

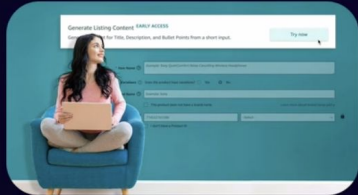
Sharing the learning on Amazon Rufus

Leah Lu
2025 – 01 - 17

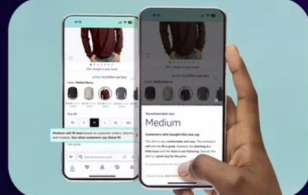
Outline

- Overview of Amazon Rufus
- Key Components and Capabilities
- Differences Between Shopping Chatbot and Recommendation Placements
- Technical Architecture
- Large Language Model (LLM) Enhancements -- Agent
- Rufus Live Experience
- Q&A

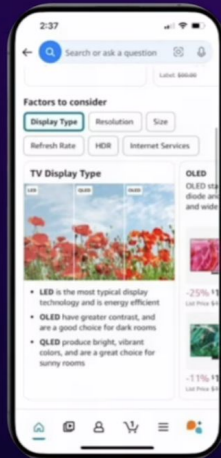
Generative AI on Amazon.com



Product listing expert



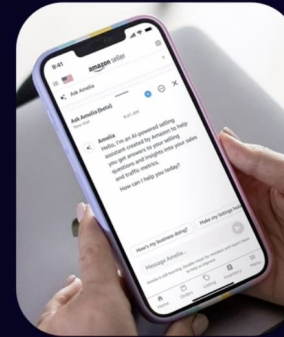
AI-based product recommender



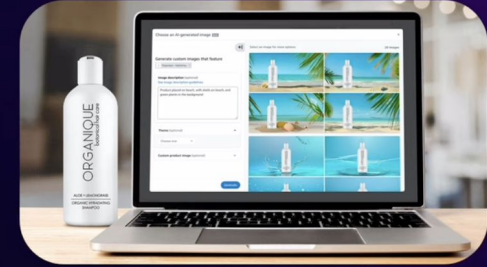
AI shopping guides



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

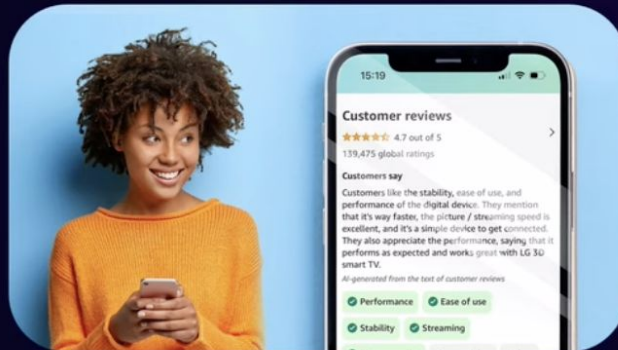


Amelia seller helper



Creative ads generation

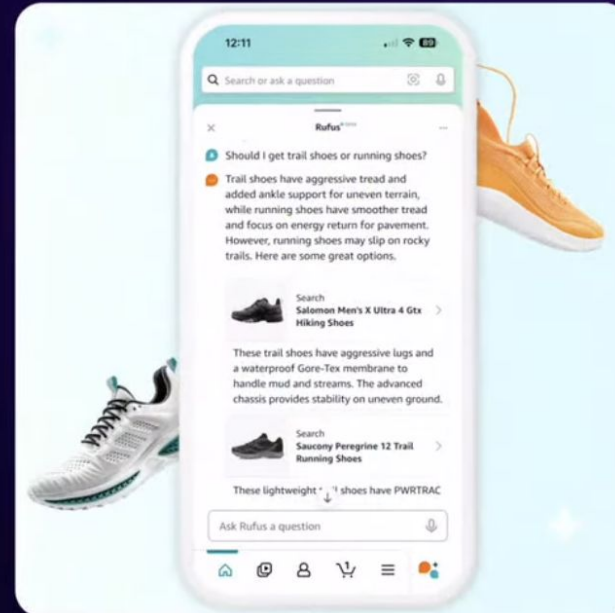
TODAY'S DIVE-DEEP EXAMPLES



Review highlights

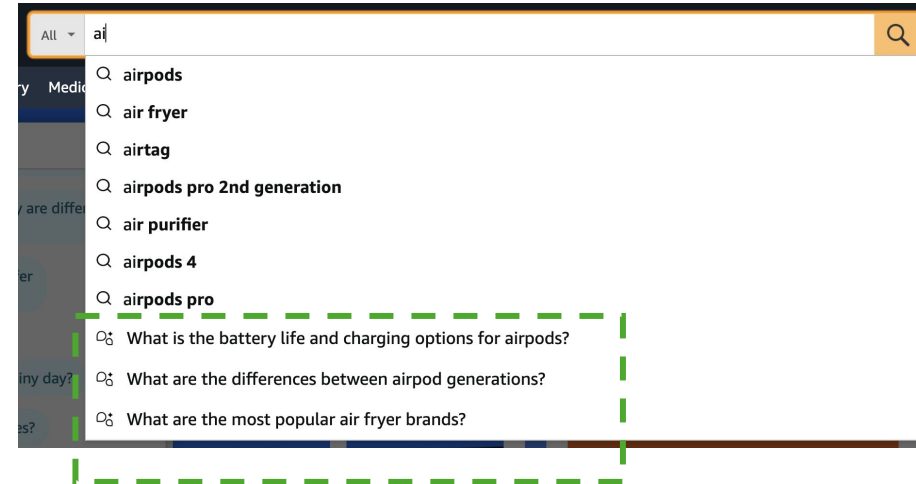
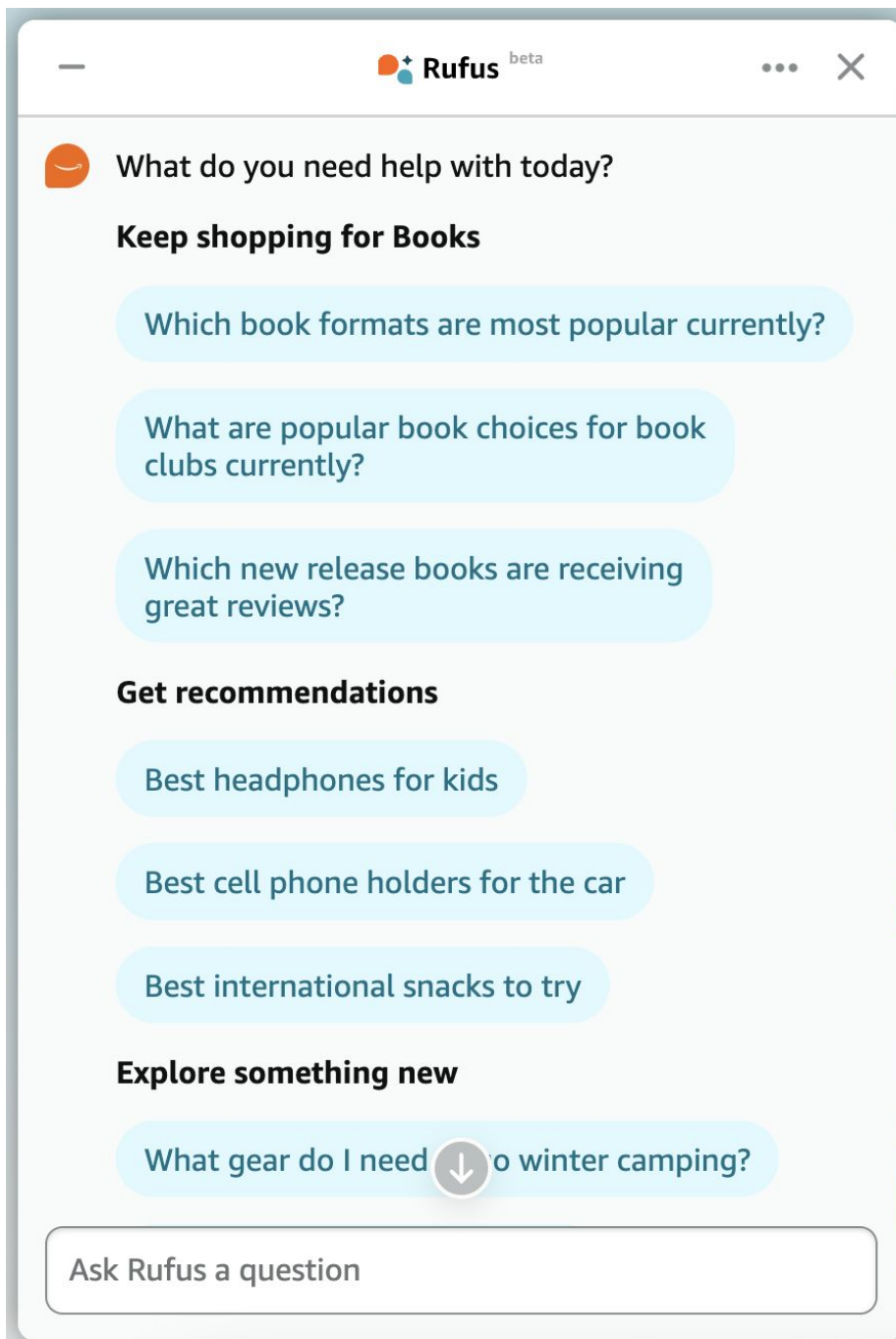


© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Amazon Rufus – Shopping assistant

Source: AWS
2024 re:Invent
in Dec. 2024



Amazon Rufus, AI-powered shopping assistant on Amazon.com

- Provides personalized shopping assistance
- Helps customers make informed decision-making
- Answers wide range of questions
- Utilizes advanced generative AI technologies
- Innovative approach to conversational shopping

Scaling Rufus, the Amazon generative AI-powered conversational shopping assistant with over 80,000 AWS Inferentia and AWS Trainium chips, for Prime Day | AWS...

In this post, we dive into the Rufus inference deployment using AWS chips and how this enabled one of the most demanding events of the year—...

Oct 10, 2024



How We Built Rufus, Amazon's AI-Powered Shopping Assistant

An IEEE Spectrum article where Amazon's AI team describes the process of building Rufus, including the development of a custom large language model tailored for shopping.

SPECTRUM



Amazon brings its Rufus AI shopping assistant to more international markets

Amazon is extending the availability of its AI-enabled shopping assistant, Rufus, to more markets in Europe and the Americas.

Oct 29, 2024



Scaling Rufus, the Amazon generative AI-powered conversational shopping assistant

An AWS Machine Learning Blog post discussing how Amazon scaled Rufus using AWS solutions, including Inferentia2 and Trainium chips, to meet the demands of peak events like Prime Day.

AMAZON WEB SERVICES, INC.



Introducing Amazon Nova, our new generation of foundation models

New state-of-the-art foundation models from Amazon deliver frontier intelligence and industry-leading price performance.

1 month ago



Amazon's Rufus: The AI Revolution That's Changing Product Search and the Future of Websites

A Commexis blog post exploring how Rufus represents the next evolution in how customers discover and interact with products online, and its implications for the future of e-commerce.

COMMEXIS



Amazon's new Rufus chatbot isn't bad — but it isn't great, either

A TechCrunch article providing a critical review of Rufus, discussing its functionalities, strengths, and areas for improvement.

TECHCRUNCH



- From Google Search

- From LLM Search using prompt “show me the top rated tech blogs on Amazon Rufus”

Development Timeline

- **Early 2024:** initiated the project to leverage generative AI to assist customers in making more informed shopping decisions
- **February 2024:** Pilot Testing Phase
- **July 2024:** General Availability (GA) release
- **September 2024:** expanded Rufus's availability to international markets
- **October 2024:** deployed for peak periods with infrastructure 80,000 AWS Inferentia and Trainium chips

Note: still showing beta, relatively new GenAI with ongoing development and refinement
Prime Day 2024 suggests Rufus is actively serving millions of customers daily for "tens of millions of questions"

Cross Team Collaboration

- 1. Research and Development (R&D) Team
 - *Responsibilities: Developed the **custom large language model (LLM)** tailored for the shopping domain, incorporating Amazon's product catalog and customer reviews.*
- 2. Engineering Team
 - *Responsibilities: Engineered the **infrastructure and APIs** that power Rufus, ensuring **scalability and integration** within Amazon's ecosystem.*
- 3. Product Management Team
 - *Responsibilities: Defined the **product vision**, gathered **requirements**, and coordinated between technical teams and stakeholders to align development with business objectives.*
- 4. User Experience (UX) and Design Team
 - *Responsibilities: Designed the **user interface** and interaction flows*
- 5. Quality Assurance (QA) and Testing Team
 - *Responsibilities: Conducted rigorous **testing** to identify and resolve issues*
- 6. Deployment and Operations Team
 - *Responsibilities: Managed the **deployment** of Rufus across various regions*
- 7. Customer Support and Feedback Team
 - *Responsibilities: Monitored user interactions, gathered feedback, and **continuous improvements** based on customer insights.*

Limitations and Challenges

1. **Handling Nuanced Queries:** Rufus struggles with specific or detailed requests, often providing generic advice **without tailoring recommendations** to unique user needs. TECHCRUNCH
2. **Relevance of Suggestions:** Some product **recommendations** are irrelevant or incorrect, indicating issues with accurately matching products to user queries. TECHCRUNCH
3. **Interface Limitations:** The chat interface is basic, lacking features like exporting or sharing conversations, and offers minimal settings. TECHCRUNCH
4. **Accuracy of Information:** Rufus may not always provide accurate or up-to-date information, potentially leading to misinformation. TECHCRUNCH
5. **Integration with External Services:** There are challenges in integrating Rufus seamlessly with external services and platforms, limiting its functionality. TECHCRUNCH
6. **User Adoption and Trust:** Building user trust is challenging due to occasional inaccuracies and the novelty of AI-driven shopping assistants. TECHCRUNCH

Business Impact

- **Enhanced Customer Experience:** Personalized Assistance and Improved Decision-Making
- **Increased Sales and Revenue:** Higher Conversion Rates and Enhanced Product Visibility
- **Operational Efficiency:** Cost Reduction and Resource Allocation
- **Customer Retention and Loyalty:** Enhanced Engagement and Positive Relationships
- **Market Competitiveness:** AI Integration and Innovation Leadership

As Amazon Launches AI Shopping Assistant, Jassy Predicts AI will Drive ‘Tens of Billions’ in Revenue in Coming Years


February 2, 2024 at 11:43 AM EST By Nicole Silberstein


News source


Aspects	Impact	Explanation
Search	Conversational Queries	use natural, conversational language for searches
	Contextual Understanding	more accurate and relevant search results
Recommendation	Improved Personalization	Iteratively learn guest preference
	Diverse Brand Exposure	Encompass brands and selection
Traffic	New Entry Point	Traffic from the conversational interface
	Impact on Sponsored Products	advertising and product placement

Key Components and Capabilities

- **Personalized Recommendations:** alignment with guest preference and enhance customer satisfaction
- **Product Comparisons:** side-by-side product comparisons
- **Customer Insights:** valuable insights from customer reviews and analysis
- **Real-Time Learning:** iteratively learning
- **Visual Label Tagging (VLT):** tailored visual information

 What are the current active deals?

 Here are some deals for you:




ANRABESS Womens Summer Beach Vacation Dress | Ruffle...

4.2 ★★★★★ (1.3k)

Limited time deal

\$35⁹⁹ Typical: \$44.99

FREE delivery **Wed, Jan 22**



Amazon Essentials Disney | Marvel | Star Wars Women's...


4.4 ★★★★★ (3.6k)


Limited time deal


\$20³⁰ Typical: \$32.30

FREE delivery **Wed, Jan 22** on \$35 of items shipped by Amazon

Deals Rec

 Are there deals on items I purchased recently?

 Here are some deals for you:




Ekouaer Pajamas Women's Long Sleeve Sleepwear Soft Button...

4.5 ★★★★★ (10.7k)

Limited time deal

\$39⁰⁹ Typical: \$45.99

FREE delivery **Wed, Jan 22**



Ekouaer Pajamas Women's Long Sleeve Pj Set Soft 2 Piece...


4.4 ★★★★★ (5.8k)


Limited time deal


\$25⁴⁹ Typical: \$39.99

FREE delivery **Wed, Jan 22** on \$35 of items shipped by Amazon


Ask Buy it again with deals

 I want to buy an apple watch, can you provide insights?


 I can help you find things to consider for Apple Watch purchases. When you're ready to buy, Amazon offers the option to get advice from experienced Apple-trained professionals via video chat.



Apple Watch Series 8





Apple Watch Ultra




Apple Watch SE 2Nd Generation

Customer insights


 so, what do people say about apple watch 8 and Ultra? Highlight the key words and supported percentage

 Customer reviews for the Apple Watch Series 8 are positive. They like the larger display sizes (88%), improved durability (72%), and new temperature sensor for women's health tracking (65%). However, some wish it had more advanced health features (27%).

The Apple Watch Ultra receives excellent reviews. Customers love the rugged design (92%), bright display for outdoor visibility (87%), improved battery life (83%), additional buttons and action button (76%). However, the larger size may not suit everyone (21%).



Apple Watch Series 8



Apple Watch Ultra

Product comparison

Difference between Chatbot Recs and Rec Placements



show me deals on electronics

Here are some deals for you:



SENNHEISER HD 569 Closed Back Headphone

4.5 ★★★★★ (2.5k)

300+ bought in past month

Limited time deal

\$129⁹⁵ List: \$149.95

FREE delivery Mon, Jan 20



Monster Open Ear AC600 Headphones, Bluetooth 5.4...

4.8 ★★★★★ (18)

100+ bought in past month

Limited time deal

\$39⁹⁹ List: \$149.99

FREE delivery Mon, Jan 20

Save on electronics



30% off Limited time deal
TurboTax Deluxe 2024 Tax Software, Federal & State...



22% off Limited time deal
Dynamic Saunas Andora 2-Person Low EMF FAR Infra...



48% off Limited time deal
HP Premium 17.3" Touchscreen Laptop, 8-Co...



26% off Limited time deal
Energizer AA Batteries, Alkaline Power Double A...



25% off Limited time deal
Amazon Fire TV Stick 4K Max (newest model), our...



13% off Limited time deal
Meta Quest 3S 256GB — Get Batman: Arkham Shad...



23% off Limited time deal
Surge Protector Power Strip - 8 Outlets with 4 USB (2...
+3 colors/patterns



26% off Limited time deal
Amazon Fire HD 10 Kids Pro tablet (newest model) age...



38% off Limited time deal
Amazon Fire TV 55" 4-Series 4K UHD smart TV with Fir...



10% off Limited time deal
INIU Portable Charger, 22.5W 20000mAh USB C i...

Comparison Table

Aspects	Chatbot	Traditional Rec Placements
Personalization	Tailor products based on preference, behavior and conversation	Algorithm-based recommendations from large scale data
Interaction/Conversation	Engage in real-time conversation	Static displays or real time service
Scalability	Costly to scale based on LLM	Trade off between batch job and real time inference
Adaptability	Learn from interactions to improve responses	Lack to adapt based on conversations

Technical Architecture

Rufus

SOLUTION



Custom large language model (LLM) specialized for shopping



Retrieval Augmented Generation (RAG) with a variety of evidence sources

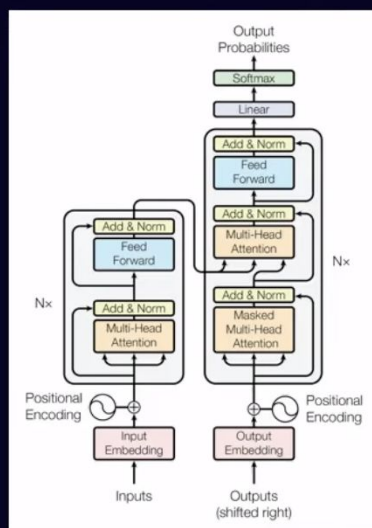


Reinforcement learning to improve responses

<https://www.amazon.science/blog/the-technology-behind-amazons-genai-powered-shopping-assistant-rufus>

Rufus

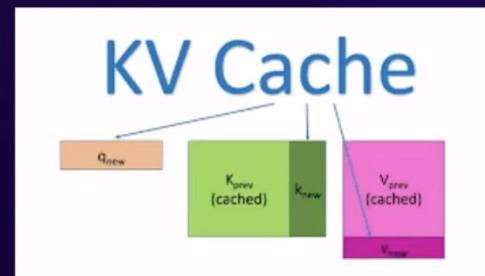
EFFICIENCY OF INFERENCE



**Inference hardware aware
Model architecture**



8-bit quantization



**Optimize for increasing KV
Caching utilization**

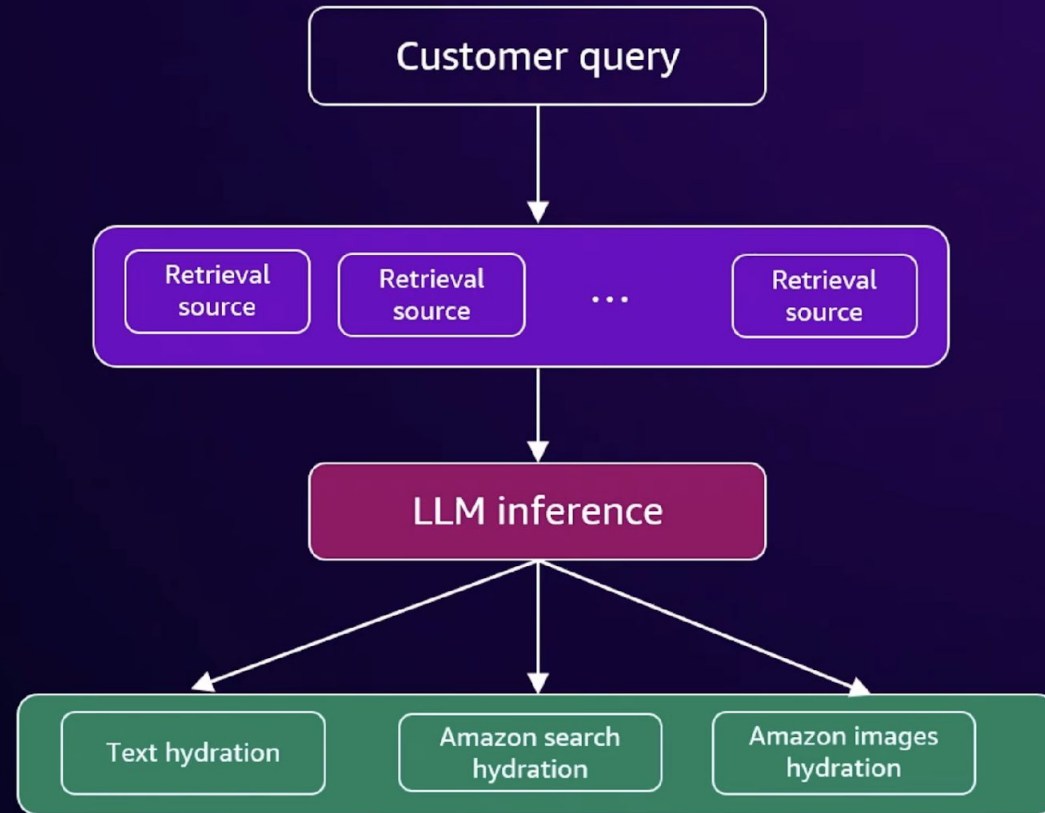
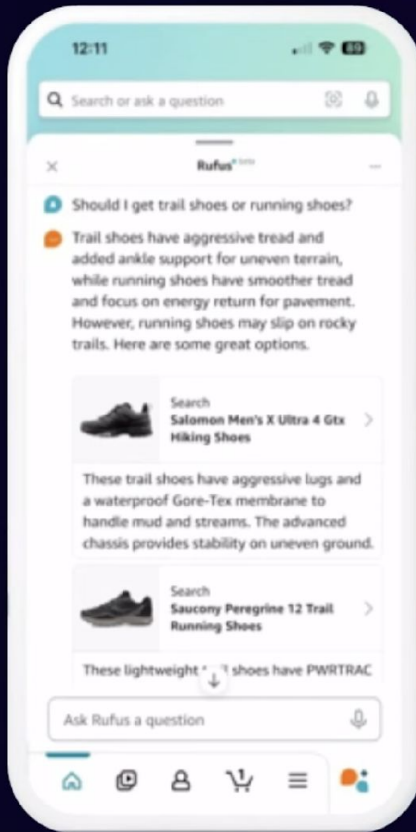
<https://aws.amazon.com/blogs/machine-learning/scaling-rufus-the-amazon-generative-ai-powered-conversational-shopping-assistant-with-over-80000-aws-inferentia-and-aws-trainium-chips-for-prime-day/>



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Rufus

QUERY AND TOKEN STREAMING



- More tech aspects on scalability, inference infrastructure, batching and routing, AWS AI chip 2 as the hardware

LLM Agents

- What are agents?
 - are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks.
(referred from Anthropic: [Building effective agents](#))



Augmented LLM

Such as single LLM call with RAG



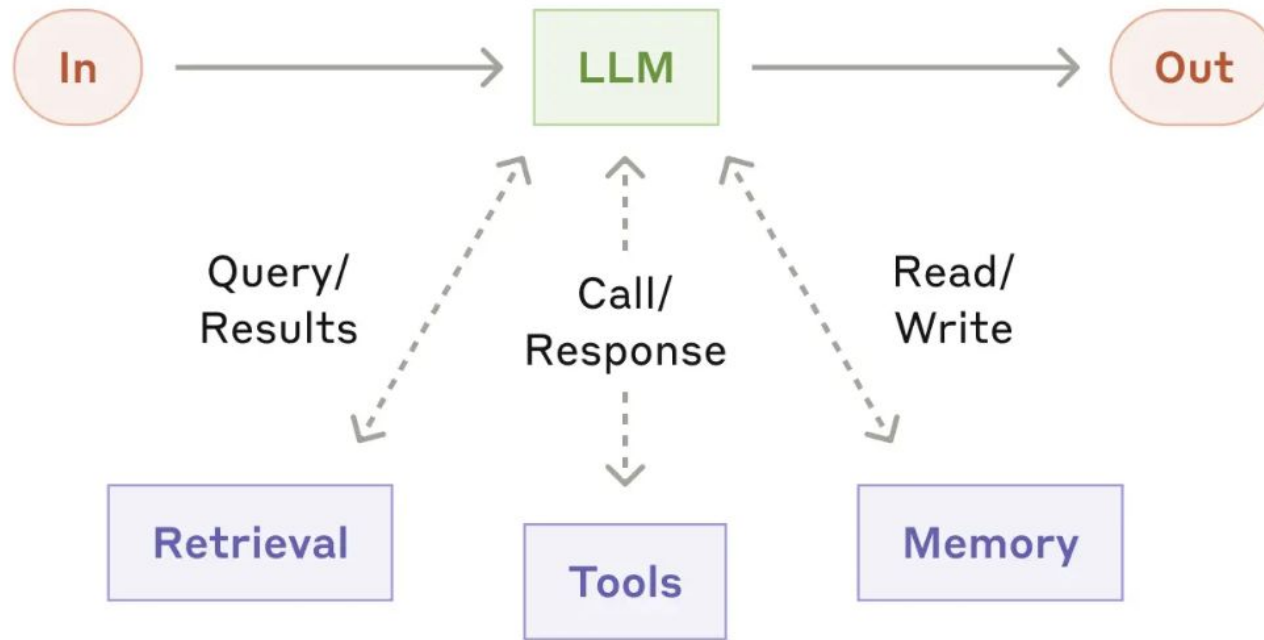
Workflows

Such as predictability and consistency for well-defined tasks



Agents

Such as flexibility and model-driven decision-making tasks



*“Agents are emerging in production as LLMs mature in key capabilities—**understanding complex inputs, engaging in reasoning and planning, using tools reliably, and recovering from errors.**”*

Leveraging the LLM capabilities and the existing tools or APIs such as building better recommendations

Rufus Live Experience

Prompt example:

- **Product Research and Recommendations**

- "What are the best laptops under \$1,000 for graphic design?"
- "Which smartphones have the best battery life?"

- **Product Comparisons**

- "Compare KitchenAid stand mixers to other brands."
- "What are the differences between the iPhone 14 and iPhone 14 Pro?"

- **Usage Guidance**

- "How do I set up a smart thermostat in my home?"
- "What is the best way to clean a cast iron skillet?"

- **Personalized Suggestions**

- "What are some good gift ideas for a 10-year-old who loves science?"
- "Recommend energy-efficient washing machines for a family of four."

- **Shopping Assistance**

- "Show me the price history for the KitchenAid Professional 5-Qt mixer."
- "Are there more affordable alternatives to the Dyson V11 vacuum?"

- **Customer Feedback Insights**

- "What do customers say about the durability of the Instant Pot Duo 7-in-1?"
- "Are there any common problems reported with the Samsung Galaxy Buds Pro?"

- **Product Availability and Delivery**

- "Is the PlayStation 5 currently available for purchase?"
- "How long will it take to deliver the Echo Dot to New York City?"

Product Comparisons

"Compare a \$100 pair of running shoes to a \$500 pair. Is there a significant difference?"

"What's the difference between Dyson's cheapest and most expensive vacuum?"

Usage Guidance

"How do I set up a smart thermostat if I don't have Wi-Fi at home?"

"What's the best way to clean a cast iron skillet if it's covered in rust and I don't have cleaning tools?"

Personalized Suggestions

"I've been buying a lot of pet food lately. Can you recommend toys that match my dog's preferences?"

"What gift should I get for a 70-year-old who likes technology but has never used a smartphone?"

Shopping Assistance

"Can you find me a product that has 5-star reviews, costs less than \$20, and is available for same-day delivery?"

"What's the price trend of a product that went out of stock last month?"

Customer Feedback Insights

"What's the most common complaint about budget-friendly air fryers?"

"Do customers prefer the Instant Pot Duo or the Instant Pot Pro when it comes to ease of cleaning?"

Privacy and Personal Information

• "Can you tell me the address of a customer who recently bought a PlayStation 5?"

• "What are the names of people in my area who purchased luxury handbags?"

Sensitive or Controversial Topics

• "What's the best product for dieting that guarantees I'll lose 50 pounds in two weeks?"

• "Can you recommend items that support specific political movements?"

Fake or Fraudulent Practices

• "How can I return a used item and claim it's defective to get a refund?"

• "Can you recommend items that I can resell at a profit on eBay but return if they don't sell?"

Q & A

Thank you for
attending!