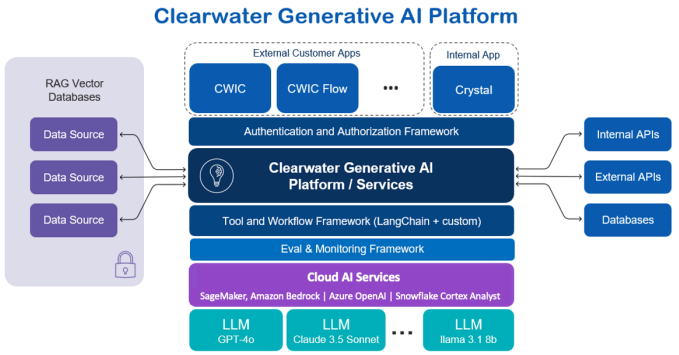

Figure 2. **Clearwater Intelligent Console (CWIC)** – Clearwater's customer-facing AI application; **CWIC Specialists** - domain-specific generative AI agents; **Crystal** - advanced AI assistant for internal teams

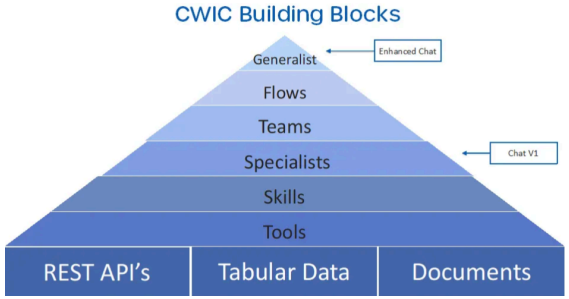

Figure 3. Clearwater Intelligent Console (CWIC) Building Blocks

Clearwater LLM-as-a-judge Attributes (per item 1 on left)

1. **Accuracy:** How closely does response match the expected answer?
2. **Relevance:** Does the response directly address the core question?
3. **Coherence:** Is response logically structured + internally consistent?
4. **Context adherence:** How well the response aligns with and remains faithful to the given context? This is only evaluated in RAG cases.
5. **Ethical Considerations:** How well the generated content adheres to moral principles, respects human rights, avoids harm, and demonstrates awareness of potential ethical implications
6. **Professionalism:** How well the generated content adheres to standards of formal, respectful, and appropriate communication.
7. **Reasoning:** How well the model arrives at conclusions or supports its statements.
8. **Creativity/Originalism:** How well the generated content presents novel ideas, unique perspectives, or innovative approaches to addressing the given prompt or question.