From Fascinations with Foundation Models to a Useful Conversational AI Application

Cheong Ang
IBM Corporation
Los Altos, CA
cheong@ibm.com

ABSTRACT

Since ChatGPT mesmerized the world with its capability to generate interesting answers, fascinations and fears around generative AI (genAI) have been compounding as new genAI capabilities from researchers and the industry frequently made headlines. Venture capitals poured $21.8B into genAI startups last year, and 36 companies hit the unicorn status. Across industries, including defense, cybersecurity and healthcare, leaders are fascinated by genAI's potential to not only surface insights in multi-modal data sources (structured, text, image, video), but also interface with humans in natural language. Their fears range from safety and privacy issues to irresponsible applications that lead to unethical decisions or cyber vulnerability exploitations. Industry leaders clamor for AI governance as organizations from the European Union to the Whitehouse published their evolving guidelines.

Application wise, beyond Q&A, multiple gaps exist toward realizing the power of genAI in a typical workflow. A Large Language Model (LLM) or Foundation Model (FM) doesn't know an organization's workflow, the data required from the user, and the enterprise system(s) to interact with to submit a request. An LLM/FM also lacks the ability to conduct multi-turn conversations to gather the information to complete such request.

Before ChatGPT, the popularity of messenger applications brought about the chatbot industry. Chatbots interpret the user intent, process their requests, and give relevant answers (Mordor, 2024). It provides a foundation to close the gaps to produce a truly conversational AI system configurable to understand workflows, and integrable into the organization's IT environment to gather insights across systems and complete work on a user's behalf. It leverages multiple LLMs/FMs as required.

This paper describes the gap-closing components and complementing LLMs/FMs in an architecture compatible with the Zero Trust security framework and AI governance guidelines. This combination takes Human-Computer Interaction to where an LLM alone cannot, enhancing the mission effectiveness of the workforce.

ABOUT THE AUTHORS

Cheong Ang, M.S., M.B.A., is a client-focused IT solution architect who has worked across industries in Federal, Healthcare, and Internet for more than 20 years. He has developed and operated a virtual care application that was serving 40,000 patients; helped to drive a data science implementation at a health system to achieve an estimated gain of 624K patient-care hours and labor expense savings of $90M/year; co-founded and built a startup to its product-market-fit stage; introduced and provided two pivotal IT platforms to a large health system as a win-win proposition among the tech supplier, health system and its patients; presented insights of AI in healthcare at HIMSS among other conferences; and co-invented patented technologies licensed to enterprises including Microsoft, Adobe, and Oracle.

From Fascinations with Foundation Models to a Useful Conversational AI Application

Cheong Ang
IBM Corporation
Los Altos, CA
cheong@ibm.com

INTRODUCTION

ChatGPT has not just turned AI into a household concept, it has gotten all organizations to experiment with or at least strategize for organization-wide adoption of some AI. There is a fear of being left behind: the potential benefits are predicted to be revolutionary. In addition, their employees have already experienced consumer-level offerings from OpenAI, Meta, Google, and others.

Organizations think business benefits, and they come either in the form of automating tasks or augmenting the workforce (Keller et al, 2024). The former brings efficiency to the workflow and the latter the possibility of fundamental changes to how work is done. Productivity (GDP per hour worked) increase is a given. The crown jewel is leapfrogging competition, whether that's another company or country.

While AI has been evolving for tens of years from the expert systems of the 80s, machine learning of the 90s, deep learning around 2010s to the current transformers-based networks since 2017, AI is still a new tech without a proven track record, business benefits wise. The primary obstacle to AI adoption, as reported by 49% of participants in a survey conducted by Gartner, is the difficulty in estimating and demonstrating the value of AI projects. This issue surpasses other barriers such as talent shortages, technical difficulties, data-related problems, lack of business alignment and trust in AI (Gartner, 2024). A prudent AI project strategy tends to target quick win(s) while building up the potential for a breakthrough, meaning the project must show some success to be funded further. Specifically, the strategy seeks to automate tasks in existing workflows with proven AI capabilities (e.g. text summarization, content generation) while adopting a platform/architecture that enables experimentation and progress toward some innovation. As AI brings with it issues associated with data security and privacy, safety, bias/fairness, explainability, transparency and accountability, an organization needs to manage and mitigate its related risks to use AI responsibly. Hence the strategy will need to consider also the people and process aspects of AI adoption. For example, retraining workers so they thrive in an AI-infused workplace and establishing metrics for risks and benefits as well as capability to monitor and course correct continually.

There is no fixed formula for the planning and execution of an AI project under such strategy. This paper describes the technical and process aspects of an AI assistant platform/architecture based on the evolved foundation of chatbot technologies. The popularity of messenger applications brought about the chatbot industry in the recent years. Chatbots interpret the user intent, process their requests and give relevant answers (Mordor, 2024) with tools curated by the organization. Such platform/architecture aligns well with this strategy by integrating into a typical organizational workflow (e.g. automating tasks in customer care), and enabling processes to monitor risks and experimentations in pursuit of innovation in many use cases.

REQUIREMENT: AUTOMATE WORK NOW, ENABLE EXPERIMENTATION FOR INNOVATION

A person who has experienced ChatGPT and other consumer-oriented AI (e.g. Meta's WhatApp AI Agent) may be under the impression that LLMs and FMs (generally, as AI models are not restricted to language) are readily deployable in business use cases. What they experienced are the chat interface of these apps. A peek at the server side (vs. the client side where the app and the consumer are) would reveal that much more went into delivering the user experience.

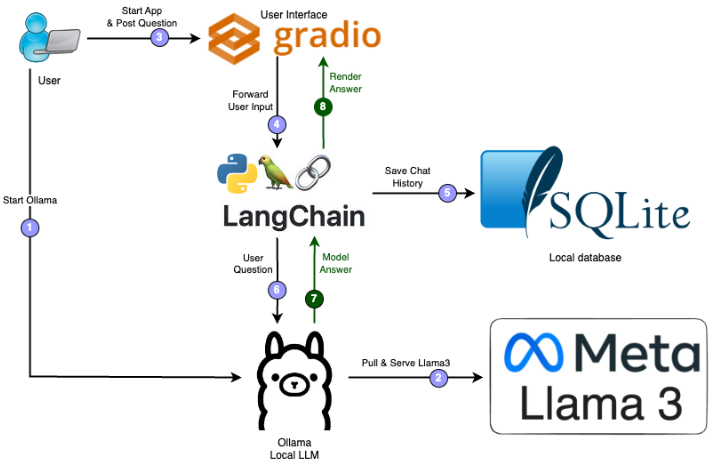

Figure 1. A Simple LLM-based Chat App (Merreider, 2024)

Figure 1 depicts an LLM-based app consisting of a simple chat app (Gradio) and an LLM running in an inference engine, Ollama. One has many choices other than Gradio and Ollama for this simple architecture, e.g. Streamlit in place of Gradio; also, the entire inference engine may be SaaS based, using a service like OpenAI instead of Ollama.

This architecture won't work for a typical enterprise deployment for many reasons. Under the strategy above, the AI system needs to robustly integrate into existing workflows, enable some wins and support experimentation for leapfrog innovation, as well as monitor specific metrics for risks and benefits to automate/guide course corrections.

Integrate into Existing Workflows

In the customer care scenario, an example of integration of AI in the workflow is answering questions and completing transactions such as booking a trip. Trained on Internet data, a typical LLM may handle the simple, single-turn question-and-answer task by giving a response "confined" to its training data. It is incapable of answering according to the organization's policies and task requirements. In fact, LLMs are not strictly confined by the training data as they are known for making up answers (i.e. hallucination or confabulation) if not invoked with a restrictive prompt (e.g. "if you don't know say 'I don't know'", or "use only the answer provided below") and parameter (i.e. low "temperature"). Also, the organization's customers are likely to converse naturally, expecting the chat app to be capable of having a lengthy, multi-turn conversation. Ideally, the chat app keeps tab of everything the customer said (i.e. context), which may involve disambiguation (clarifying any potential misunderstanding), digression (to another task or to handle request for additional information), collecting the necessary data (to complete a task), and cancellation (of the current intent). If the conversation is voice-based, the chat app needs to handle turn-taking and backchanneling (e.g. short utterances expressing acknowledgement such as "uh-huh") (Wang et al, 2024) appropriately to ensure a smooth user experience. The state-of-the-art chat apps OpenAI and Google have demonstrated also use multimodal inputs including audio, visual, and other device-sensor signals to provide additional context for its actions. In short, the most advanced chat app is getting human-like, reacting using all available information and even speaking with a tone that reflects the appropriate emotions.

Enable Some Wins and Support Experimentation for Leapfrog Innovation

Much of the AI-based innovation is expected to hinge on the intelligent outputs of the AI models and the actions influenced by these outputs. The actions may be fully AI automated or involve humans as in the case where the AI provides decision support with "human in the loop". The term "agent" has been used in the industry to describe such combination: AI with intelligence which can perform actions.

Agents may specialize in their respective capabilities such as web search and report writer. Several agents, each playing a different role, may be combined to serve a higher-level function such as market research. Agentic specialization also allows each agent to be optimally equipped with its respective smaller, domain-specific AI model, which tends to outperform large, general-purpose models and can run on small-footprint devices.

A platform capable of combining agents for different functions can serve more than one use case. This flexibility enables an organization to deploy mature agents and AI models for quick wins and explore new agents and AI models for moonshots.

Monitor Specific Metrics of Risks and Benefits to Help Automate/Guide Course Corrections

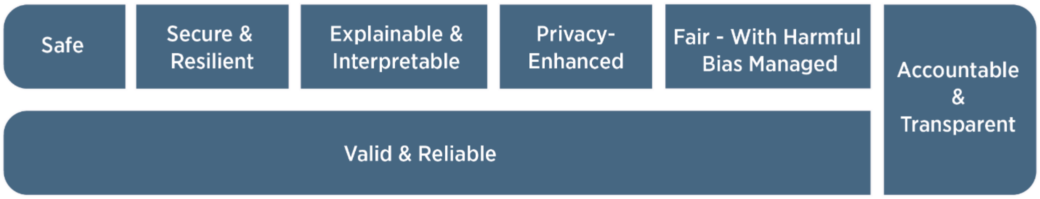
Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable

Figure 2. Characteristics of trustworthy AI systems (NIST, 2023)

Risks are worrisome across the industry. In a survey of 100 Fortune 1000 executives who are reporting to their respective CIOs, all the executives have concerns about the genAI security risks, 51% of them worry about copyright and legal exposure, and 47% data privacy violations (PagerDuty, 2024). In the early 2023, the National Institute of Standards and Technology (NIST) published its AI Risk Management Framework (AI RMF) to help organizations determine a process and corresponding metrics and tools to track risks and enable appropriate course corrections. Figure 2 shows the characteristics of trustworthy AI systems in the NIST AI RMF. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics (NIST, 2023). Amazon Web Services (AWS) also put forth a genAI security risk scoping matrix that is pragmatic to the type of models involved (Saner et al, 2023). It is prudent for an organization to have a central AI governance board that defines the policies and establishes governance procedures and tools. While a detailed discussion of risks in the NIST AI RMF is beyond the scope of this paper, the basic considerations of risks and benefits should include the characteristics depicted in Figure 2. For example, for the Secure characteristic, implementation of security authentication and access authorization impose restrictions on who can use a certain AI app/model. Also, implementation of cybersecurity capability for threat detection and response protects the organization from cyberattacks. Similarly, for the Explainable & Interpretable, Privacy-Enhanced, and Fair-With Harmful Bias Managed characteristics, implementation of data governance and model governance establishes governance policies and tools, logs data, monitors metrics, and automates compliance activities, which enables auditability of AI-related decisions and actions. Auditability of AI-related decisions and actions allows the organization to continually assess AI-related workflows and operations for risks and returns, course correcting as appropriate, to realize the foundational characteristic of the NIST AI RMF: Valid and Reliable. In addition, evaluating select data, metrics, and AI-related decisions and actions against business impacts in terms of ROI as well as ethical and responsible use of AI serves to uphold the Accountable and Transparent characteristic.

THE PLATFORM/ARCHITECTURE AND PROCESSES

Let’s consider the following architecture as one that satisfies the criteria of the strategy above:

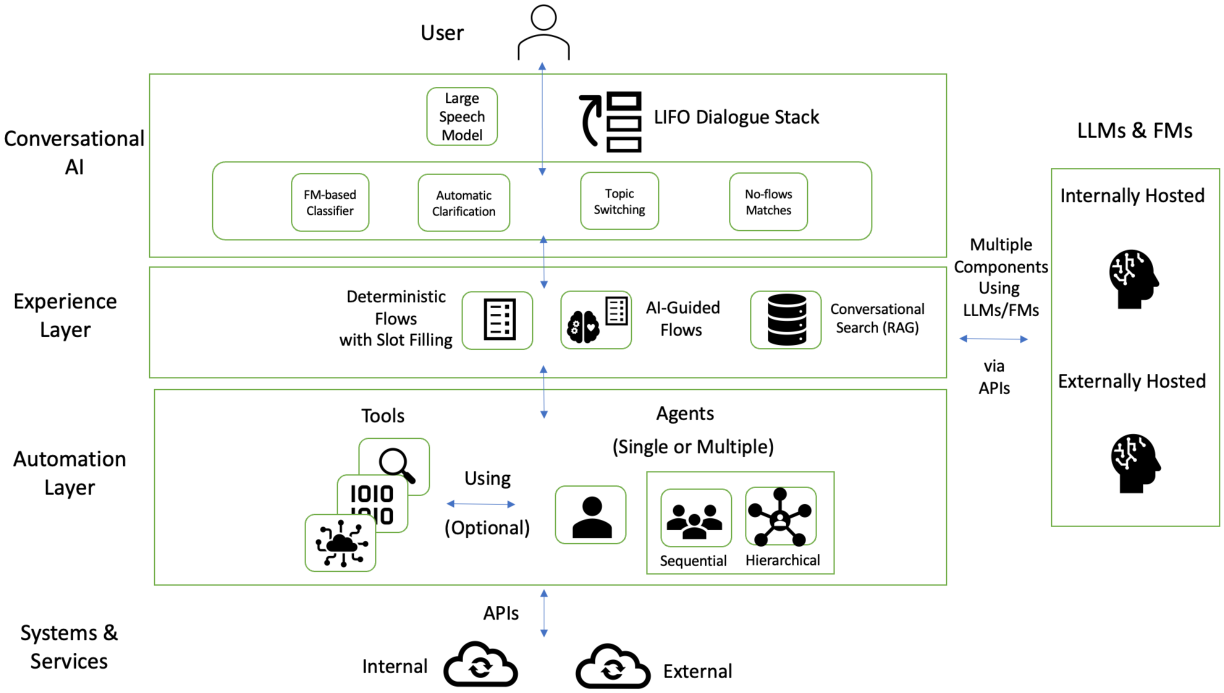

Figure 3. Architecture of a Conversational AI System

The Components

The architecture above interfaces with humans via a conversational AI or chat component for a few reasons. Prior to ChatGPT, multiple genAI models (BERT, BLOOM, PaLM, GPT, GPT 2, etc.) have been used to support various use cases via APIs in python code. It requires programming skills to experience them and put them to use. ChatGPT launched AI to its popularity via a chat interface. Furthermore, texting has turned into the foundational communication mechanism as Short Messaging Service (SMS), messengers, web chat, and smartphone apps evolved over the years.

However, the chat interface in the architecture diagram may support more than texting with the AI found in the initial version of ChatGPT. OpenAI, Google, and others have demonstrated AI smartphone and glasses that incorporate what the AI sees including a diagram on a whiteboard or a building in the real world, and hears such as background music, via camera and microphone for video and audio inputs to respond to the user's requests. Separately, Dr. Fei Fei Li showed a spatial AI setup in which a user instructed a robotic arm to prepare recipe ingredients and put them in a pot using a noninvasive EEG cap with brain signal sensors placed on the user's head (Li, 2024). The input has evolved from textual, including setups that use a tech between the user and the AI to transform speech to text or image to text, to truly multimodal – the non-textual inputs are turned into input vector embeddings, a set of numbers that represents the gist of the inputs, in the context of the AI system without being converted to text first. Further, the AI may get additional context by accessing other data including websites, databases, and sensors (e.g. thermal, GPS, lidar) to gain a deeper understanding of the conversation and act accordingly. Multi-modality brings exciting possibilities; equally so is the evolution of AI's ability to converse like a human. The user expects the conversation with AI to be smooth – tolerating nuances including interruptions, cancellations due to changes of mind, and digressions (Bocklisch et al, 2024) – and helpful in terms of AI's ability to understand the user, provide relevant answers, and accomplish tasks for the user.

Referring to Figure 3, the conversational AI component works with the Flows in the Experience Layer to deliver the user experience and/or task automation. A Flow may be viewed as the specification for a tailored user experience. Flows are modular and may be triggered by the conversational AI component as conditions match. Flows may be intended for specific purposes such as helping the user with an HR topic or guiding the user in filing in an IT trouble ticket. The former example is Q&A-centric while the latter intends to collect data ("fill slot") for submission to certain system(s). Flows enable the organization to compartmentalize the supported workflows and be deliberate about what the AI can and should do.

The conversational AI component would try to identify candidate Flows as it chats with the user, using all historical context in the current chat session and beyond, including the user's profile information. The organization may still have a default ("No-flows Matches") component that can be as chatty or restrictive, e.g. urging the user to select from a limited set of menu items, as the organization would like. It is important to point out two differences between this setup and the traditional chatbot architecture, which aims to identify a user intent and then focus on completing the workflow behind the intent. While this setup still performs user intent identification, the conversational AI keeps a running list of candidate Flows and guides the user to complete them (or cancel if confirmed). It is a departure from the more single-minded approach of attempting to match an intent. As a result, the conversation is more natural since the AI can use any information it has learned even before an intent is identified to both fill the slots of an intent and infer an intent or its cancellation without explicitly asking the user. This setup may be implemented with a Last-In-First-Out (LIFO) Dialogue Stack (Bocklisch et al, 2024) that keeps the immediate instructions such as reacting to the user's current question or comment at the top of the stack, above the Flows-invocation instructions that might have been added when the user previously expressed some possible intents, e.g. several possible credit card related Flows. This way, the conversation AI can address the user's question or comment first as it has not had the chance to gather the required data, confirm the intents such as clarifying whether the user wants to freeze, unfreeze, or cancel a credit card, and execute the corresponding Flows. After addressing the topic at hand, the conversational AI digresses to the previous topic(s) or Flow(s), maintaining a natural conversational flow with the user.

It is worth differentiating Flow and agent. Recall above that an agent is AI with intelligence which can perform actions. More specifically, an agent is a service that processes its inputs using an LLM/FM, tapping predefined tools (e.g. web search and/or other API calls to some systems such as HR Management System) and optionally short-term or even long-term memory, to produce outputs. Typically, an agent serves a specific role, e.g. researcher, web search, report writer. An application can call on the web search agent to search the web and provide its summary. The industry has also explored having multiple agents collaborate on a task. For example, to answer a user's question on "what is new in AI in 2024?", the researcher agent starts with planning, calls the web search agent with relevant queries, reviews the results, and calls the web search agent again on any new ideas that surface in the reviews. Then the researcher agent engages with the report writer agent iteratively – generate report, critique report, refine report based on the feedback – to get to a satisfactory report.

A Flow, as mentioned previously, focuses on a tailored user experience such as a specific conversation flow or workflow pattern. While a Flow can invoke an LLM/FM or a tool directly, it may also call on an agent to complete a task. The important point is that the organization may separate a Flow that defines the user's experience from the tool(s) or agent(s) the Flow uses to perform work. This allows for experimenting, e.g. A/B Testing, with different user experiences to achieve the same work or experimenting with new tools/agents and approaches such as multi-agent collaboration above to accomplish work with the same user experience.

As the LIFO Dialogue Stack-based design is highly scalable in comparison to the intents-based design (Bocklisch et al, 2024), one may have many combinations of Flows, tools, and agents deployed behind the same conversational AI component for a variety of purposes, including production, beta, and R&D, to support innovation efforts.

The Processes

While the architecture consisting of conversational AI, Flows, agents, and LLMs/FMs explains the inner working, the quality and auditability of the AI lie in the processes and the governance infrastructure around them. OpenAI, Google, and Microsoft have shown off the magic of a multimodal AI system. Let it see what is on your screen and it can debug your code and even help you in playing Minecraft. Allow it on your computer, and it can recall for you a PowerPoint slide with the purple text you remember seeing or the tagline your colleague came up with in a virtual meeting a week ago. It also sees the webpage displayed in your browser, and can enter data, scroll, and click to interact with various applications on your behalf. Clearly, AI-driven automation is not limited to API calls as the AI understands what it sees in an application and can interact with the user interface designed for human. Nvidia has also demonstrated using an FM trained in a simulated world to drive an autonomous robot in the real world. The robotic interactions and the physics in a world are just additional modalities in the FM. All these beg the question: what if the AI on the computer and/or the robot don't act in our interest?

Some experts simplify how the transformer-based AI system works: it's a new way of data representation that enables appropriate predictions of what should come next. Some went as far as calling an LLM a glorified auto-complete system that predicts the probable next word. Obviously, the core of such a system is the data behind the LLM. The backend processes that admit the data (DataSecOps) (Figure 4), train and verify the model (ModelSecOps) (Figure 5) and expose the model endpoints or integrate the model into applications (DevSecOps) (Figure 6) need to come together seamlessly and continually so through the lifecycle of the AI system for the system to work well.

These processes are crucial in delivering the Valid & Reliable foundational characteristic in the NIST AI RMF (Figure 2). Hence the datasets, the models, and the APIs and/or applications will have passed their respective quality-assurance gates via automated and/or human-in-the-loop testing and approval. Another critical observation is that the 3 processes are interconnected:

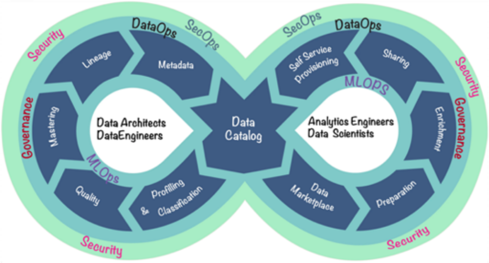

Figure 4. DataSecOps Process (Miyake, 2023)

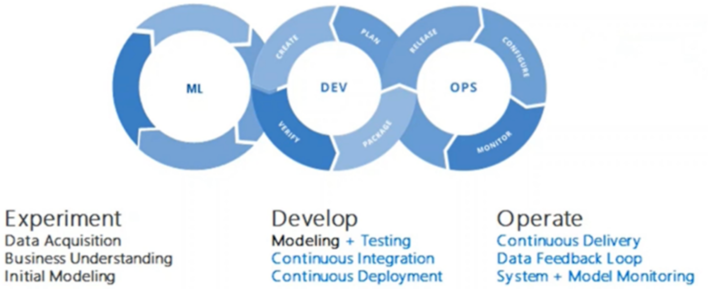

Figure 5. MLSecOps Process (Zahra, 2019)

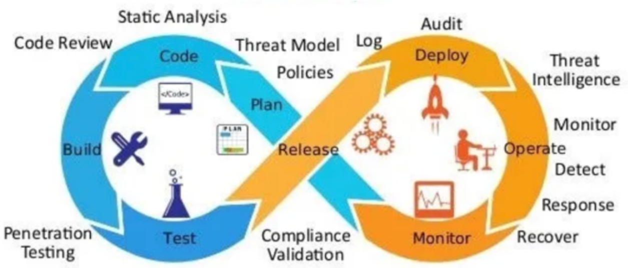

Figure 6. DevSecOps Process (Singh, 2023)

- DataSecOps: aiming to curate the appropriate datasets for the AI system, this process incorporates not only select sourced data but also feedback data from DevSecOps and possibly ModelSecOps.
- ModelSecOps: targeting to produce AI models that meet both business and operational requirements, this process uses the approved datasets from DataSecOps and provides approved models to DevSecOps and feedback to DataSecOps.
- DevSecOps: supporting the development and deployment of the AI APIs and/or applications, this process uses the approved models from ModelSecOps and provides feedback to DataSecOps and possibly also ModelSecOps.

These processes support an AI system in the background throughout its lifecycle, ideally completing with automation that logs the audit trail of the data provenance, workflows, approvals, inferencing activities and their corresponding inputs, among others, to continuously keep tab of the development, evolution, and health of the AI, crucial to not just the operations but also the governance and security of the AI system.

DISCUSSIONS

The two examples below are actual AWS patterns that provide more details on how the architecture/platform described may automate work now and enable future innovation while providing governance and ensuring security.

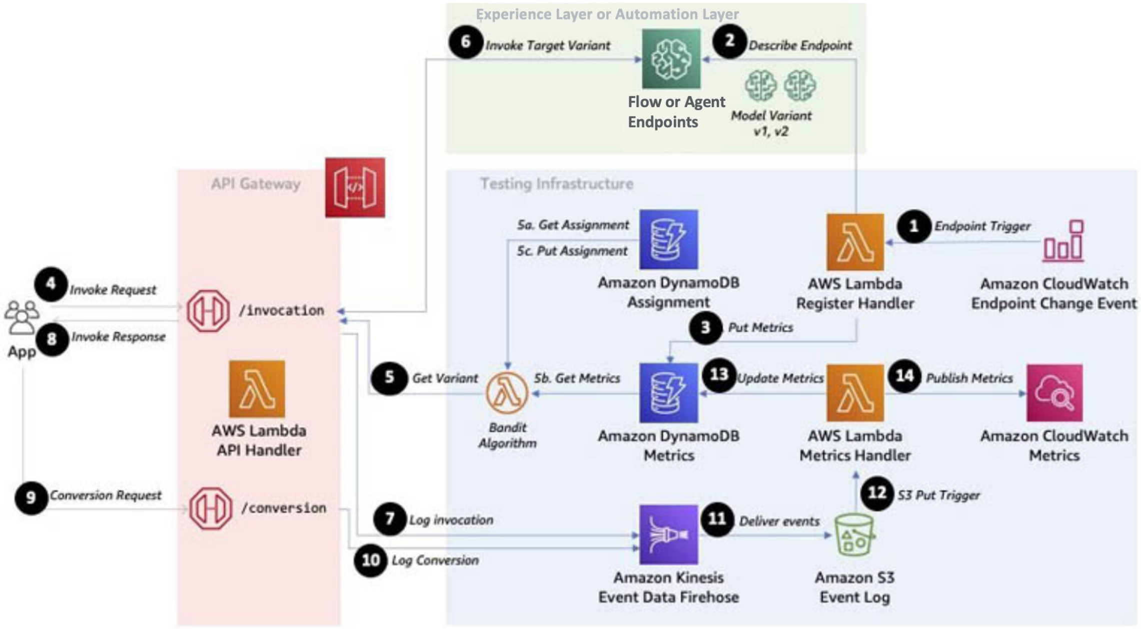

Figure 7. Architecture for Incorporating A/B Testing in an App in AWS Cloud (adapted) (Bright, 2021)

Automate Work Now and Enable Future Innovation

To support experimentation, the capability to route some user traffic to the test setup(s) and leave the rest in the production setup is critical. One way to achieve it is with reverse-proxy based network routing, which in Figure 7 the role is fulfilled by the AWS API Gateway. The conversation AI component serves the "App" in Figure 7 and is embedded in the organization's workflow(s). When the conversation AI invokes a specific Flow or agent, the API Gateway consults the Bandit Algorithm for the variant of the Flow or agent to route the user traffic to. The Bandit Algorithm selects among Flow/action variants sequentially based on the probability of a selection being optimal via Thompson sampling to efficiently balance the trade-off between exploration and exploitation (Klarich et al, 2024). Note that this test setup does not only log an invocation, i.e. which variant is called by the conversation AI, but also its corresponding conversion. The organization defines what a conversion is; for example, when the Flow variant successfully guided the user to complete the target task, or the agent variant's output won a favorable user feedback. Coupling with instrumentation / dashboarding, one for a specific use case, to gauge Key Performance Indicators (KPIs), the organization will be able to compare the baseline metrics, benchmark metrics, and experiment metrics.

Experimentation may be carried out on various aspects of the system – Flows, tools, agents, and LLMs/FMs – in Figure 3, but the discussion here focuses on the Flows and agents. The idea is to route a percentage of the user traffic from the conversational AI to the test variants of the existing production Flow and/or agent per the experimental designs. Recall that a Flow specifies a series of interactions to define a conversational experience, and an agent has ability to plan and may consist of memory, tools, LLM, among others. Also, an agent may coordinate (hierarchically) or collaborate (sequentially) with other agents to accomplish a goal, possibly optimizing for a specific reward function. An experimental design may try different combinations, e.g. within an agent, testing its tools and LLM; and at the agent level, testing the planning ability and multi-agent interactions in search for one that delivers optimal results/KPIs.

Imagine a Flow that takes the user through a series of prompts for the user to complete a loan pre-approval. The existing Flow may include going over every question needed for the pre-approval with the user. An improved Flow may leverage retrieving and verifying information the organization (e.g. bank) already has, e.g. the user's assets, credit score, and even incomes, and work with the user on only several remaining questions. In addition, the user's past and/or current interactions with the conversation AI component might have already provided information regarding new updates about the user's household. This conversational Flow variant may be tested alongside the existing one in an A/B Testing setting enabled by the routing capability above.

In conjunction with agent(s), a Flow may improve a workflow, e.g. upon verifying that the user would like to save on her car loan, agents may be retrieving details about the car and available loans with APIs from external parties such as the Department of Motor Vehicles and optimizing for the best offer to present to the user. At times, improvements may go beyond workflows into the work itself. Using the example above, while refinancing a car loan (the work) may save some money, consolidating multiple loans under a home equity loan (the new work) may be more optimal for the user, a proposal a financial analyst agent might have come up with in the collaboration of multiple Agents. There is even a possibility for the organization (e.g. bank) to innovate on its value creation. For example, in consideration for offering a 529 college savings plan to the user, the financial analyst agent above may consult a tax expert agent, which may incorporate in the conversation some tax advice and services, a new value-add the organization may consider offering.

The Governance and Security

The Center for Security and Emerging Technology (CSET) states 3 critical points relevant to AI safety: robustness, assurance, and specification. Robustness guarantees that a system continues to operate within safe limits even in unfamiliar settings; assurance seeks to establish that it can be analyzed and understood easily by human operators; and specification is concerned with ensuring that its behavior aligns with the system designer's intentions (Rudner et al, 2021). The following discusses how the architecture and processes come together to provide robustness, assurance, and specification.

If one were to audit how a decision is made, the investigation will likely touch on contributions of specific features/attributes to the model's inference or recommendation. In the case of approval or denial of a loan, the investigation may include whether protected attributes such as race and gender influence the decision. Referring to the NIST AI RMF, this is related to the Explainable & Interpretable and Accountable & Transparent characteristics. Subsequent questions may touch on if or why the presence of such bias – the Fair – with Harmful Bias Managed characteristic. Peeling back the onion: what gates are in place, or whether a sound process was in place to ensure other characteristics including Safe, Secure & Resilient, and Privacy-Enhanced are met? This questions how the model was trained and what datasets were used, among others. Equally important is whether and how the AI system is kept up to date once deployed by monitoring and addressing any issue with data and model drifts and incorporating performance feedback in terms of business impacts, e.g. fairness, regulatory compliance, and productivity goals. The controls and gates are reflected in the ModelSecOps and DevSecOps diagrams (Figures 5 and 6) as their ***monitor*** activities.

While the expectation is that AI platforms like AWS SageMaker and Azure Machine Learning would log the model training and validation activities in the process of approving and publishing the model into a model registry, AI platform tools such as IBM watsonx.governance goes a step further by automating tracking of the data, model, and inferencing activities as part of the DataSecOps, ModelSecOps, and DevSecOps processes to provide factsheets organized by use cases such as loan application. To monitor the AI system's operations, an AI platform may include an evaluation store that logs model inferences and provides the operational insights into issues mentioned above (e.g. drifts, bias). Overlaying the model inferences and insights from an evaluation store with additional business data such as customer and product, the organization may compute business metrics related to AI-driven customer interactions, e.g. effectiveness of a loan promotion in customer-care chat sessions.

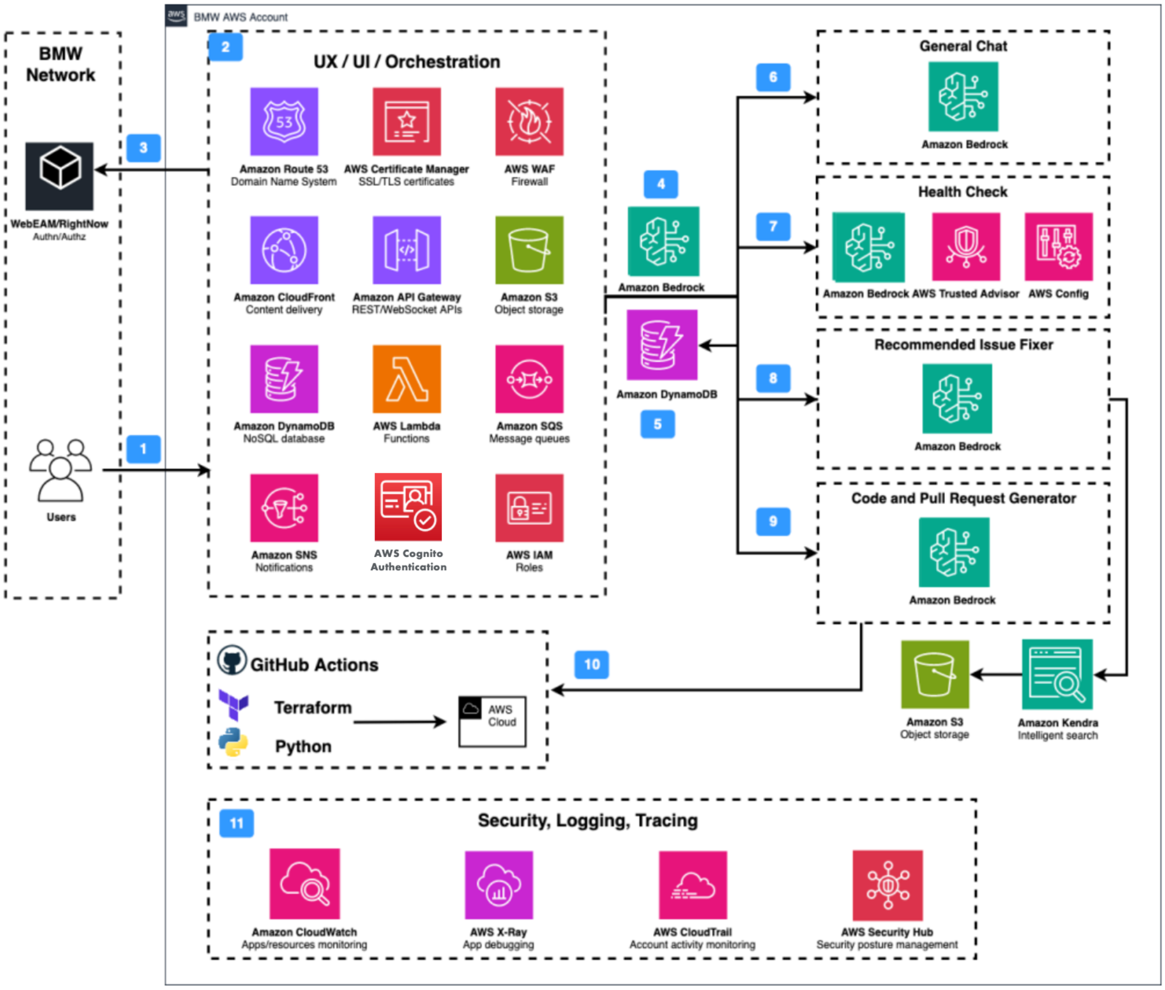

Figure 8. Architecture of BMW's Infrastructure Optimization GenAI Assistant (adapted) (Kohl et al, 2024)

Privacy and security are front and center in the NIST AI RMF. While the DataSecOps process should establish the necessary precaution regarding any inappropriate use of Personally Identifiable Information (PII), privacy and security may be exposed on the network or cloud where the storage and compute resources reside. In addition, models and software libraries may be exposed to supply chain risks. From the perspective of chief information security officers (CISOs), consumption of genAI applications in business experiments and unmanaged employee adoption creates new attack surfaces and risks on exposure of individual privacy, sensitive data and organizational intellectual property (D'Hoinne et al, 2023). Figure 8 is the pattern of the conversational AI assistant solution BMW Group used to help its DevOps teams streamline infrastructure optimization efforts. The pattern implements many aspects of the Zero Trust (ZT) framework (Syed, 2022) for secure deployment of an AI system. The deployment is on the AWS cloud, but the concepts of authentication, authorization, user roles, access policies, endpoint protection, etc. discussed below are applicable on other clouds.

Identity-based access management: When the user makes a request, it's routed to Amazon Cognito for authentication. Then an AWS Lambda-based authorizer helps determine the authorization from the identity layer, which is managed by the DynamoDB table policy. If the client has access, the relevant access such as the AWS Identity and Access Management (IAM) role or API key for the agent's endpoint are fetched from AWS Secrets Manager (Chattha et al, 2023). The setup segments and isolates the resources (shown as icons within the BMW AWS Account) with appropriate placements in AWS Virtual Private Clouds and access controls defined via AWS Security Groups and AWS IAM policies. This setup reflects several considerations (pillars) of the ZT framework: User, Application & Workload and Network & Environment.

Logging to enable auditability as well as security and compliance monitoring: Detailed logs record user interactions, model requests, and system responses. These logs provide valuable information for troubleshooting, tracking user behavior, and reinforcing transparency and accountability (Chattha et al, 2023). These logs and the processes around the system including Threat Model, Threat Intelligence, Detect, Response, Recover shown in DevSecOps (Figure 6) facilitate security analysis of events, activities and user behaviors (the ZT Visibility & Analytics pillar), as well as automation of security response based on defined processes and security policies (the ZT Automation & Orchestration pillar).

Agents-based architecture: Each agent (e.g. Health Check, Recommended Issue Fixer) in Figure 8 is an intelligent system designed to reason, make decisions, and take actions using the LLM and available tools — the interfaces to services, functions, and APIs, such as abilities to use search mechanisms or execute the code (Kohl et al, 2024). Such multi-agent system brings several advantages. First, it fosters modular development, debugging, and testing of the system. Secondly, multi-agent design enables responsibility separation between different components or functions of the system. This makes the agents more controllable and secure as each agent's behavior, inputs and outputs, can be separately monitored, tested, and equipped with security guardrails (Kohl et al, 2024). Security considerations include protection of API endpoints and countermeasures to attacks targeting the AI API endpoints. The latter includes data poisoning (as an AI system may incorporate data seen during inference time to continually update its model) and prompt injections that purposefully lead the AI to act outside of the scope it was designed for.

The governance and security features in the architecture and processes above are in place to ensure safe and explainable operations for the system's intended purposes to fulfill CSET's robustness, assurance, and specification AI safety criteria.

CONCLUSION

The future AI would bring is uncertain. But AI-related efforts have heated up competition among nations, organizations, and even individuals as they are leveraging it for productivity improvement and eyeing leapfrog innovation. The ROIs are not obvious but are expected to be substantial. Certainty is only predicted for those that are not onboard – as in "workers with AI will replace workers without".

A platform that enables pursuits of innovation while delivering productivity improvement is well aligned with the approach of solving big problems with small wins. The architecture/platform described in this paper achieves small wins by integrating into the existing workflows with conversational AI, enabling deployments of both production Flows and agents for the winning use cases, and test Flows and agents to support experimental designs with routing of specific user traffic. The DataSecOps, ModelSecOps, and DevSecOps processes and tools monitor specific metrics of risks and KPIs to automate or guide course corrections. The architecture/platform and processes take governance and security guidance of the NIST AI RMF and ZT framework into considerations to cap the downside risks. This helps to fulfill CSET's robustness, assurance, and specification AI safety criteria.

A series of small wins with concrete, complete, implemented outcome of moderate importance leads to incremental commitment and action. Small wins also attract allies as they don't appear to be highly risky and a zero-sum game, lowering resistance to subsequent proposals or fundings (Taylor, 2020). In financial analysis, it is rational to invest in projects with high expected payoffs. Accomplishing small wins while on course toward big win(s) is essentially realizing parts of the total expected payoff by turning the probabilities of these parts into certainties, effectively increasing the total expected payoff. This increases the chance that the organization would continue to fund the project.

While there is no one fixed formula that ensures the success of an AI project, the strategy, architecture/platform, and processes set forth provide the details and rationale for a path forward under the uncertainty of fast-paced technological advancement and changing regulatory compliance landscape, allowing the flexibility to adapt as the AI era unfolds.

REFERENCES

Bocklisch T, Werkmeister T, Varshneya D, Nichol A, (2024). Task-Oriented Dialogue with In-Context Learning. Retrieved from https://arxiv.org/pdf/2402.12234

Bright J, (2021). Dynamic A/B Testing for Machine Learning Models with Amazon SageMaker MLOps Projects. Retrieved from https://aws.amazon.com/blogs/machine-learning/dynamic-a-b-testing-for-machine-learning-models-with-amazon-sagemaker-mlops-projects/

Chattha T, Di Francesco P, Hwang J, (2023). Create a Generative AI Gateway to allow secure and compliant consumption of foundation models. Retrieved from https://aws.amazon.com/blogs/machine-learning/create-a-generative-ai-gateway-to-allow-secure-and-compliant-consumption-of-foundation-models/

D'Hoinne J, Litan A, Firstbrook P, (2023). 4 Ways Generative AI Will Impact CISOs and Their Teams. Retrieved from https://www.gartner.com/doc/reprints?id=1-2EI17OKR&ct=230719&st=sb

Gartner, (2024). Gartner Survey Finds Generative AI Is Now the Most Frequently Deployed AI Solution in Organizations. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations

Keller C, Babic M, Utsav A, de Jesus Assis A, Goehring B, (2024). AI Revolution: Productivity Boom and Beyond. Retrieved from https://www.ib.barclays/content/dam/barclaysmicrosites/ibpublic/documents/our-insights/AI-impact-series/ImpactSeries_12_brochure.pdf

Klarich K, Goldman B, Kramer T, Riley P, Walters W, (2024). Thompson Sampling─An Efficient Method for Searching Ultralarge Synthesis on Demand Databases. https://pubs.acs.org/doi/10.1021/acs.jcim.3c01790

Kohl J, Wöhlke A, Jones B, Mueller C, Engelhardt D, Zinovyeva L, Castro N, Saxena S, Khodjaev S, (2024). BMW Group Develops a GenAI Assistant to Accelerate Infrastructure Optimization on AWS. Retrieved from https://aws.amazon.com/blogs/industries/bmw-group-develops-a-genai-assistant-to-accelerate-infrastructure-optimization-on-aws/

Li, Fei Fei (2024). With Spatial Intelligence, AI Will Understand the Real World | Fei-Fei Li | TED. Retrieved from https://www.youtube.com/watch?v=y8NtMZ7VGmU

Merreider M, (2024). Bringing Large Language Models to Your Local Environment with Ollama: A Step-by-Step Guide. Retrieved from https://medium.com/@miroslavmerreider/bringing-large-language-models-to-your-local-environment-with-ollama-a-step-by-step-guide-to-e1389eb4b338

Miyake D, (2023). DataSecOps: Delivering Secure Data Products. Part 1/2. Retrieved from https://blog.devgenius.io/datasecops-delivering-secure-data-products-part-1-2-1c9bc273c12c

Mordor Intelligence (2024). Global Chatbot Market. Retrieved from https://www.mordorintelligence.com/industry-reports/global-chatbot-market

NIST (National Institute of Standards and Technology), (2023). NIST AI 100-1 Artificial Intelligence Risk Management Framework (AI RMF 1.0). Retrieved from https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

PagerDuty (2024). Fortune 1000 Technology Executives Maintain a Distrust of GenAI – Security Risks & Ethics Keep Them up at Night https://www.pagerduty.com/assets/whitepaper-generative-ai-survey.pdf

Rudner T, Toner H, (2021). Key Concepts in AI Safety: An Overview. Retrieved from https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-an-overview/

Saner M, Lapidakis M, (2023). Securing Generative AI: an Introduction to the Generative AI Security Scoping Matrix. Retrieved from https://aws.amazon.com/blogs/security/securing-generative-ai-an-introduction-to-the-generative-ai-security-scoping-matrix/

Singh R, (2023). Top DevSecOps Tools for 2023 Open Source Solutions for Enterprises. Retrieved from https://ranjaniitian.medium.com/top-devsecops-tools-for-2023-open-source-solutions-for-enterprises-7c146f80b325

Syed N, Shah S, Shaghaghi A, Anwar A, Baig Z, Doss R, et al, (2022). Zero Trust Architecture (ZTA): A Comprehensive Survey. Retrieved from https://ieeexplore.ieee.org/document/9773102

Taylor B, (2020). To Solve Big Problems Look For Small Wins. Retrieved from https://hbr.org/2020/06/to-solve-big-problems-look-for-small-wins

Wang J, Chen L, Khare A, Raju A, Dheram P, He D, Wu M, Stolcke A, Ravichandran V, (2024). Turn-Taking and Backchannel Prediction with Acoutic and Large Language Model Fusion. Retrieved from https://arxiv.org/pdf/2401.14717

Wilkinson L, (2024). Enterprises struggle to show the value of AI projects. Retrieved from https://www.ciodive.com/news/generative-ai-adoption-barrier-project-value/716504/

Zahra A, (2019). MLOps Explained. Retrieved from https://www.c-sharpcorner.com/blogs/mlops