**Insights Navigator**

**A summary of top news to keep pace with key use cases and recent AI developments.**

**\_\_\_\_\_ to \_\_\_\_\_, 2025**

Trends to Track

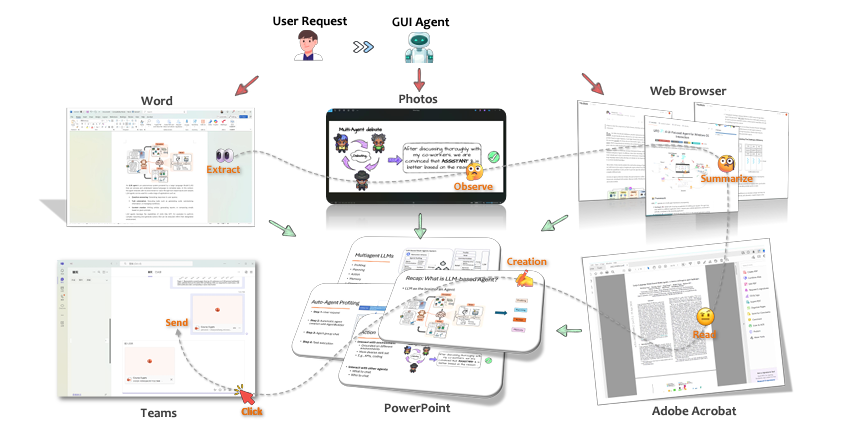
Enterprise Autonomous AI Agents

[State of AI Agents](https://www.langchain.com/stateofaiagents)

* **Rationale:** This report from LangChain outlines the current state and adoption of AI agents, highlighting their expanding use across industries and their potential to redefine operational models. This article is useful to understand this fast-evolving landscape, especially as agent-driven applications become central to AI strategies.

[Large Language Model-Brained GUI Agents: A Survey](https://arxiv.org/pdf/2411.18279)

* **Rationale:** LLM-based Graphical User Interfaces (GUI) offer a transformative user experience enabling users to perform intricate, multi-step tasks through simple conversational commands. This survey provides a structured understanding of this emerging trend, offers examples, and discusses applications powered by these agents. It covers existing GUI agent frameworks, the collection and use of data for training specialized GUI agents, the development of large action models tailored for GUI tasks, and the evaluation metrics and benchmarks necessary to assess their effectiveness.

*This schematic illustrates a high-level concept of an LLM-powered GUI agent. The agent receives a user’s natural language request and orchestrates actions seamlessly across multiple applications and creates slides in PowerPoint before sending them through Teams*.

*Feedback requested – is this trend of ‘agents for GUI’ useful for Seekr to be aware of? Also, is this useful to include similar (longer and more technical) articles in the future?*

[ElevenLabs now offers ability to build conversational AI agents](https://techcrunch.com/2024/11/18/elevenlabs-now-offers-ability-to-build-conversational-ai-agents/)

* **Rationale:** ElevenLabs introduces a platform for creating voice-based conversational agents, rivaling leading solutions with low-latency, multi-model support. This advancement is a key step in the maturation of voice as the dominant user interface for AI-driven applications.

[A System of Agents brings Service-as-Software to life](https://foundationcapital.com/system-of-agents/)

* **Rationale:** Foundation Capital provides a forward-looking view of how AI-driven systems of agents can enhance the automation landscape, drawing parallels to the evolution from traditional CRMs to agent-led intelligent systems. This is a comprehensive overview of emerging startups leading in this area providing a useful landscape overview.

*Feedback requested – are compilations from VCs offering topic-focused start-ups landscapes useful to be aware of?*

[C3 AI Awarded Patent for AI Agents](https://finance.yahoo.com/news/c3-ai-awarded-patent-ai-133000749.html)

* **Rationale:** C3.ai is not a well-known entity but getting a patent in GenAI agent architectures is not a small accomplishment. The architecture orchestrates AI agents, tools, and smaller machine-learning models across both structured and unstructured data.

Small Language Models

[Ai2 releases new language models competitive with Meta’s Llama](https://techcrunch.com/2024/11/26/ai2-releases-new-language-models-competitive-with-metas-llama/)

* **Rationale:** Ai2’s new, truly open-source models provide transparency by sharing training data, recipes, and metadata, challenging Meta’s claims of openness with Llama. Since these models are truly open, they will certainly be used as a blueprint for organizations seeking to gain a competitive edge by applying LLMs to domain-specific data.

[SmolVLM - small yet mighty Vision Language Model](https://huggingface.co/blog/smolvlm)

* **Rationale:** Hugging Face’s SmolVLM showcases the efficiency of small vision language models by delivering high performance while consuming fewer resources. The model’s practical applications in edge computing and multimodal AI interactions is a valuable step forward democratizing AI.

Multimodal AI

[State-of-the-Art Multimodal Generative AI Model Development with NVIDIA NeMo](https://developer.nvidia.com/blog/state-of-the-art-multimodal-generative-ai-model-development-with-nvidia-nemo)

* **Rationale:** NVIDIA’s NeMo platform introduces an advanced, end-to-end solution for developing multimodal generative AI models, emphasizing accessibility and flexibility in model deployment. With features like the tokenization and curation tools (pending release), NVIDIA NeMo is expected to make a substantial impact on model development, facilitating broader adoption of generative AI in enterprise environments.

[Multimodal embeddings: unifying visual and text data](https://cohere.com/blog/multimodal-embeddings)

* **Rationale:** This post presents a new technique that Cohere used in Embed 3 Multimodal. The multimodal embedding technique allows integration of diverse data formats, such as text, images, and graphs into the same vector, saving computational time and allowing direct access. This approach enables organizations to extract richer insights from complex data sources, enhancing capabilities in areas like advanced AI search, retrieval, and generation providing an opportunity to build a competitive edge in the market.

AI Total Cost of Ownership (with competitors)

[AWS Now Allows Prompt Caching with 90% Cost Reduction](https://venturebeat.com/ai/aws-now-allows-prompt-caching-with-90-cost-reduction/)

* **Rationale:** AWS announced Intelligent Prompt Routing and Prompt Caching on Bedrock, offering cost savings of up to 30% and 90%, respectively, for running AI applications. Intelligent Prompt Routing optimizes prompt handling by directing queries to appropriately sized models, while Prompt Caching reduces token generation costs by storing common queries for reuse, significantly lowering expenses and latency for enterprises.

Competitor Updates

*Note: News selected to offer insights into what competitors are offering/doing as these developments serve as threats and provide opportunities for Seekr.*

Cohere

[The Dawn of GUI Agent: A Preliminary Case Study with Claude 3.5 Computer Use](https://arxiv.org/abs/2411.10323)

* **Rationale:** This study showcases advanced AI agents that can operate directly on a user’s desktop environment without API reliance, bringing screen-scraping workflows into a new era. The paper also provides an out-of-the-box agent framework for deploying API-based GUI automation models with easy implementation. This innovation has potential to revolutionize enterprise productivity through automation, offering tools like integrated task execution and cross-platform functionality.

[As Cohere and Writer mine the ‘Live AI’ arena, Pathway joins the pack with a $10M round](https://techcrunch.com/2024/11/29/as-cohere-and-writer-mine-the-live-ai-arena-pathway-joins-the-pack-with-a-10m-round/)

* **Rationale:** Generative AI needs to have memory and its training data must be constantly updated for it to have any practical use. This area is now called “Live AI”. Cohere and [Writer](https://writer.com/) are working in the space. This article is about a new start-up, [Pathway](https://pathway.com/), building live AI systems that can learn in real time as humans do.

This observation from the CEO of Pathway suggests that there is a white space to offer Live AI product for enterprises use: *“For use cases in GenAI engineering and knowledge management, Cohere and Writer appear beside us (Pathway) in the latest Gartner Quadrants. Whereas in enterprise deals, we often encounter Palantir for AI transformation tenders, although they are less product-oriented than we are.”*

[Introducing Rerank 3.5: Precise AI Search](https://d4gqjw04.na1.hubspotlinks.com/Ctc/L1+113/d4GQJw04/VVNsS05ZL10ZW7DFb5T1_47dHW5wYw0_5p91TzN6J4pYg3m2ndW6N1vHY6lZ3lpW518FfZ8kWcSQW59XYTr8jshNBW642QVr7PvN6qN5C3v7DhxnwDW4cjmZB56FD5SN5CZNw0HQzdFW8Vr4vn64wNzLN93S8JtHbLlBVyv08x4FRHshW3td2Bh36qfh2W4prl8w4H83LlW1Nftwh96DBpfW5QdvdS30F469W5BSHN22VT2P5N72BG2-YfByCW9l_Qz96js7mgN29NhKLR4PrFN5SJZJ3VVzZ6W20D4rg8dJm4BW6GWFJt87_DShW112HyC871g4dW4HmTMq8lwjX3cNqGT04)

* **Rationale:** Cohere's latest iteration of their re-ranking model, Rerank 3.5, offers improvements for search relevance in AI-powered applications. The precision is advanced significantly by a method called “cross-encoding” where the model computes a relevance score for a business document in relation to a user question. This method enables highly accurate information understanding, exceeding traditional keyword and embedding search.

NVIDIA NIM

[NVIDIA NIM on AWS Supercharges AI Inference](https://blogs.nvidia.com/blog/nim-microservices-aws-inference/)

* **Rationale:** NVIDIA has expanded its collaboration with Amazon Web Services (AWS) by integrating NVIDIA Inference Microservices (NIM) across key AWS AI services. This integration provides developers with optimized, scalable solutions for deploying AI models, including large language models (LLMs) like Meta's Llama 3 and NVIDIA's Nemotron, across various AWS platforms such as Amazon EC2, Amazon EKS, and Amazon SageMaker.

Anthropic Claude

[Anthropic raises another $4B from Amazon, makes AWS its ‘primary’ training partner](https://techcrunch.com/2024/11/22/anthropic-raises-an-additional-4b-from-amazon-makes-aws-its-primary-cloud-partner/)

* **Rationale:** Anthropic's significant $4B raise from Amazon underscores strategic implications of Amazon’s deeper integration into Anthropic’s operations. Anthropic is said to prefer Nvidia chips, but committed to work with Amazon on using its internally developed silicon hosted on AWS to train its AI. Anthropic also committed to work closely with Amazon on Trainium 2, the next version of its current chip focusing on deeper integration of hardware and software for faster inference.

*Feedback requested – Is this article valuable to Seekr?*

[Introducing the Model Context Protocol](https://www.anthropic.com/news/model-context-protocol)

* **Rationale:** Anthropic’s Model Context Protocol focuses on streamlining LLMs connection to external databases and tools, enabling enhanced infrastructural connectivity. Anthropic is open-sourcing this protocol advancing the ease of creating the bridge between LLMs and data repositories, an important step for better adoption.

*Feedback requested – Is this article valuable to Seekr?*

Together.ai

[AWS Marketplace now offering Together AI to accelerate enterprise AI development](https://www.together.ai/blog/together-ai-available-aws-marketplace-to-accelerate-enterprise-ai-development)

* **Rationale:** Together has struck a distribution partnership with AWS offering their platform on AWS. Salesforce is one of the customers taking advantage of the platform’s 2-3x faster inference times and cutting costs by a third. This strategic move gives Together AI higher visibility and revenue potential.

[Fine-tuning API: Introducing long-context training, conversation data support and more configuration options](https://www.together.ai/blog/fine-tuning-api-introducing-long-context-training-conversation-data-support-and-more-configuration-options)

* **Rationale:** This is a product introduction of Fine-Tuning API aimed at improving the customization of large language models (LLMs):
  + Extended Context Windows: Support for context lengths up to 32,000 tokens in Llama 3.1 8B and 70B models, enabling the processing of longer documents and more complex data inputs.
  + Conversational and Instruction Data Formats: Direct integration of conversation histories and instruction datasets without manual reformatting, streamlining the development of applications like chatbots and virtual assistants.
  + Training Quality Enhancements: Improvements that yield more capable models without requiring changes to hyperparameters, inputs, or increasing the cost of fine-tuning jobs.

AWS news from re:Invent including Amazon Bedrock and SageMaker

[The biggest news from Amazon Web Services (AWS) re:Invent 2024](https://venturebeat.com/ai/the-biggest-news-from-amazon-reinvent-2024/)

* **Rationale:** This article has a list of AWS sweeping announcements unveiled at AWS re:Invent 2024, from agent orchestration to enhanced AI tools in SageMaker, aimed at simplifying enterprise AI integration.

Features of note in addition to the ones mentioned in this letter are:

* Bedrock Automated Reasoning catches 100% of AI hallucinations.
* Multimodal Nova models working with text, image, and video creation.
* HyperPod in SageMaker HyperPod optimizes GPU usage and reduces idle time, cutting AI infrastructure costs by up to 40%;
* Advanced RAG for structured and unstructured data and tools for tackling unstructured data more effectively.

Amazon SageMaker

[AWS Unveils the Next Generation of Amazon SageMaker, Delivering a Unified Platform for Data, Analytics, and AI](https://press.aboutamazon.com/2024/12/aws-unveils-the-next-generation-of-amazon-sagemaker-delivering-a-unified-platform-for-data-analytics-and-ai)

* **Rationale:** Amazon SageMaker becomes a comprehensive data platform aimed at simplifying data access for building AI applications. It unifies capabilities for fast SQL analytics, petabyte-scale big data processing, data exploration and integration, model development and training, and generative artificial intelligence (AI). This release confirms the market trend of unifying data repositories and AI model training environments for faster AI application development.

Amazon Bedrock

[Newly enhanced Amazon Connect adds generative AI, WhatsApp Business, and secure data collection](https://aws.amazon.com/blogs/aws/newly-enhanced-amazon-connect-adds-generative-ai-whatsapp-business-and-secure-data-collection/)

* **Rationale:** Amazon Connect's new features integrate generative AI capabilities, WhatsApp Business connectivity, and enhanced data collection tools, providing a coherent platform for customer engagement in enterprises. We believe that this platform will have a broad impact on call centers and enterprises globally, making it a ‘must know case’ in integrating AI for operational efficiency.

Microsoft Copilot (Autonomous AI Agents)

[Copilot Studio is enhancing its platform with knowledge improvements, Azure AI integration, and more](https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/copilot-studio-is-enhancing-its-platform-with-knowledge-improvements-azure-ai-integration-and-more/)

* **Rationale:** Microsoft's updates to Copilot Studio, including multimodal AI, agent previews, and deeper Azure AI integration, position the platform as a leading low-code solution. **“Autonomous agents** are now in preview. Makers can now build agents that work on their behalf, without having to prompt the agent, saving human hours and increasing efficiency. They can create these agents from scratch or configure agents that are prebuilt in Copilot Studio.” These enhancements focus on democratizing AI for non-technical users and emphasize enterprise-wide control and scalability.

[Microsoft’s agentic AI tool OmniParser rockets up the open source charts](https://venturebeat.com/ai/microsofts-agentic-ai-tool-omniparser-rockets-up-the-open-source-charts/)

* **Rationale:** With tools like Anthropic's Computer Use starting to enable LLMs to take control of your desktop, OmniParser could end up being a big piece of the agent system puzzle. The rate at which this new tool is being adopted is a big signal towards the importance of this development. It is worth being aware of it.

Company Use Cases

Conversational AI Assistants

[How InsuranceDekho transformed insurance agent interactions using Amazon Bedrock and generative AI](https://aws.amazon.com/blogs/machine-learning/how-insurancedekho-transformed-insurance-agent-interactions-using-amazon-bedrock-and-generative-ai/)

* **Rationale:** This case study highlights how InsuranceDekho used Amazon Bedrock to achieve an 80% reduction in response times and improve cross-selling opportunities. The detailed architecture and clear business outcomes make it a valuable model for organizations aiming to enhance customer interactions through generative AI conversational assistants.

[How Amazon Finance Automation built a generative AI Q&A chat assistant using Amazon Bedrock | Amazon Web Services](https://aws.amazon.com/blogs/machine-learning/how-amazon-finance-automation-built-a-generative-ai-qa-chat-assistant-using-amazon-bedrock/)

* **Rationale:** This use case from Amazon Finance Automation unit showcases a step-by-step process for developing a generative AI Q&A assistant using RAG (Retrieval-Augmented Generation) on Bedrock. By improving performance from 49% to 86% RAG accuracy, this article provides valuable insights on successful iterative performance optimization.

Mass Data Summarization

[FinRobot: AI Agent for Equity Research and Valuation with Large Language Models](https://arxiv.org/abs/2411.08804?s=09)

* Rationale: This research highlights a multi-agent AI system for equity research, emphasizing open-sourced chain of thought (CoT) reasoning and evaluation frameworks. The Data-CoT Agent aggregates diverse data sources for robust financial integration and the Concept-CoT Agent mimics an analysis reasoning to generate actionable insights. This application of LLMs to summarize and understand contexts has potential to revolutionize content-heavy tasks beyond finance, aligning with trends in AI-driven decision-making.

*Feedback requested – Are longer research articles like this useful to be aware of?*

Personalized Product Recommendation Engines

[Syngenta develops a generative AI assistant to support sales representatives using Amazon Bedrock Agents](https://aws.amazon.com/blogs/machine-learning/syngenta-develops-a-generative-ai-assistant-to-support-sales-representatives-using-amazon-bedrock-agents/)

* **Rationale:** Syngenta introduced Cropwise AI, an innovative platform leveraging Amazon Bedrock Agents to provide precision agriculture solutions to their agents in the field. Cropwise AI enables personalized recommendations at scale, tailoring seed choices to align with local conditions and specific farm needs, creating a more precise and accessible selection process. The use case offers insights into the implementation that can be useful for leaders developing similar personalized product recommendation applications in other industries.