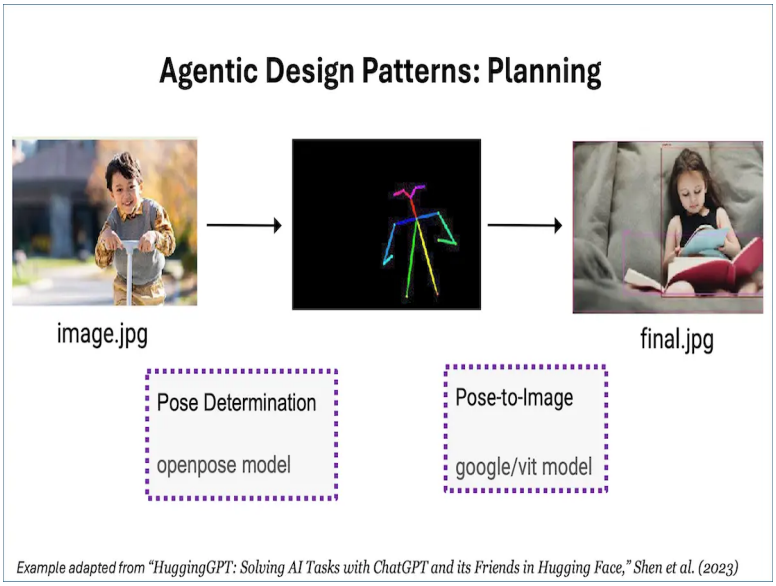
While technological advancements in the lower layers of the AI stack are crucial, the true potential of AI is unlocked through the development and deployment of innovative applications. This layered perspective highlights the interdependence of various components within the AI ecosystem. Generative

components within the AI ecosystem. Generative AI significantly accelerates the development and iteration of machine learning models, allowing for rapid prototyping and experimentation. This speed allows researchers and developers to test multiple hypotheses efficiently, leading to faster

hypotheses efficiently, leading to faster innovation.

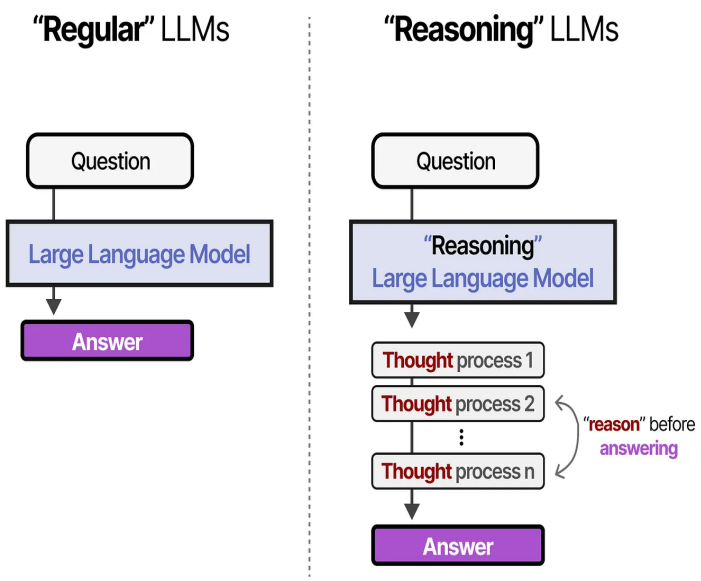
Agentic workflows represent a shift in AI interaction, moving beyond single-prompt responses to iterative and deliberative processes that mirror human problem-solving. These workflows involve multiple steps like research, planning, execution, and refinement, often leading to significantly improved



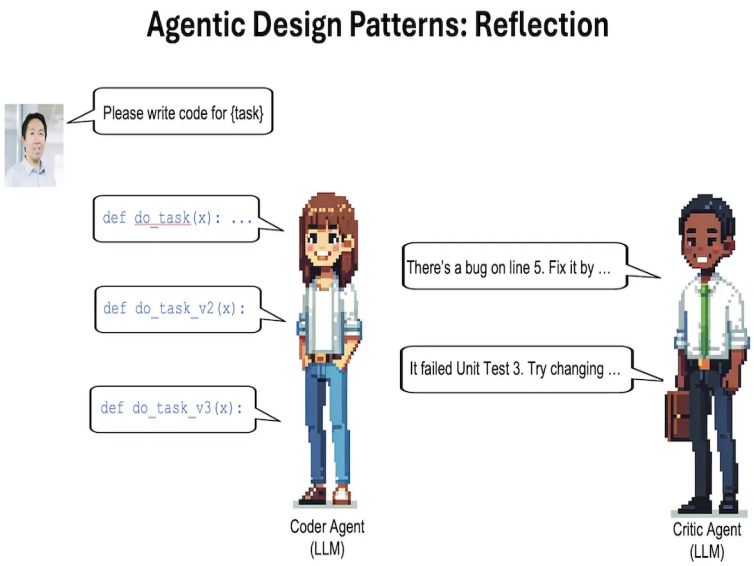
often leading to significantly improved outcomes, especially for complex tasks requiring nuanced reasoning.

The "agentic orchestration layer" is emerging to facilitate the development and management of these complex agent-based applications. Agentic workflows don't necessarily rely solely on LLMs; they can incorporate diverse modules, including those outside the realm of LLMs, to achieve a common goal:

the realm of LLMs, to achieve a common goal: generating a more comprehensive and refined response to a given query.

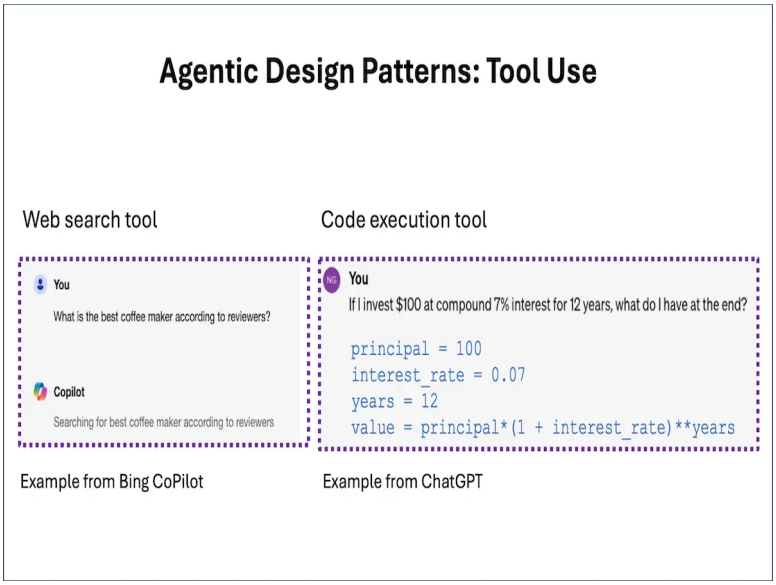


The Reflection design pattern empowers AI agents to enhance their output quality through self-evaluation and improvement. Similar to peer review or debugging, this involves an agent analyzing its own work, identifying errors or areas for refinement, and iteratively revising its output. This



and iteratively revising its output. This process can occur within a single agent or involve multiple agents, with one generating content and another providing constructive criticism.

The Tool Use design pattern enables AI agents to expand their capabilities by leveraging external tools and APIs. These tools can encompass a wide range of functionalities, from web search engines for information retrieval to code interpreters for computational tasks. By integrating with these



computational tasks. By integrating with these external resources, AI agents become more versatile and capable of tackling complex tasks that require specialized knowledge or interaction with the external world.

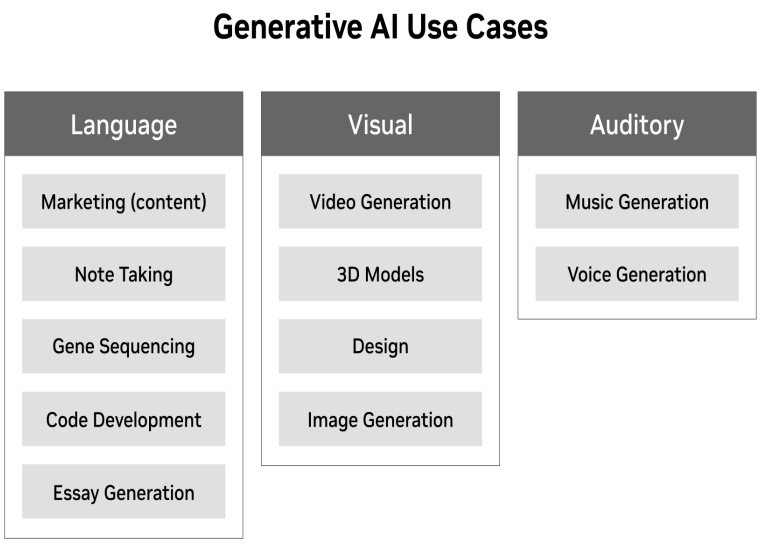
This decomposition allows AI agents to approach intricate problems in a structured and organized manner, similar to project management methodologies. By planning the necessary steps to achieve a complex objective, agents can better manage complexity and increase the likelihood of successful task

and increase the likelihood of successful task completion. This planning process can be static, with the entire plan defined upfront, or dynamic, where the agent adjusts its plan based on new information or unexpected events. This approach is similar to how teams of researchers or engineers might

to how teams of researchers or engineers might collaborate on a project, each with specialized roles and expertise.

Visual AI technologies enable computers to interpret and understand visual information from images and videos, leading to diverse applications across autonomous vehicles, medical imaging analysis, quality control, and retail analytics. Effectively processing visual data alongside other

processing visual data alongside other unstructured data types like audio and text is crucial for building more powerful and versatile AI systems. Agentic AI, built upon compound LLMs, takes this a step further. This approach involves strategically linking multiple LLMs in sequence to enhance task



linking multiple LLMs in sequence to enhance task outcomes. For example, one LLM might draft initial content, another critically review it, and a third revise based on the critique, creating an iterative refinement process. This demonstrates the value of combining specialized AI functions,

the value of combining specialized AI functions, achieving higher quality and sophistication in output compared to single LLMs, ultimately moving towards AI systems capable of reflection and improvement during task execution.

Agentic AI takes compound LLMs a step further by incorporating diverse "agents" beyond just language models. These agents can encompass various computational tools and resources, like data retrieval systems or APIs, each selected for its specific capabilities. A crucial aspect is the orchestrator

A crucial aspect is the orchestrator agent, which dynamically manages the workflow, directing the appropriate agents based on the task's evolving needs. This dynamic approach allows for adaptation to changing conditions, leading to more flexible and robust problem-solving compared to rigid,

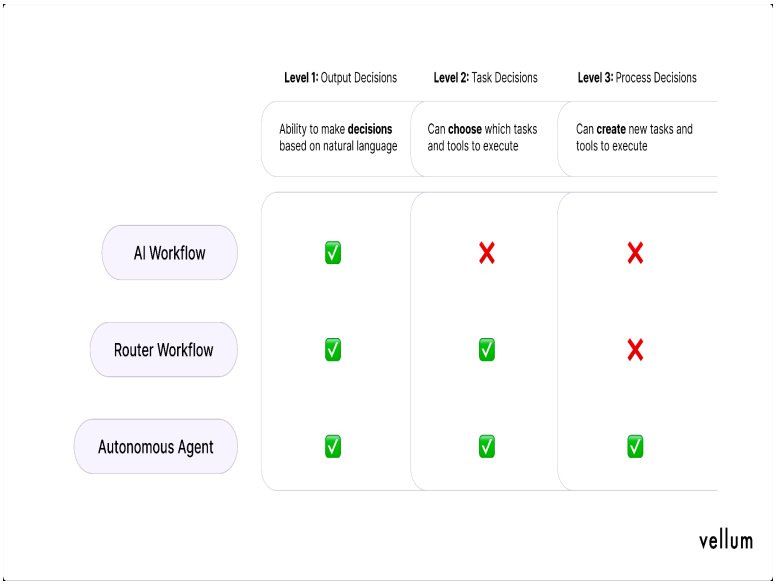
and robust problem-solving compared to rigid, pre-defined workflows.

Agentic AI systems demonstrate enhanced adaptability by modifying their strategies based on real-time feedback or changing environmental factors. This flexibility makes them more resilient and effective in dynamic, real-world situations compared to less adaptable AI architectures.

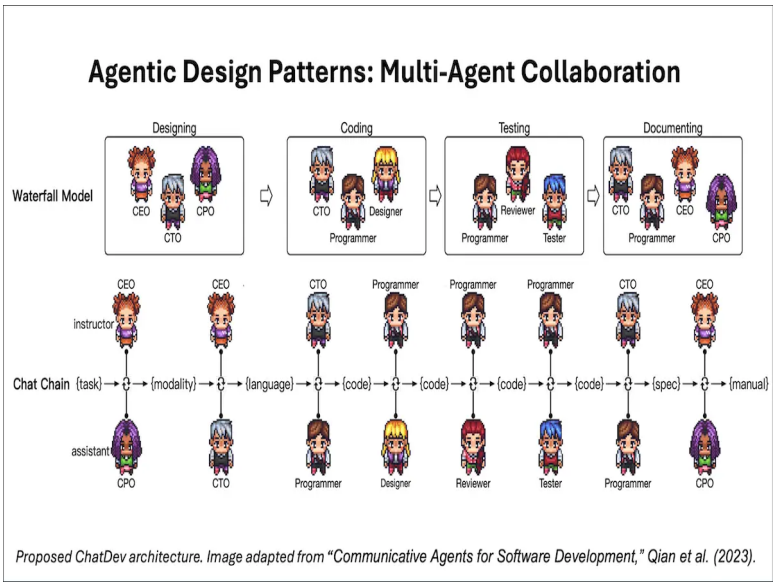
Conversational Workflow Orchestration in AI systems, as seen in platforms like AutoGen, involves the seamless interaction and communication between multiple AI agents. These agents engage in looping dialogues, collaborating and exchanging information until a specific task is successfully

information until a specific task is successfully completed. The ability to orchestrate these conversational workflows allows for more complex and sophisticated AI applications.

Conversational Workflow Orchestration allows for looping dialogues between agents to complete tasks, making it suitable for complex processes like code generation or debugging.



Agents can be extended to connect with external tools, APIs, or custom code, enabling them to perform various actions. For example, an agent could call Python code, execute shell commands, or interact with web APIs. This allows for the creation of sophisticated workflows involving multiple agents



Proposed ChatDev architecture. Image adapted from "Communicative Agents for Software Development," Qian et al. (2023).

sophisticated workflows involving multiple agents and external resources.

CrewAI allows you to define a sequence of agents to work together, like a production line. This is ideal for tasks like content creation, document processing, or automated research workflows. For example, you could have a "researcher" agent gather information, a "summarizer" agent condense it, and

a "summarizer" agent condense it, and a final agent present the findings. This agentic approach breaks down complex tasks into manageable steps, leveraging the strengths of different agents.

Agentic workflows leverage multiple components, often including Large Language Models (LLMs), to generate improved responses to queries. Compounded LLMs, a specific type of agentic workflow, involve iteratively refining a draft output through the feedback of one or more critic LLMs. The process

feedback of one or more critic LLMs. The process begins with an initial LLM generating a draft response, which is then evaluated by a critic LLM. The critic's feedback, whether suggesting improvements or highlighting areas needing attention, is used by a third LLM to update and refine the draft.

by a third LLM to update and refine the draft. This iterative process can continue for multiple rounds, with human feedback potentially integrated as a critic at any stage. The ultimate goal of agentic workflows is to produce a higher-quality, more accurate, and more insightful response by

more accurate, and more insightful response by leveraging the strengths of different LLM modules and incorporating feedback loops.