Scalable Conversational AI Architecture for Financial Services: A Case Study in Mortgage Industry Digital Transformation

Jul 10th, 2025

Venkata Santosh Sajjan Alla
Sr. Solutions Architect,
AWS Financial Services

Axel Larrson
Principal Solutions Architect,
AWS Financial Services

Manali Sapre
Sr. Director AI Engg,
Rocket Mortgage

Seshidhar Raghupathi
Software Architect,
Rocket Mortgage

Abstract

This paper presents the design, implementation, and evaluation of a scalable multi-agent conversational AI architecture deployed in production at a major financial technology company serving the mortgage industry. Our solution leverages Agentic AI to create a unified, domain-specific multi-agent system capable of real-time customer support, lead qualification, and transaction facilitation across web and mobile platforms. The architecture employs eight specialized agents orchestrated through a unified API interface, incorporating knowledge bases, action groups, and intelligent routing mechanisms. Performance evaluation demonstrates significant operational improvements including a 300% increase in conversion rates from web traffic to closed loans, 85% reduction in customer care escalations, 45% reduction in servicing specialist transfers, and 68% customer satisfaction ratings. Key technical contributions include a modular agent architecture design, cross-region inference implementation for scalability, and practical insights for production deployment of agentic AI systems in regulated financial environments. Our findings provide actionable guidance for financial services organizations implementing conversational AI at enterprise scale, demonstrating how cloud-native multi-agent architectures can transform customer engagement while maintaining regulatory compliance and operational efficiency.

Introduction

The digital transformation of financial services has fundamentally altered customer expectations for service delivery, with consumers increasingly demanding immediate, personalized, and contextually aware support across all interaction channels. In the mortgage and home financing sector, this challenge is particularly acute due to the complexity of financial products, regulatory requirements, and the emotionally significant nature of home ownership decisions. Traditional customer service models, while effective for simple inquiries, struggle to scale when faced with the multifaceted nature of mortgage origination, servicing, and refinancing processes that often require deep domain expertise and real-time access to customer data and product information.

The emergence of large language models (LLMs) [1] and conversational AI[2] technologies has created unprecedented opportunities to reimagine customer service delivery in financial services. However, the implementation of AI-driven customer support systems in regulated industries presents unique challenges including data privacy requirements, accuracy demands, and the need for seamless integration with existing enterprise systems. While general-purpose chatbots have shown promise in simple customer service scenarios, the mortgage industry's complex workflows, regulatory constraints, and high-stakes decision-making processes require more sophisticated approaches. Recent advances in agentic AI frameworks[3], have introduced capabilities that extend beyond traditional conversational interfaces to include reasoning, planning, and action execution. These systems employ the Reasoning and Acting (ReAct) framework to interpret user intent, decompose complex tasks into actionable steps, and interact with backend systems to complete transactions and retrieve personalized information. However, limited research exists on the practical implementation of such systems in production financial services environments, particularly regarding architecture design decisions, performance optimization, and operational considerations for enterprise-scale deployment.

The mortgage industry presents a compelling use case for advanced conversational AI due to several factors: (1) the lengthy and complex nature of the home buying process, (2) the need for 24/7 availability during time-sensitive market conditions, (3) the requirement for personalized guidance based on individual financial situations, and (4) the opportunity to streamline operations while maintaining regulatory compliance. Despite these potential benefits, few comprehensive studies have examined the end-to-end implementation of multi-agent conversational AI systems in this domain, particularly with quantified business impact metrics and operational lessons learned.

This paper addresses this gap by presenting a comprehensive case study of a production multi-agent conversational AI system deployed at Rocket Companies, a leading financial technology company in the mortgage industry.

Overview

This article provides a structured overview of the best practices for building secure, lightweight, and reliable container images. It explores the fundamental advantages of adopting microservices architecture and containerization in modern application development, focusing on how these technologies facilitate scalability, faster development cycles, and technological autonomy. The article delves into critical aspects of container image management, including the selection of trusted base images, the importance of regular updates, and the implementation of container image signing for enhanced security. It also addresses strategies for optimizing image size and performance, such as limiting the number of layers, employing multi-stage builds, and utilizing minimal base images. Furthermore, the article underscores the significance of secure secrets management and reducing the attack surface of container images. It introduces Amazon Elastic Container Registry (Amazon ECR) as a robust, secure, and fully managed solution for storing and managing container images, highlighting its features for image lifecycle management, replication, and vulnerability scanning. Through this comprehensive guide, readers will gain

insights into constructing container images that not only meet the demands of modern application deployment but also adhere to the highest standards of security and efficiency.

We'll look at the following approaches to build better images. While this isn't an exhaustive list, these topics provide a good base for your image builds, and you can adopt them as you wish.

Business Context and Challenge

Rocket Companies is a Detroit-based financial technology company with a mission to "Help Everyone Home." Beyond mortgage lending, Rocket's services span the entire home ownership journey including home search, purchasing, financing, and home equity utilization. The company has achieved growth by simplifying complex processes and empowering clients through intuitive, technology-driven solutions. Rocket's integrated web and mobile applications combine home search, financing, and servicing capabilities in a unified experience. Through data analytics leveraging 11 petabytes of data combined with advanced automation, Rocket accelerates processes from loan approval to servicing while maintaining personalized client interactions at scale.

The advent of generative AI[4] presented Rocket with an opportunity to address a persistent challenge: home buying remains an overwhelming experience for many clients. This led to a fundamental question - how can the company provide the same trusted guidance clients expect at any hour, across any communication channel? The home financing process involves complex decisions, time-sensitive market conditions, and requires access to personalized financial information, making traditional chatbot solutions inadequate for the sophisticated support requirements.

The challenge required a solution that could deliver 24/7 multilingual assistance, maintain contextual awareness across interactions, provide real-time answers about mortgage options and processes, enable guided self-service actions, and seamlessly transition to human support when necessary. The solution needed to integrate with Rocket's existing digital infrastructure while meeting regulatory compliance requirements for financial services.

Solution Overview

Amazon Bedrock Agents[5] is a fully managed, cloud-based capability that enables organizations to build, test, and scale agentic AI applications on Amazon Web Services (AWS)[6]. The platform provides built-in integrations and security features that accelerate development from proof-of-concept to production deployment. These agents extend foundation models[7] using the Reasoning and Acting (ReAct) framework, enabling them to interpret user intent, plan and execute tasks, and integrate with enterprise data and APIs.

The agents utilize foundation models to analyze user requests, decompose them into actionable steps, retrieve relevant data, and trigger downstream APIs to complete tasks. This capability enables movement beyond passive support into proactive assistance, helping clients navigate complex financial processes in real-time.

Key capabilities of Amazon Bedrock Agents implemented in the solution include:

Agent Instructions: Define the agent's objective and role (e.g., mortgage servicing expert), enabling goal-oriented behavior aligned with specific domain expertise.

Amazon Bedrock Knowledge Bases: Provide fast, accurate retrieval of information from proprietary documents

and learning resources, ensuring responses are grounded in authoritative content.

Action Groups: Define secure operations such as lead submission or payment scheduling that agents can execute through interaction with backend services.

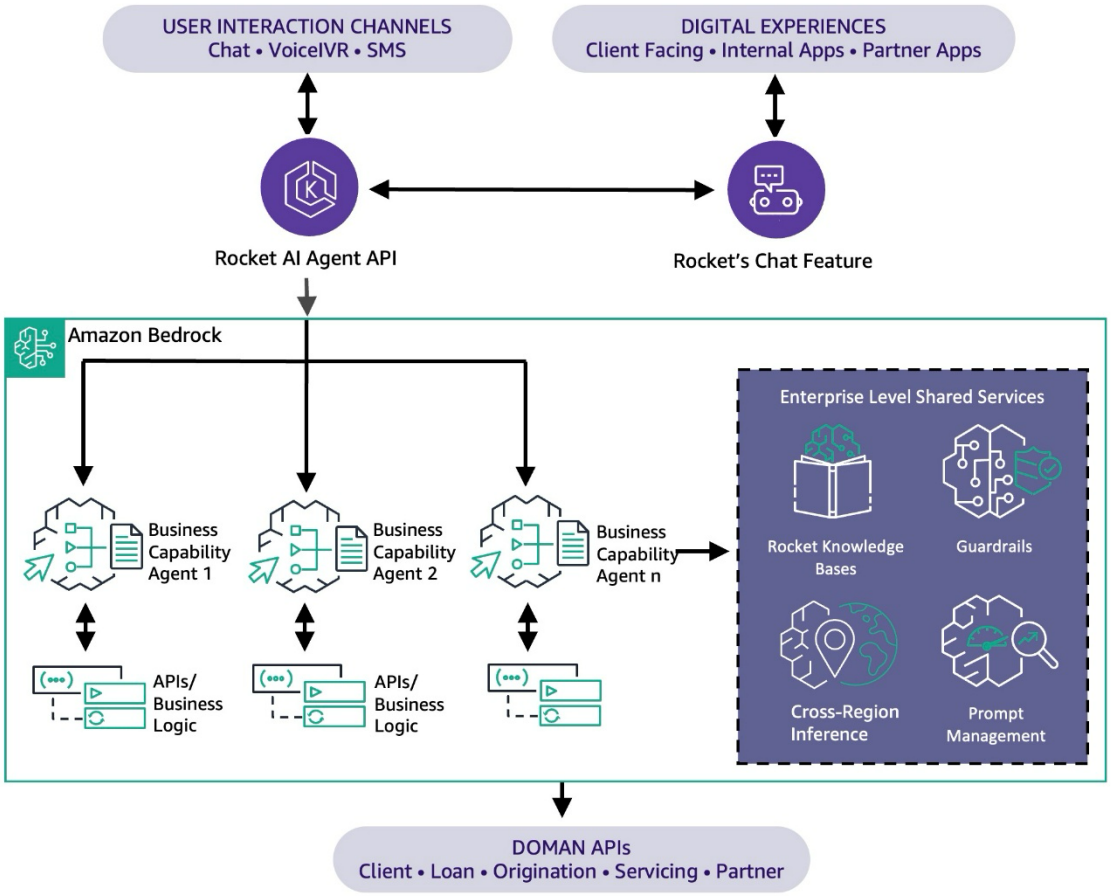
Agent Memory: Memory retention capabilities allow agents to maintain contextual awareness across multiple conversation turns, enhancing user experience through more natural, personalized interactions.

Amazon Bedrock Guardrails: Support responsible AI implementation by ensuring agents operate within appropriate topic boundaries and maintain compliance with organizational policies.

The combination of structured reasoning with cross-system action capabilities enables agents to deliver measurable outcomes rather than merely providing information responses.

Architecture and Implementation

The Rocket AI Agent represents a centralized capability deployed across Rocket's digital properties, designed for scale, flexibility, and domain-specific precision. The architecture core consists of eight domain-specific agents, each focused on distinct functions such as loan origination, servicing, and broker support. These agents operate collaboratively behind a unified interface to provide seamless, context-aware assistance.



Here are three foundational elements that shape Rocket AI Agent's architecture:

- **Client Initiation:** Clients access the chat functionality through Rocket's mobile applications or web platforms, providing multiple entry points for user engagement.
- **Rocket AI Agent API:** A unified API interface manages agent interactions and provides consistent functionality across all chat implementations.
- **Agent Routing:** The AI Agent API directs requests to the appropriate Amazon Bedrock agent based on static criteria such as the originating web or mobile property, or through LLM-based intent identification for dynamic routing decisions.
- **Agent Processing:** Individual agents decompose tasks into subtasks, determine optimal execution sequences, and coordinate actions with knowledge retrieval processes.
- **Task Execution:** Agents utilize Rocket's proprietary data through knowledge bases to retrieve information, deliver results to clients, and perform actions to complete requested tasks.
- **Guardrails Implementation:** Enforce responsible AI[8] policies by restricting topics and language that deviate from experience objectives, ensuring compliance and appropriate interaction boundaries.
- **Prompt Management:** Enables management of a comprehensive prompt library for AI agents and optimization of prompts for specific foundation models.

This modular, scalable architecture design enables efficient and consistent service delivery across diverse client needs throughout the homeownership lifecycle. The system maintains separation of concerns while providing unified user experiences across multiple Rocket digital properties.

Results and Business Impact

The implementation of Rocket AI Agent has demonstrated significant improvements across client journey metrics and internal operational efficiency:

Conversion and Lead Generation

The deployment resulted in a threefold increase in conversion rates from web traffic to closed loans. The AI agent's 24/7 availability enables lead capture during non-traditional business hours, significantly expanding the window for client engagement and qualification.

Operational Efficiency

Analysis of chat containment metrics revealed substantial efficiency gains. The AI assistant implementation for prospective client support resulted in an 85% decrease in transfers to customer care and a 45% decrease in transfers to servicing specialists. This reduction in human agent handoffs has freed team capacity to focus on complex, high-impact client interactions requiring specialized expertise.

Customer Satisfaction

Customer satisfaction (CSAT) scores demonstrated strong positive reception, with 68% of clients providing high satisfaction ratings across servicing and origination chat interactions. Primary satisfaction drivers include rapid

response times, clear communication, and information accuracy, contributing to enhanced client trust and reduced process friction.

Client Engagement

The system has driven stronger client engagement through intuitive, personalized self-service capabilities, with users completing more tasks independently. The AI agents adapt to each client's position in the homeownership journey and their preferences, offering escalation to human bankers on client terms.

Service Accessibility

Expanded language support, including Spanish-language assistance, has improved service accessibility for diverse client demographics, supporting Rocket's mission to "Help Everyone Home" through inclusive service delivery.

System Integration

Rocket has successfully deployed AI agents across digital services including the servicing portal and third-party broker systems, providing continuity of experience across all client engagement touchpoints. This consistent, on-brand support delivery transforms the client homeownership experience through scalable, intelligent engagement capabilities.

Technical Insights and Lessons Learned

The design, development and deployment process revealed several critical insights that shaped both technical strategy and overall client experience. These findings provide guidance for organizations implementing generative AI applications at enterprise scale.

Data Quality and Curation

The quality of generative AI responses correlates directly with source data quality and structure. The implementation utilized Amazon Bedrock Knowledge Bases, which internally leverages Amazon Kendra[9] for retrieval across content libraries including FAQs, compliance documents, and servicing workflows. Careful curation of enterprise knowledge bases proved essential for maintaining response accuracy and relevance.

Agent Scope Optimization

Analysis revealed that assigning each agent a focused scope of 3-5 actions resulted in more maintainable, testable, and high-performing systems. For example, the payment agent focuses exclusively on tasks such as payment scheduling and due date provision, while the refinance agent handles rate simulations and lead capture. Each agent's capabilities utilize Amazon Bedrock action groups with well-documented interfaces and separately monitored task resolution rates.

Graceful Escalation Strategy

Escalation mechanisms represent a critical component of user trust rather than system failure. The implementation incorporated uncertainty thresholds using confidence scores and specific keyword triggers to detect interactions requiring human assistance. In such cases, the AI agent proactively transitions sessions to live support agents or

provides users with escalation options, avoiding frustrating conversational loops and ensuring complex or sensitive interactions receive appropriate human attention.

User Behavior Evolution

Real-world usage patterns prove dynamic, with clients interacting with systems in unexpected ways and evolving interaction patterns over time. Investment in observability and user feedback mechanisms proved essential for rapid adaptation to changing usage patterns and requirements.

Cross-Region Inference Implementation

To provide scalable, resilient model performance, cross-Region inference[10] was enabled early in development. This architecture allows inference requests to route to optimal AWS Regions within supported geography, improving latency and model availability through automatic load distribution based on capacity. During peak traffic periods such as product launches or interest rate fluctuations, this architecture prevented regional service quota bottlenecks, maintained system responsiveness, and increased throughput by utilizing compute capacity across multiple AWS Regions[11]. The result is consistent user experience under variable and unpredictable load conditions.

These insights demonstrate that while generative AI enables powerful capabilities, thoughtful implementation remains essential for delivering sustainable value and maintaining trusted user experiences.

Future Developments

The current implementation represents the initial phase of agentic AI potential realization at Rocket. Building on domain-specific agent success, the next development phase focuses on scaling capabilities through multi-agent collaboration[12] powered by Amazon Bedrock Agents. This evolution will enable agent orchestration across domains to deliver intelligent, end-to-end experiences that reflect the complexity of actual client journeys.

Multi-Agent Collaboration Benefits:

End-to-End Personalization: Multiple domain-specific agents (refinance, servicing, loan options) will share context and coordinate to deliver tailored, intelligent responses that evolve with client homeownership journeys in real-time.

Back-Office Integration: Agents capable of invoking secure backend APIs and workflows will enable automation of back-office operations including document verification, follow-up processes, and lead routing, improving speed, accuracy, and operational efficiency.

Context Switching: Fluid movement between servicing, origination, and refinancing functions within single chat sessions will provide seamless user experiences across business domains.

Workflow Orchestration: Capability to handle multistep tasks spanning multiple Rocket business units will enable comprehensive process automation and client support.

Multi-agent orchestration establishes the foundation for consistently available, deeply personalized assistance that advances beyond question answering to drive meaningful outcomes from home search through loan closing and

beyond. This represents the next phase in Rocket's mission to "Help Everyone Home" through advanced AI-driven client engagement.

Conclusion

In this post, The Rocket AI Agent implementation demonstrates a successful transformation of client engagement through agentic AI technology. By combining Amazon Bedrock Agents with proprietary data and backend system integration, the solution delivers scalable, intelligent, and accessible client support available 24/7 without traditional service limitations.

The quantified results including threefold conversion rate increases, 85% reduction in customer care escalations, and 68% customer satisfaction ratings - validate the business impact of thoughtfully implemented conversational AI in financial services. The technical insights gained, particularly regarding data curation, agent scope optimization, and cross-region scalability, provide actionable guidance for similar implementations in regulated industries.

This implementation serves as a comprehensive example of how cloud-native multi-agent architectures can transform customer engagement while maintaining regulatory compliance and operational efficiency. The lessons learned and architectural approaches documented provide a foundation for financial services organizations pursuing AI-driven customer service transformation at enterprise scale.

References

[1] Large language models, also known as LLMs, are very large deep learning models that are pre-trained on vast amounts of data. https://aws.amazon.com/what-is/large-language-model/

[2] Conversational artificial intelligence (AI) is a technology that makes software capable of understanding and responding to voice-based or text-based human conversations. https://aws.amazon.com/what-is/conversational-ai/

[3] Agentic AI marks the evolution from reactive assistants to proactive, autonomous systems that can understand, decide, and act with minimal oversight. https://aws.amazon.com/ai/agentic-ai

[4] Generative artificial intelligence (generative AI) is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music. https://aws.amazon.com/what-is/generative-ai/

[5] Amazon Bedrock Agents uses the reasoning of foundation models (FMs), APIs, and data to break down user requests, gathers relevant information, and efficiently completes tasks freeing teams to focus on high-value work. https://aws.amazon.com/bedrock/agents/

[6] Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—use AWS to be more agile, lower costs, and innovate faster. https://aws.amazon.com/

[7] Trained on massive datasets, foundation models (FMs) are large deep learning neural networks that have changed the way data scientists approach machine learning (ML). https://aws.amazon.com/what-is/foundation-models/

[8] Responsible AI is an approach to developing, deploying, and using AI systems that align with ethical principles and societal values. https://aws.amazon.com/ai/responsible-ai/

[9] Amazon Kendra, an intelligent enterprise search service, to traditional search solutions. We outline how Amazon Kendra's ML models, accuracy, and ease of use make it easier for customers and employees to find the information they need when they need it. https://aws.amazon.com/kendra/

[10] Cross-Region inference automatically selects the optimal AWS Region within your geography to process your inference request. This improves customer experience by maximizing available resources and model availability. https://docs.aws.amazon.com/bedrock/latest/userguide/cross-region-inference.html

[11] A named set of AWS resources that's in the same geographical area. A Region is comprised of at least three Availability Zones. AWS Regions are divided into partitions. https://docs.aws.amazon.com/glossary/latest/reference/glos-chap.html#region

[12] Multi-agent collaboration in Amazon Bedrock is a capability that allows multiple specialized AI agents to work together under the coordination of a supervisor agent to handle complex multi-step tasks that require different areas of expertise. https://aws.amazon.com/blogs/aws/introducing-multi-agent-collaboration-capability-for-amazon-bedrock/