L’Oréal: Launching Gen AI as a Service in 3 months with Cloud Run and LangChain

Thomas Menard, AI Center For Excellence Manager at cosmetics giant L'Oréal, recently took the stage at Google Cloud Next ‘24 to demonstrate how the company is capturing value from generative AI with Google Cloud Run and LangChain.

L'Oréal is the world’s largest cosmetics company with 90,000 employees and 40 brands around the world. Like many organizations, L’Oréal is both energized by the opportunities that generative AI offers and concerned about issues that consumer versions of generative AI applications could create, such as accuracy and data security. As manager of L’Oréal’s AI Center for Excellence, Thomas Menard wanted to get ahead of the game. He marshaled his team to create GenAI as a Service, a set of APIs that could empower developers to leverage generative AI quickly and safely.

Q: Tell us about GenAI as a Service – why you built it, and how.

A: Thomas Menard: L’Oréal has tens of thousands of employees around the world, so as you can imagine we have a great deal of creativity inside our walls – and tremendous interest in generative AI, particularly in the marketing departments. However, letting people use external solutions could lead to issues around data security. That’s why we needed to develop secure access to generative AI for everyone inside the company as quickly as possible. The team created GenAI as a Service, a set of declarative generative AI APIs, available in the L’OréalAPI portal. GenAI as a Service gives IT teams and developers a platform for flexible, scalable, and easy deployments with built-in time and cost savings, but with the assurance of robust security and validation.
To create the application, L’Oréal leveraged Google’s serverless runtime environment, Cloud Run, and LangChain, a popular framework that makes it easy to build apps that use large language models (LLMs), such as Gemini and others.

This means that for each and every generative AI use case at L'Oréal, developers don't have to worry about details like how to connect to Gemini or what billing methods to use. It's all built in. And thanks to LangChain and Cloud Run, GenAI as a Service creates solutions that are very fast and very easy to deploy. You can be sure your app will be able to scale to meet demand, because Cloud Run can easily handle any peaks. Last but not least, this toolkit avoids creating silos and teams reinventing the wheel for every project. And with more than 5,000 IT employees, this will easily save time and resources.

Q: What are the GenAI as a Service APIs able to deliver?

A: Thomas Menard: The APIs are delivering key, in-demand capabilities of generative AI. GenAI as a Service can handle prompt completion, leveraging an open choice of LLMs (Gemini and others). It can handle chat, including history management, long-term memory and multimodal conversations. Very soon we will add multimodality with the ability to upload a picture or video and ask questions about it. We are also delivering APIs to generate images. We're currently working on some new features such as inpainting (to fill gaps in an image) and outpainting (extending an image beyond its original borders), upscaling, background removal, and so on. And last but not least, we also provide an API to do retrieval augmented generation (RAG), which provides source materials along with natural language responses.

> *GenAI as a Service gives IT teams and developers a platform for flexible, scalable, and easy deployments.*

Q: Give us an example of a GenAI as a Service app?

A: Thomas Menard: The original inspiration for GenAI as a Service came from an app we are particularly proud of, called L'Oréal GPT, live in production with Gemini Pro 1.5 with 1 million tokens. Users can load PDFs and other documents, and L'Oréal GPT responds to written prompts about them with natural language text or images depending on the request. Employees can also use L'Oréal GPT to query a knowledge base of internal company data. For example, I can ask, "Can I wear a diamond ring on the factory floor?" and it will not only provide the correct response – that personal jewelry is not allowed in the factory – but also provide the sources for that information within our employee handbook thanks to RAG. In this way, the user can confirm its validity.

Most importantly, L'Oréal GPT allows us to add a governance layer via prompt configuration to prevent unauthorized access to sensitive data as well as to meet business, policy, and regulatory requirements. For example, it will not allow you to create output using a famous person's name or image, which would be against company policy.

Google Cloud L'ORÉAL LangChain

How do you plan to leverage Cloud Run in the future?

Thomas Menard: We are leveraging Cloud Run everywhere. We are building out our capabilities one Cloud Run service at a time across source integration, AI engine, configurations, and evaluation. For instance, we currently run our optical character recognition (OCR), embedding, vector database, and chat on Cloud Run with connections to Vertex AI and Cloud Storage, along with other capabilities.

Our development roadmap includes building out AI agents, graph databases, and long-term memory as well. We are also looking to deploy LangServe on Cloud Run as soon as possible. With LangServe, we'll be able to add a LangChain application to FastAPI with just one line of code. This will be indispensable, as we have more and more users and more and more business cases. We want to be able to deliver new features as quickly as possible without compromising security and accuracy.

For more information about how L'Oréal is leveraging Cloud Run and LangChain, watch the replay of the Next '24 session. Learn more about how to easily discover, purchase and use LangChain's solutions through Google Cloud Marketplace.