



JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Lecture 4

BU.330.760 Generative AI for Business

Minghong Xu, PhD.
Associate Professor

Reflections



» Language models vs. reasoning models

- “Speak out” the thinking process
- Chain of thought: show the steps
- But more tokens generated, more expensive
- Techniques such as DeepSeek can partially mitigate the cost issue
- *Question: is showing the steps the only way human can develop reasoning capabilities?*

» Prompt Engineering

- You will need it in Agentic AI



Today's Agenda

- » Agentic AI
- » Physical AI
- » Business case kickoff

- » Homework 1 grading is released
- » Don't forget to submit Homework 2

Automation



- » Ultimate goal of AI: achieve automation
- » From assist you to work as your agent



AI Agents



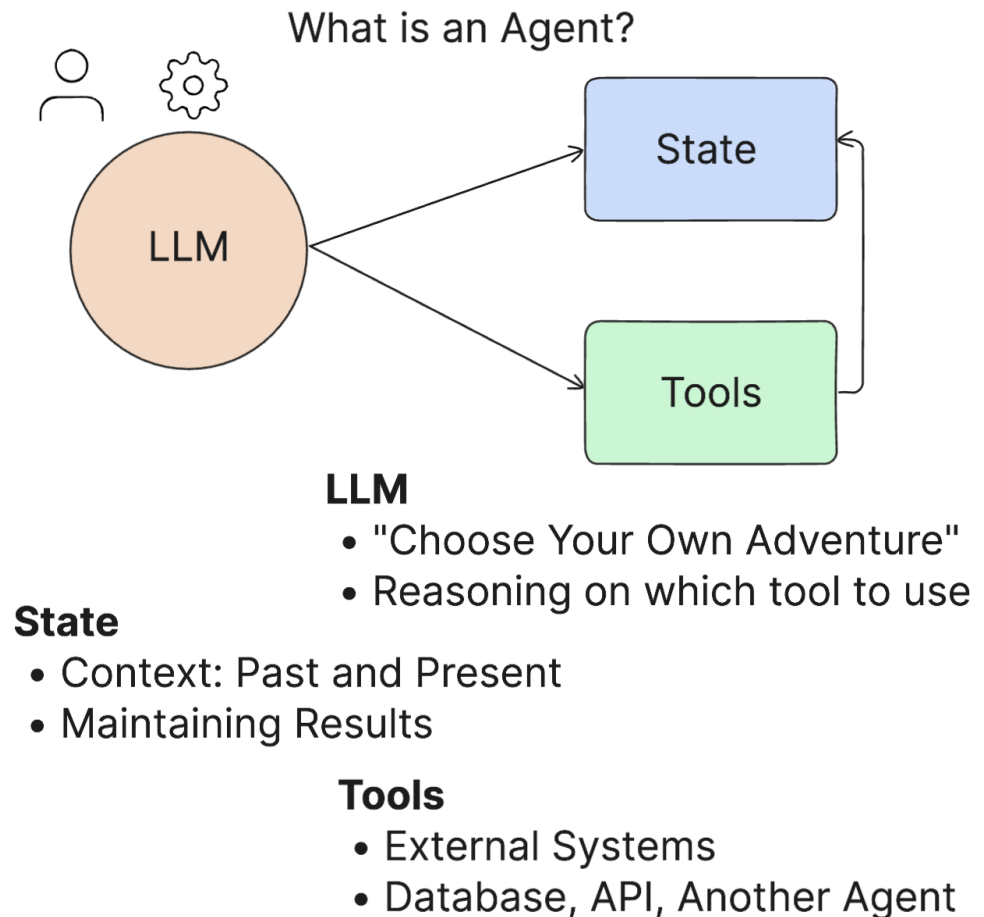
- » A person legally empowered to act on behalf of another person or entity
- » Designate AI for this role

More Mathematical Definition



» AI Agents allow Large Language Models (LLMs) to perform tasks by giving them access to a state and tools

- State: context of work
- Tools: database, API, application or another LLM



Why AI Agents



What can they do?



The Batch > Letters > Article

The Dawning Age of Agents

LLM-based agents that act autonomously are making rapid progress. Here's what we have to look forward to.

Letters Technical Insights

Published Mar 06, 2024
Reading time 3 min read

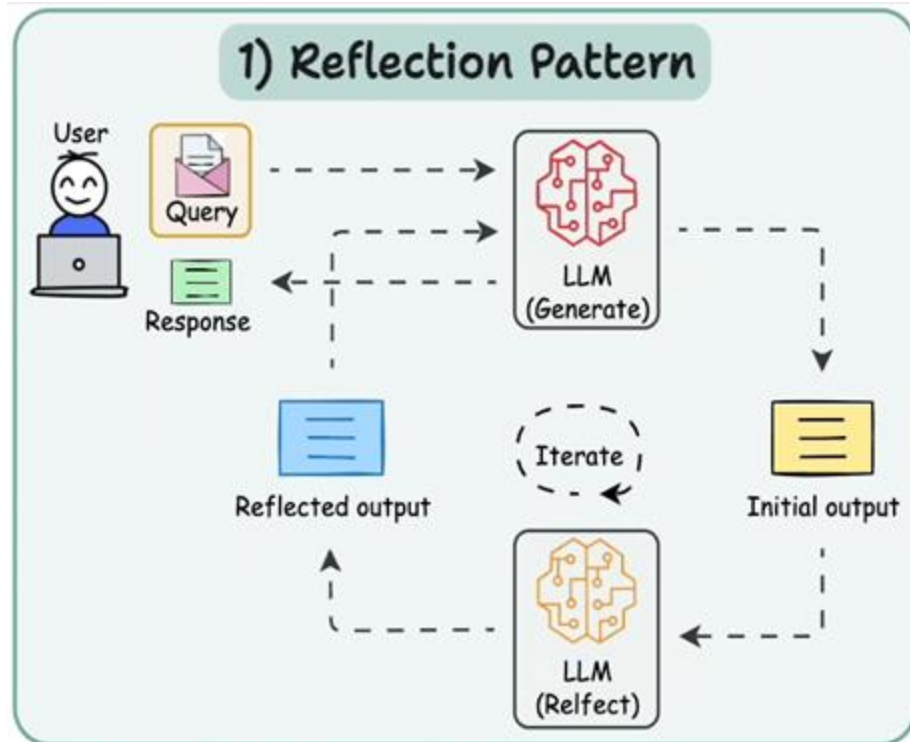


<https://www.deeplearning.ai/the-batch/the-dawning-age-of-agents/>

- » Plan out autonomously, execute sequences of actions
- » Search and fetch webpages, dispatch a product to user, control a robot

Reflection Pattern

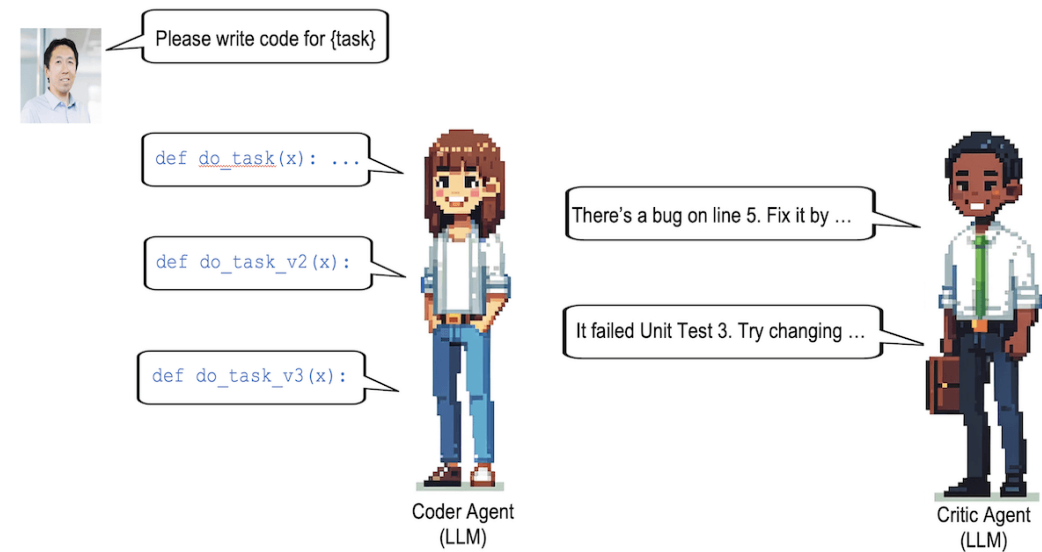
- » AI reviews its work to spot mistakes or areas for improvement
- » Iterate until it produces final response



<https://blog.dailydoseofds.com/p/5-agentic-ai-design-patterns>

Minghong Xu, PhD.

Agentic Design Patterns: Reflection

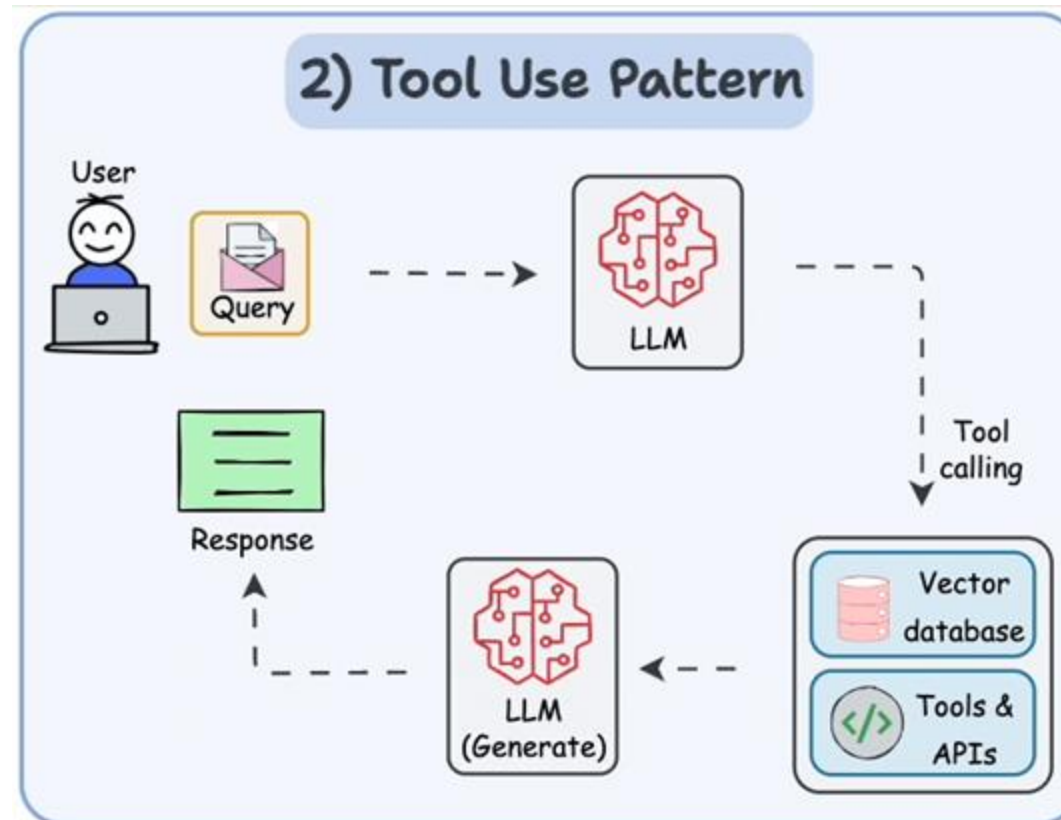


<https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-2-reflection/>

Tool Use Pattern

- » Tools allow LLMs to gather more information by:
 - web search
 - querying a vector database
 - executing Python scripts
 - invoking APIs, or any other function
- » Rely only on a pre-trained transformer to generate output tokens is limiting

<https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-3-tool-use/>



<https://blog.dailydoseofds.com/p/5-agentic-ai-design-patterns>

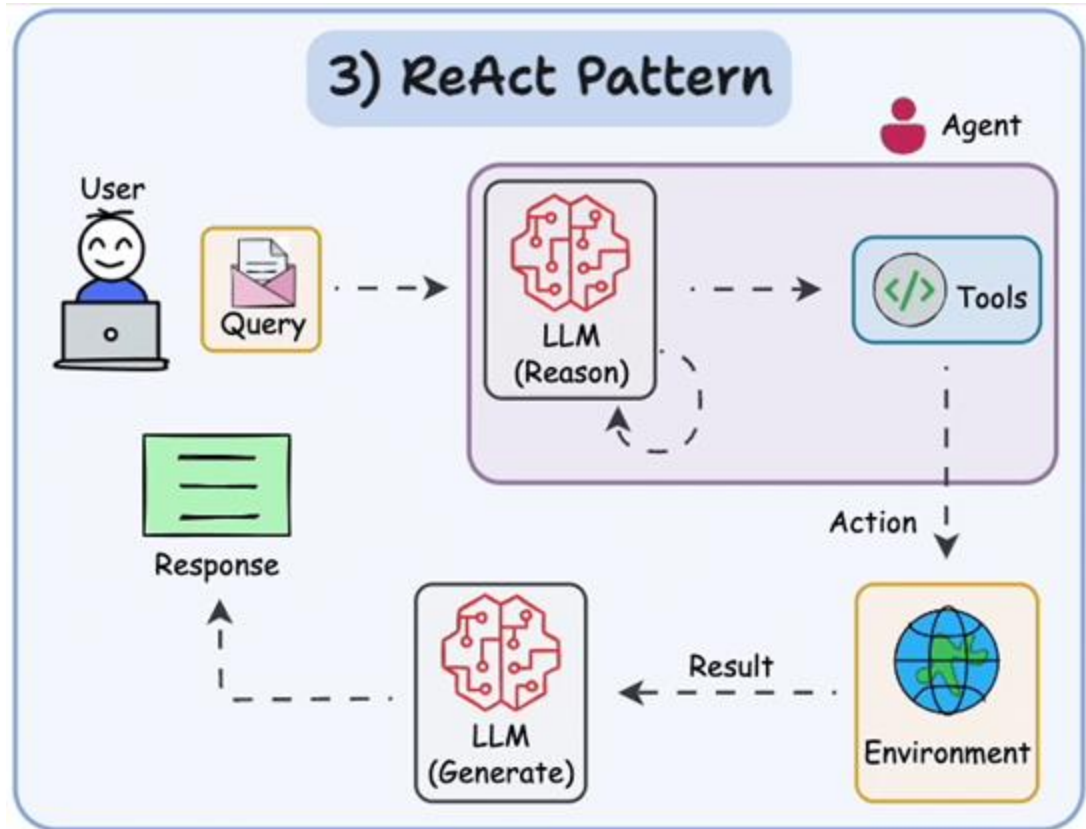


Tools can be used

- » Search different sources: Wiki, arXiv, etc.
- » Productivity tools: send email, read/write calendar entries, etc.
- » Image tools
 - Before the availability of multimodal models
- » Ask LLM to automatically choose the right tool
- » Or a subset of tools if there are hundreds of tools available

Reason and Act Pattern

- » ReAct combines the previous two patterns:
- The Agent can reflect on the generated outputs
 - It can interact with the world using tools

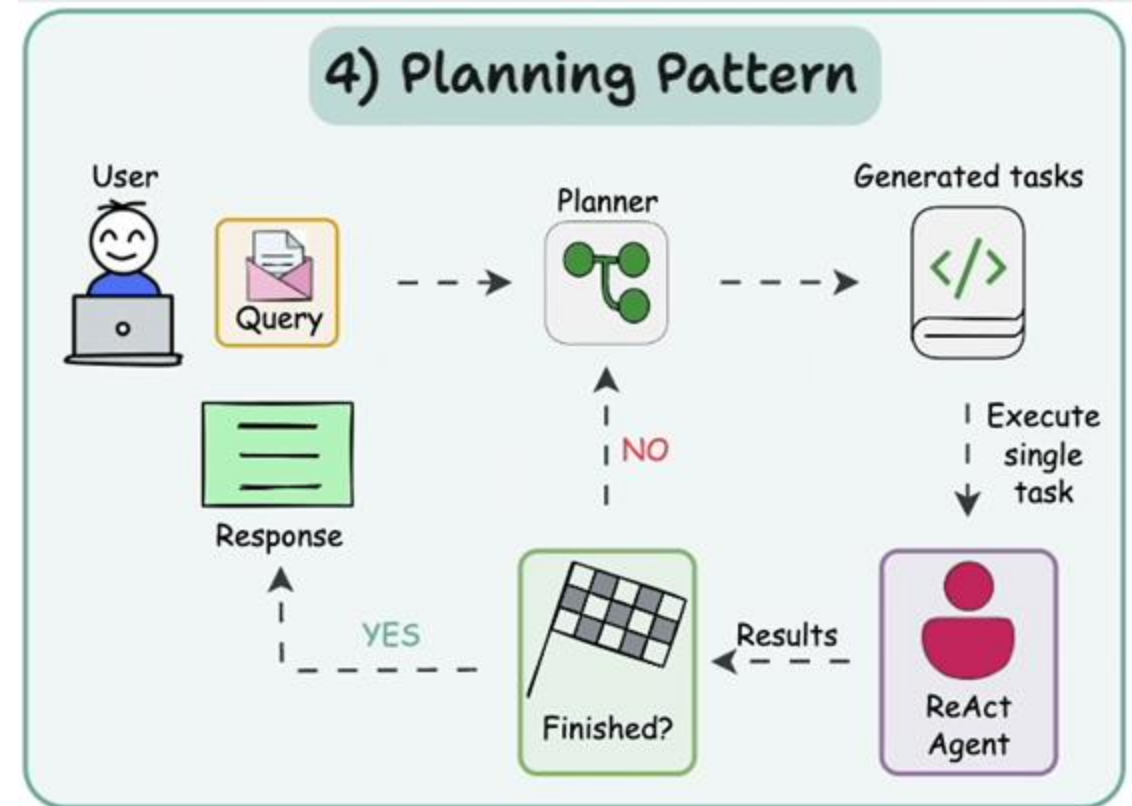


<https://blog.dailydoseofds.com/p/5-agentic-ai-design-patterns>

Planning Pattern

- » Planning: design and execute a multistep plan to achieve a goal
 - write an outline for an essay->do online research->write a draft, and so on
- » Allow agent to decide dynamically what steps to take
- » May lead to less predictable results

<https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-4-planning/>

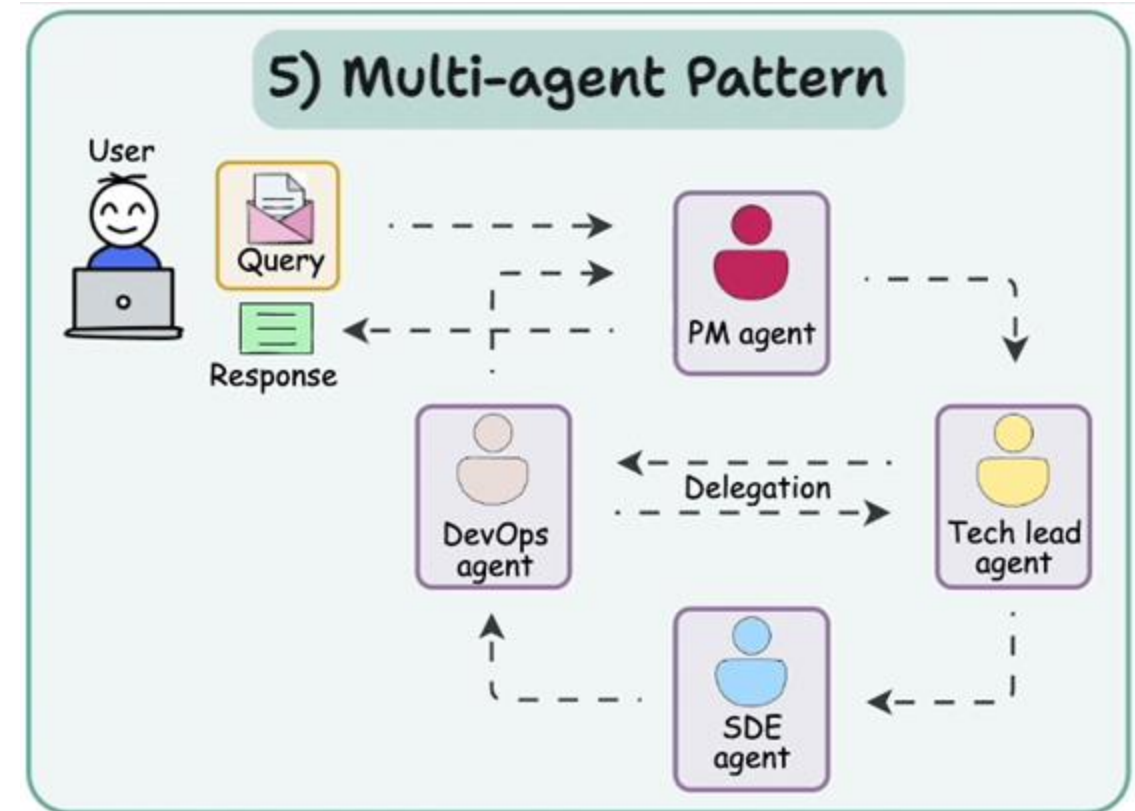


<https://blog.dailydoseofds.com/p/5-agentic-ai-design-patterns>

Multi-Agent Pattern

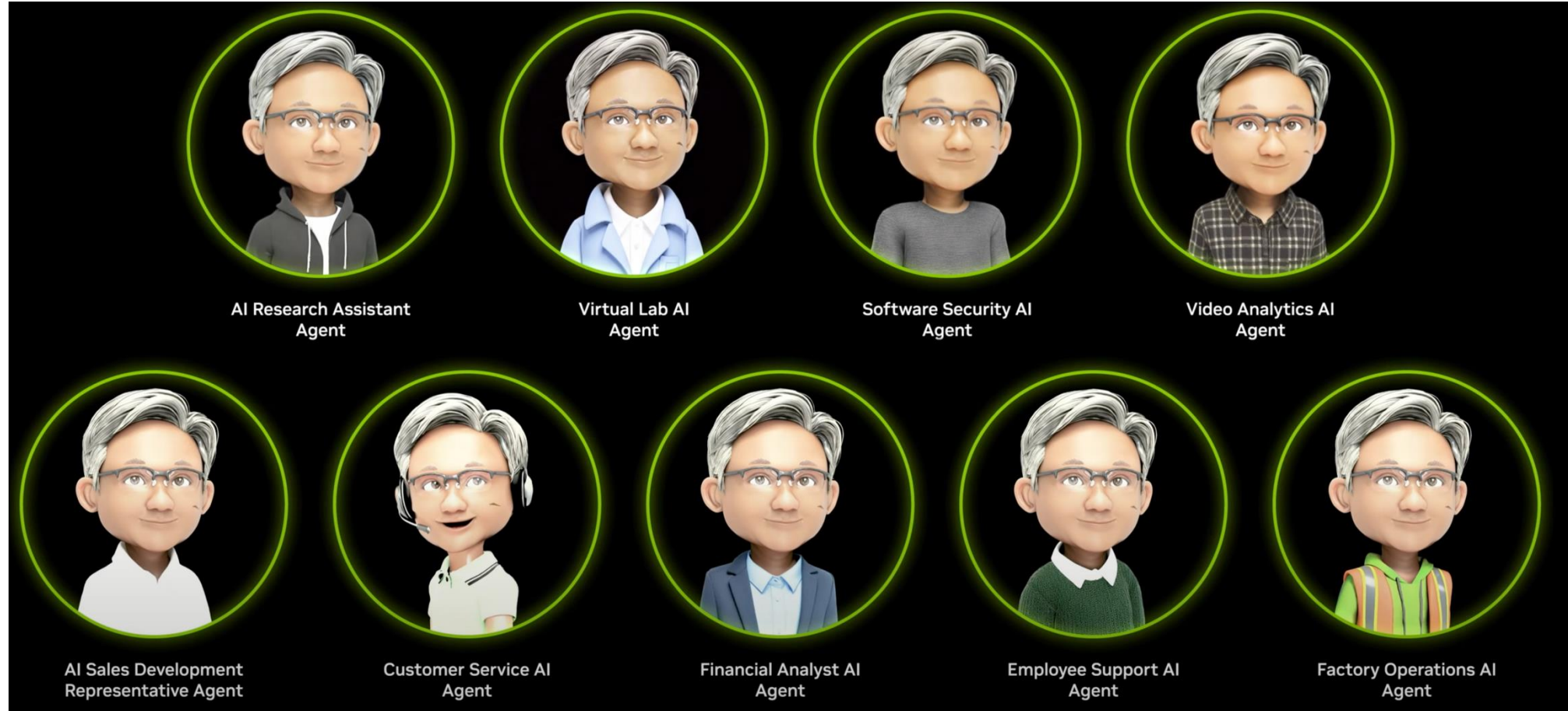
- » Multi-agent collaboration: more than one AI agent work together
 - split up tasks, each agent assigned a dedicated role and task
 - discuss and debate ideas, access tools
- » Result: higher-quality output, more intelligent system
- » Similar to human teaming, but lower damage from mismanaging

<https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-5-multi-agent-collaboration/>



<https://blog.dailydoseofds.com/p/5-agentic-ai-design-patterns>

IT: The NEW HR



Jensen Huang's Keynote at CES 2025

OpenAI Deep Research



» *Demo Time*

Why Agentic



The Batch > Letters > Article

Building Models That Learn From Themselves

AI developers are hungry for more high-quality training data. The combination of agentic workflows and inexpensive token generation could supply it.

Letters Technical Insights

Published May 1, 2024
Reading time 2 min read



<https://www.deeplearning.ai/the-batch/building-models-that-learn-from-themselves/>

- » Humans can learn from each other
- » Perhaps LLMs can too
- » Train smaller models directly on the output of large models



Andrew Warns of Model Collapse

- » <https://www.deeplearning.ai/the-batch/study-reveals-serious-defects-in-models-trained-on-their-own-content/>
- » An LLM can't learn much by training on data it generated directly
- » Like a supervised learning algorithm can't learn from trying to predict labels it generated by itself
- » Training a model repeatedly on the output of an earlier version of itself can result in model collapse
- » Like human not repeat ourselves, gain will be minimal

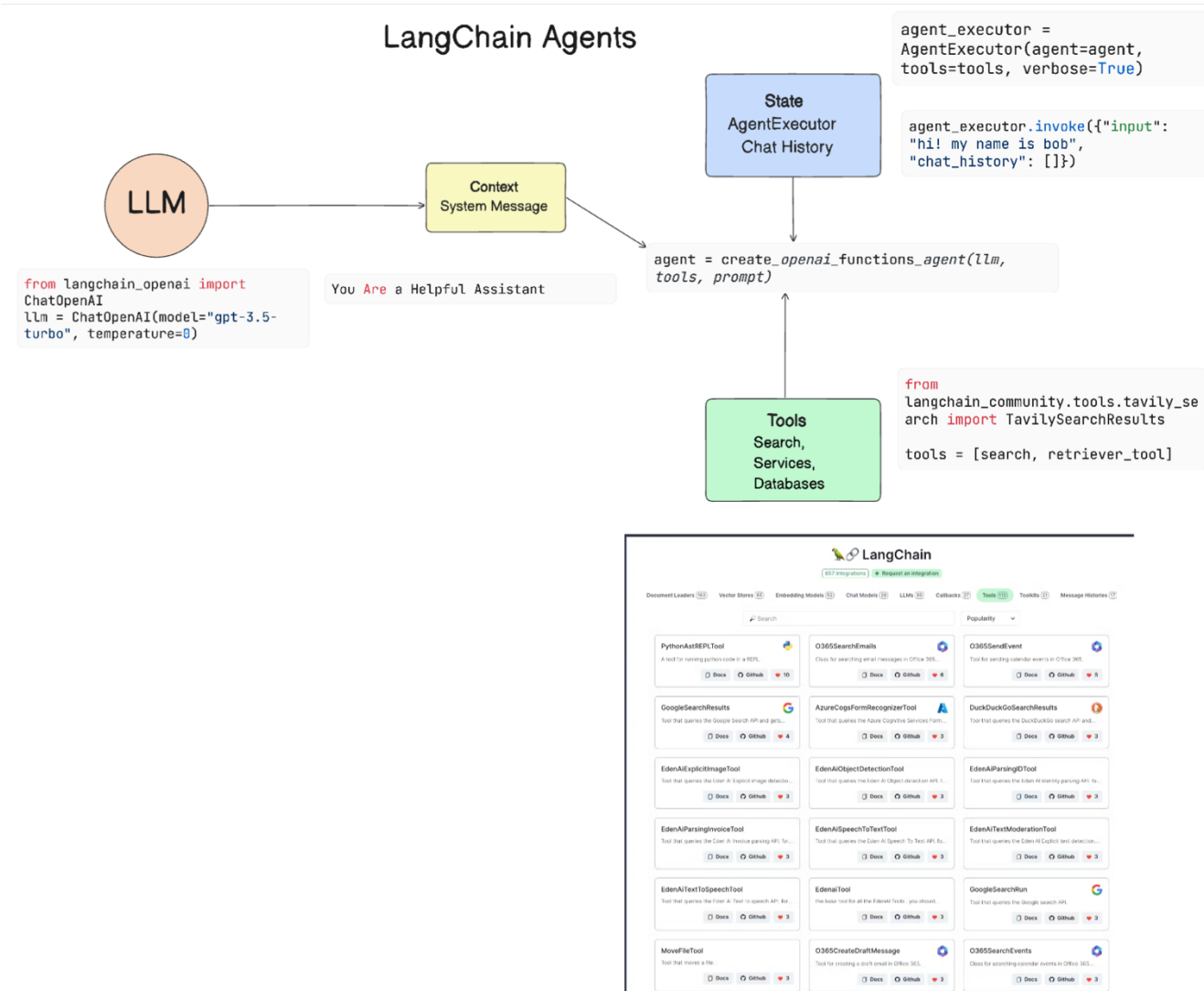


AI Agent Tools

LangChain Agents (Optional)



» https://python.langchain.com/docs/how_to/#agents



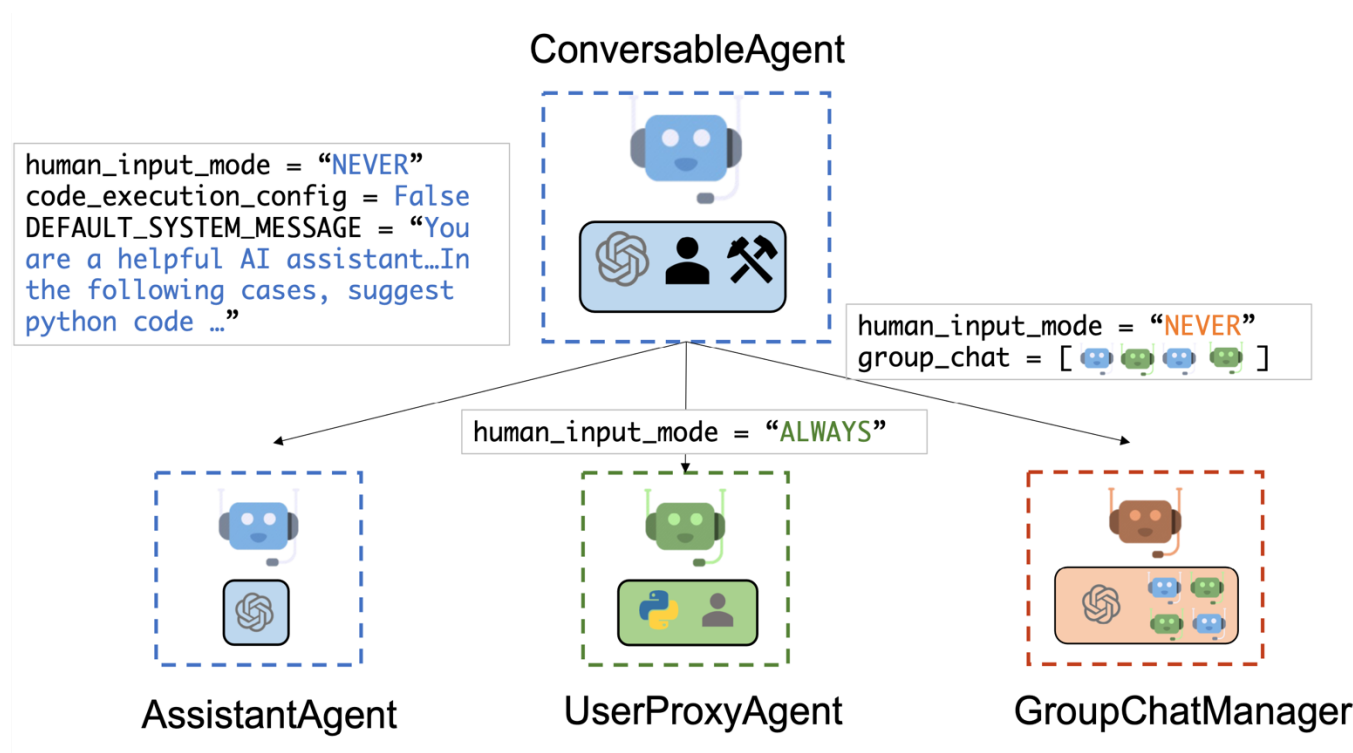
AutoGen



» <https://microsoft.github.io/autogen/docs/Examples>

» Main focus: conversations

- Agents are both **conversable** and **customizable**



ConversableAgent



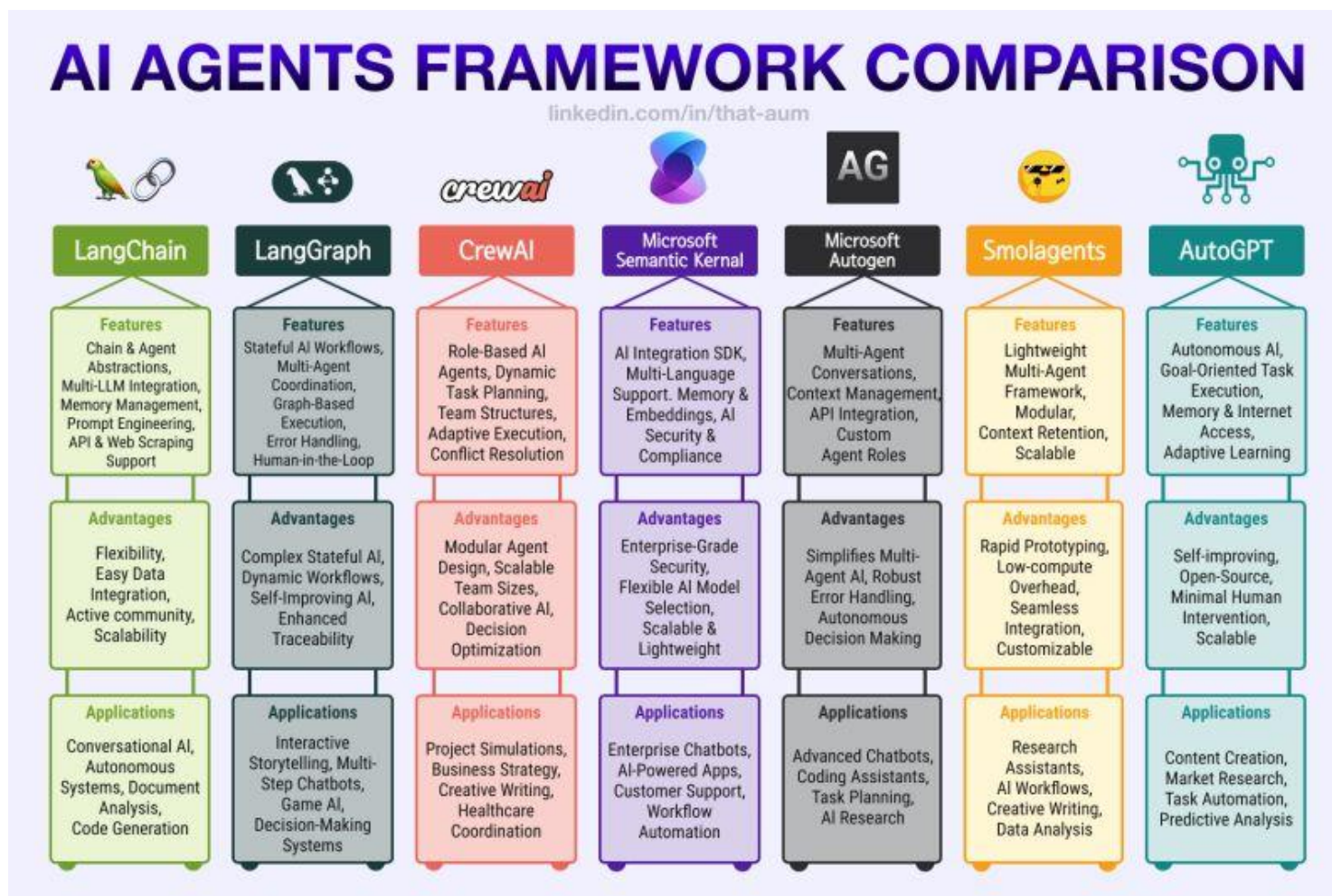
- » Generic class for Agents, capable of conversing with each other to jointly finish a task
 - AssistantAgent: work as your AI assistant, can write code, suggest corrections, fix bugs, etc.
 - UserProxyAgent: proxy agent for humans, soliciting human input as the agent's reply
- » llm_config: LLM configuration, set as a dictionary
- » register_reply(): register your own reply functions

Lab 3 AutoGen



- » Mock interview
- » Stock comparison

Other Tools





Low/No Code Tools

- » Amazon PartyRock: <https://partyrock.aws/home>
 - Free
 - Good for you to explore ideas before coding your own Python solution
- » Google Vertex AI Agent Builder:
<https://cloud.google.com/products/agent-builder>
- » Microsoft AutoGen Studio:
<https://microsoft.github.io/autogen/0.2/docs/autogen-studio/getting-started>

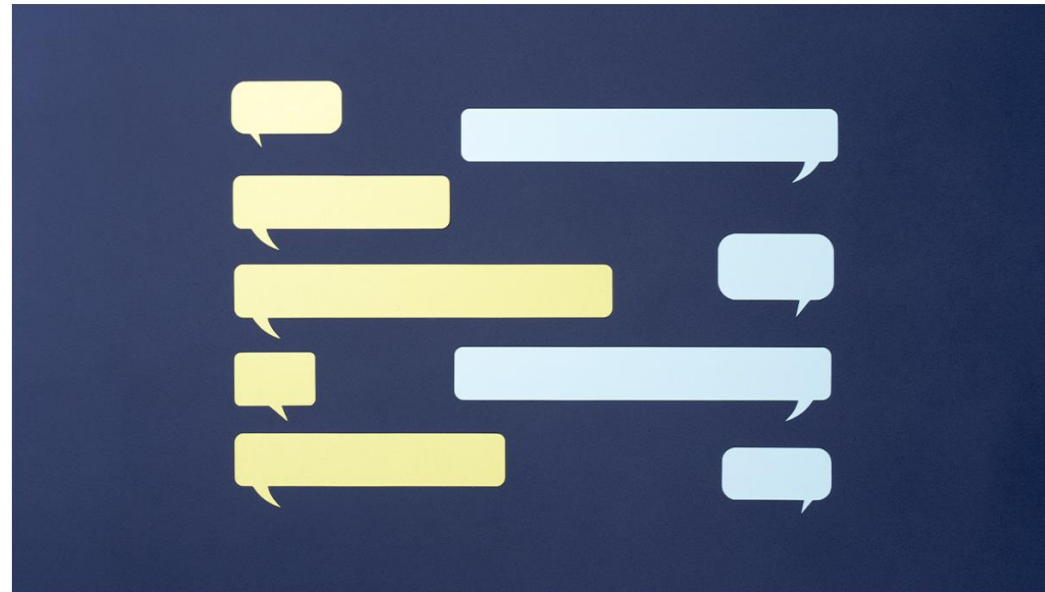
Use Case: Negotiation Bot



How Walmart Automated Supplier Negotiations

by Remko Van Hoek, Michael DeWitt, Mary Lacity, and Travis Johnson

November 08, 2022



MirageC/Getty Images

<https://hbr.org/2022/11/how-walmart-automated-supplier-negotiations>



Supply Chain Negotiations of Walmart

- » Large procurement operations, 100,000+ suppliers
- » Impossible to conduct focused negotiations with all of them
- » Have to use cookie-cutter terms

- » Walmart solution with AI: chatbot negotiator, Pactum AI
 - On behalf of Walmart
 - Negotiate with human suppliers
 - Pilot in Jan 2021 by Walmart Canada
 - Improve payment terms, secure additional discounts



Automated Procurement Negotiation

» Walmart's target:

- Negotiate early payment discounts
- Extended payment terms without discounts

» Walmart's offer:

- Option for suppliers to change termination terms
- Opportunities for growth in assortment and sales volume

» Walmart's gain from pilot:

- Close 64% deals, well above 20% target
- 1.5% in savings on the spend
- Extension of payment terms to an average of 35 days



Suppliers' Feedback

» Pros:

- Easy to use
- Ability to make a counteroffer
- Respond at their own pace

» Cons:

- Prefer face to face
- Too verbose



Lessons to Learn

- » Move to a production pilot quickly
 - Gartner: AI journey for many companies languishes in the proof-of-concept phase, focus on technical capabilities instead of business goals
- » Start with indirect-spend categories and pre-approved suppliers
 - Minimize the risks of disruption
- » Decide on acceptable trade-offs
 - Define the boundaries of what the buyer is willing to concede in exchange for what it wants
- » Scale by extending geographies, categories, and use cases
 - Different regions, products, suppliers, languages...
 - Scaling provide additional cases for improvement

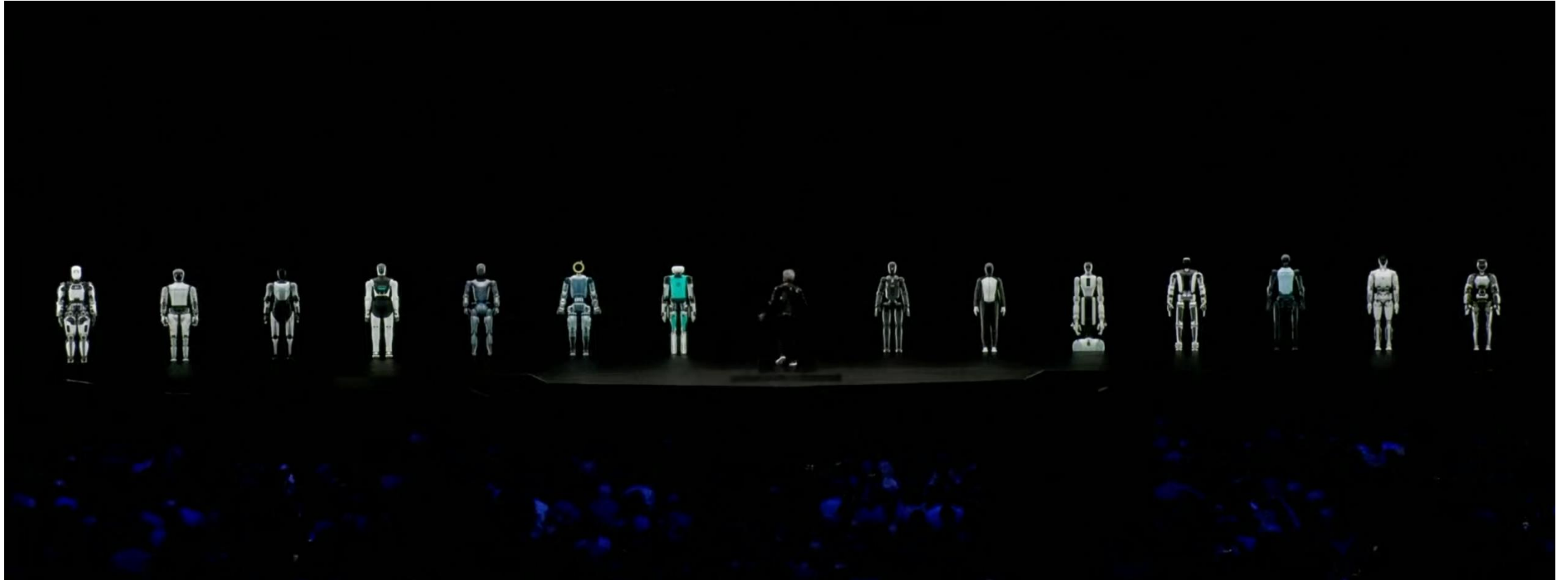
Use Case: Insurance Underwriting





Physical AI

Robots' “ChatGPT” Moment



https://www.youtube.com/live/XASnBeNKg6A?si=JiUPBY6WOB-j_-AK&t=6770

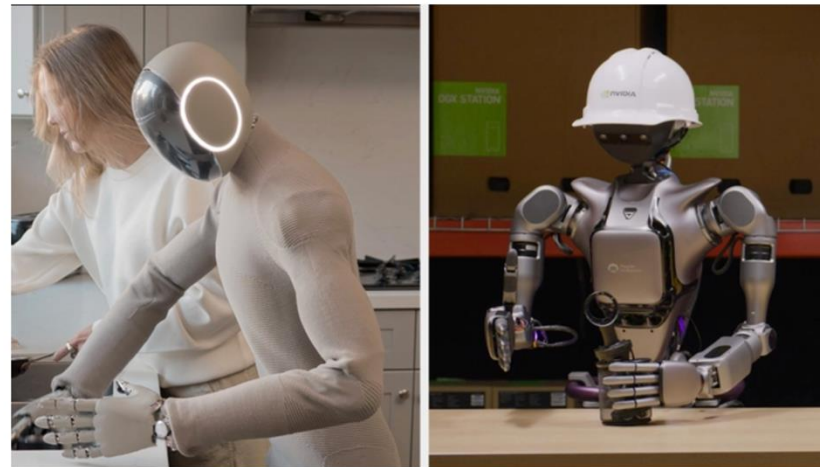
Robot Foundation Model



NVIDIA Announces Isaac GR00T N1 — the World's First Open Humanoid Robot Foundation Model — and Simulation Frameworks to Speed Robot Development

- › Now Available, Fully Customizable Foundation Model Brings Generalized Skills and Reasoning to Humanoid Robots
- › NVIDIA, Google DeepMind and Disney Research Collaborate to Develop Next-Generation Open-Source Newton Physics Engine
- › New Omniverse Blueprint for Synthetic Data Generation and Open-Source Dataset Jumpstart Physical AI Data Flywheel

March 18, 2025



<https://nvidianews.nvidia.com/news/nvidia-isaac-gr00t-n1-open-humanoid-robot-foundation-model-simulation-frameworks>

How This Time is Different

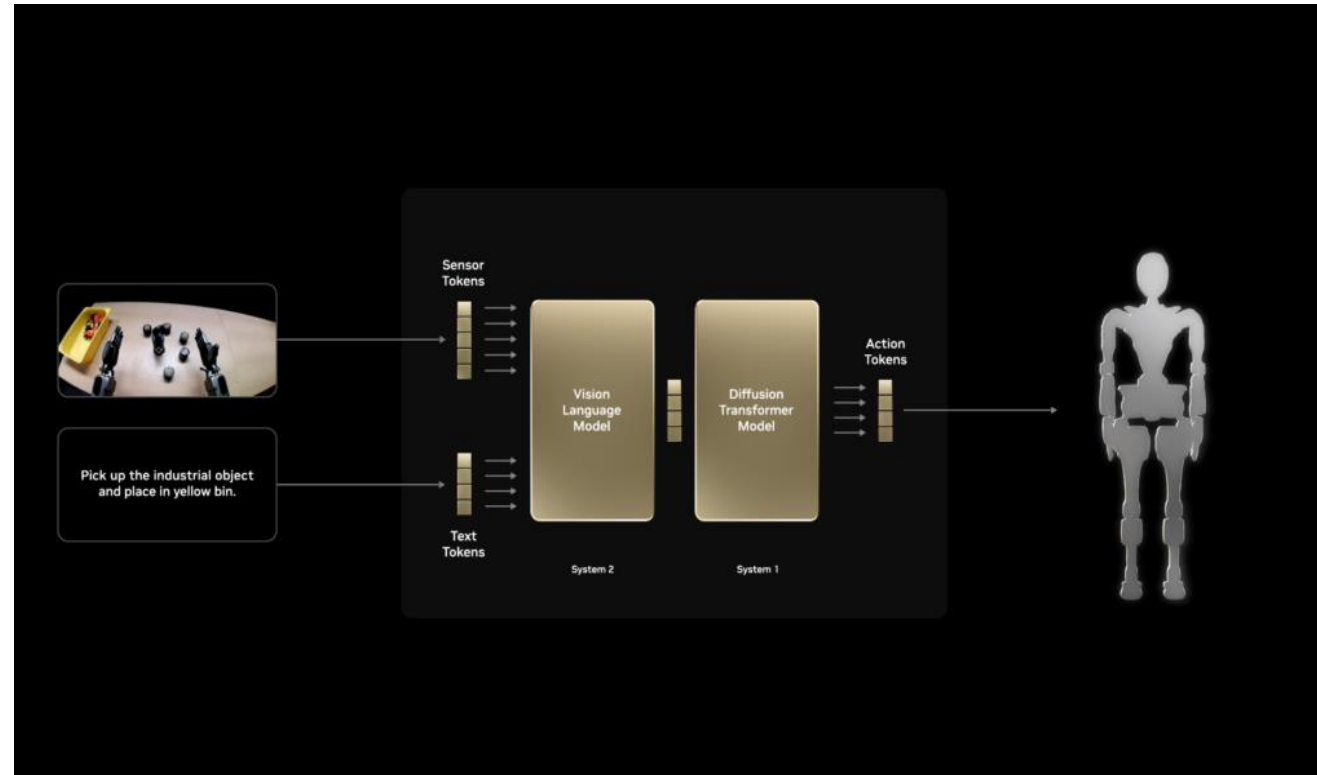


Aspect	Nvidia Issac GR00T N1 (2025)	Previous Robotics AI Approaches (Pre-2025)
Architecture	Dual-system model: cognitive planner + reactive controller (next page)	Typically single-policy or modular controllers per task; no “thinking vs. acting”
Learning Paradigm	Foundation model trained via imitation learning from human demos, scaled with enormous synthetic data. Robots learns from watching humans, then fine-tunes to new tasks with extra data	Learned task-by-task, requiring large task-specific datasets or manual programming. Imitation learning on limited data. No large pre-trained “robot brain”
Skill Generalization	Pretrained with broad manipulation skills applicable to many scenarios. Adapt to novel situations with some additional training	Narrow scope per model. Skills weren’t easily transferable
Data Scale & Sources	Internet videos of human behavior, synthetic trajectories from simulation, plus real teleoperation data. Open-sourced datasets	Data was a bottleneck. Simulation use was smaller scale. Typically proprietary
Performance	Smoother and more fluid motions, higher success rates	Jerkier movements or brittle skills. Noticeable gaps in fluidity and reliability
Accessibility	Open-source	Internal to companies or labs

Dual System



- » **Vision-Language Model (System 2):** interpret the environment through vision and language instructions, enabling robots to reason about their environment and instructions, and plan the right actions



- » **Diffusion Transformer (System 1):** generates continuous actions to control the robot's movements, translating the action plan made by System 2 into precise, continuous robot movements

<https://developer.nvidia.com/blog/accelerate-generalist-humanoid-robot-development-with-nvidia-isaac-gr00t-n1/>

Artificial General Intelligence (AGI)



EDITORS' PICK | INNOVATION > CONSUMER TECH

OpenAI CEO Sam Altman: 'We Know How To Build AGI'

By [John Koetsier](#), Senior Contributor. ⓘ Journalist, analyst, author, podcaster.

[Follow Author](#)

Jan 06, 2025 at 01:15pm EST

Tesla's Musk predicts AI will be smarter than the smartest human next year

By Reuters

April 8, 2024 4:09 PM EDT · Updated a year ago



'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years

Geoffrey Hinton says there is 10% to 20% chance AI will lead to human extinction in three decades, as change moves fast

● **'We need dramatic changes': is societal collapse inevitable?**






📷 Geoffrey Hinton said humans will be like toddlers compared with the intelligence of highly power AI systems. Photograph: Pontus Lundahl/TT News Agency/AFP/Getty Images

AGI Hype

» Yann LeCun at HBC

» <https://youtu.be/UmxlgLEscBs?si=FxnvOPdjInlg49g&t=1653>

**Andrew Ng**  • Following
Founder of DeepLearning.AI; Managing General Partner of AI Fund; Ex...
2w • Edited • 

Last Friday on Pi Day, we held AI Dev 25, a new conference for AI Developers. Tickets had (unfortunately) sold out shortly after we announced their availability, but I came away energized by the day of coding and technical discussions with fellow AI Builders! Let me share here my observations from the event.

I'd decided to start AI Dev because while there're great academic AI conferences that disseminate research work (such as NeurIPS, ICML and ICLR) and also great meetings held by individual companies, often focused on each company's product offerings, there were few vendor-neutral conferences for AI developers. With the wide range of AI tools now available, there is a rich set of opportunities for developers to build new things (and to share ideas on how to build things!), but also a need for a neutral forum that helps developers do so.

Based on an informal poll, about half the attendees had traveled to San Francisco from outside the Bay Area for this meeting, including many who had come from overseas. I was thrilled by the enthusiasm to be part of this AI Builder community. To everyone who came, thank you!

Other aspects of the event that struck me:

- First, agentic AI continues to be a strong theme. The topic attendees most wanted to hear about (based on free text responses to our in-person survey at the start of the event) was agents!
- Google's Paige Bailey talked about embedding AI in everything and using a wide range of models to do so. I also particularly enjoyed her demos of Astra and Deep Research agents.
- Meta's Amit Sangani talked compellingly as usual about open models. Specifically, he described developers fine-tuning smaller models on specific data, resulting in superior performance than with large general purpose models. While there're still many companies using fine-tuning that should really just be prompting, I'm also seeing continued growth of fine-tuning in applications that are reaching scale and that are becoming valuable.
- Many speakers also spoke about the importance of being pragmatic about what problems we are solving, as opposed to buying into the AGI hype. For example, Nebius' Roman Chernin put it simply: Focusing on solving real problems is important!



Next Week



- » BloombergGPT
- » RAG
- » Adversarial Attacks, Risk and Governance



References

» Andrew Ng's DeepLearning AI

<https://www.deeplearning.ai/the-batch/tag/letters/>

» Microsoft's Generative AI for Beginners

<https://github.com/microsoft/generative-ai-for-beginners/tree/main>