Human agents are a very handy thing. Travel agents can book you a nice cruise and take care of all the logistics. Insurance agents can protect your car and home and occasionally make you smile at their commercials. Secret agents can discover your enemy's secrets and imbibe their martinis, shaken not stirred.

So it should be no surprise that employing agents is the next big thing in AI solutions development. Agentic workflows in AI take multiple forms, but the gist of the approach is to use specialized AI agents to take on different tasks and have them collaborate to generate the best possible outcome, all autonomously. Current research shows that this consistently produces better results than the traditional request and response model, and most cutting-edge development techniques in AI now employ agentic workflows.

There's a catch, of course. Agentic workflows by nature require many more Large Language Model (LLM) interactions and many more tokens to be generated, in some cases orders of magnitude more. This exposes a potentially fatal flaw in scaling agentic AI applications: speed. If token generation is too slow it gums up the entire process.

For example, enter a complex request into your favorite consumer LLM service (ChatGPT, Gemini, etc.). It will probably take a few seconds before the response is complete. Now imagine it's not you, the end user, waiting for that response, but rather one of the string of agents at work in a multi-agent AI solution. Multiply that few seconds by 10 or even 100, and you'll see how the carefully crafted agentic workflow AI solution falls apart. As AI business technologists consider using agentic workflows to tackle their biggest challenges, they also need to ensure their inference strategy is up to the task, speedwise. Otherwise, they may build something cool and powerful only to find that it can't scale.

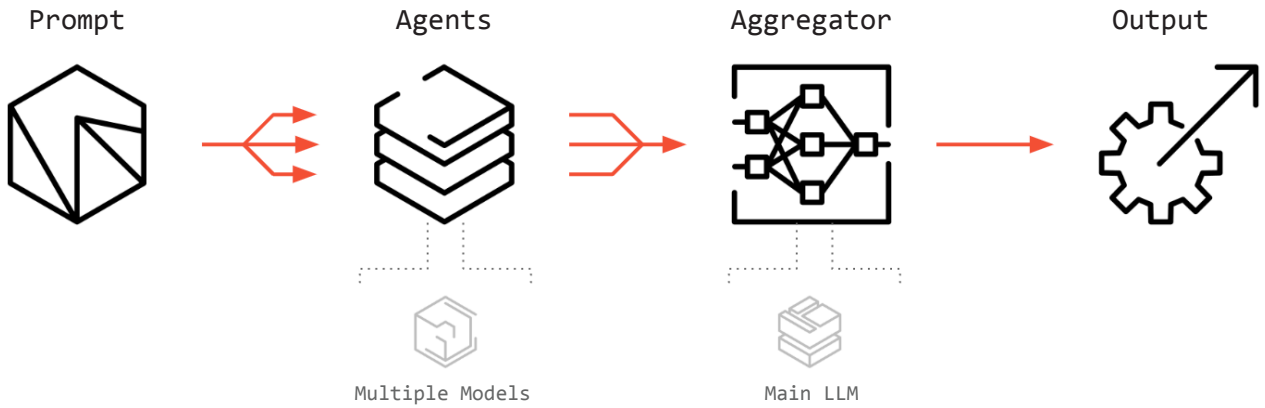
> As AI business technologists consider using agentic workflows to tackle their biggest challenges, they also need to ensure their inference strategy is up to the task, speedwise. Otherwise, **they may build something cool and powerful only to find that it can't scale.**

Groq® LPU™ AI inference technology is the best solution for the task. Its ultra-fast performance, coupled with high affordability and energy efficiency, means Groq technology is seemingly tailor-made for AI agentic workflow solutions – and in a way, it was. Groq was founded with the vision of making inference much faster and more affordable, opening up the world of scaled AI solutions to a wide universe of applications and users.

Let's dive in.

What Are Agentic Workflows?

In a typical AI workflow, the person will make a query to an AI solution, get a response, and that's it. For example, you might ask the AI program to write an email to a particular sales prospect. You must then manually review the response, and from there can ask the AI to refine the response, maybe making it shorter or wittier. Agentic AI workflows take a more sophisticated approach. They are more goal-oriented than task-oriented and can act autonomously to get to the best outcome. In the email example, you might have one agent tasked with drafting the email, another tasked with editing for grammar and brevity, and yet another tasked with asking for your review or sending it.



THERE ARE FOUR GENERAL CATEGORIES OF AI AGENTIC WORKFLOWS[1]:

1 **Reflection: A solution that checks its own work and iterates based on its findings.**

For example, a solution might have a "coder agent" that generates code and a "critic agent" that works to debug it. They could be using the same LLM model, or different models. People using AI engage in reflection all the time: you get a response then ask the model to modify it in someway. Using agents to reflect automates and speeds up the process.

2 **Tool Use: AI solutions that employs tools to expand their capabilities.**

One example of this is how a team of researchers created a model, called Toolformer, trained in how and when to access simple tools such as a calculator, internet search engines, a translation system, and a calendar. These are simple tasks for a human but can be challenging for advanced LLMs. Collaborating with a Toolformer-type model as an agent can vastly improve outcomes and expand the capabilities of an AI solution.

3 **Planning: AI solutions that combine several tasks (usually across different LLMs) to achieve an objective.**

This is what real-life agents do all the time. You tell your travel agent to find you a one week cruise in September that offers the best combination of luxury and price, and they get to work on the trip. That planning may entail several steps: checking calendars to find the best time for you and your partner to go, finding the right cruise and flights, then checking to see if those choices are the best for your criteria.

An AI travel agent could use the same planning process. It automatically breaks down an objective (plan a trip) into a series of tasks (calendar, booking, quality check), and then executes these tasks. The same planning process could be applied more generally to any type of research project, e.g. finding the right academic papers to support a thesis. A planning AI solution is one that can take a complex or vague objective and break it down into a series of tasks to be executed.

4 **Multi-agent Collaboration: AI solutions that leverage multiple agents to collaborate on a task, with each agent having a different role and potentially using a different model.**

The aforementioned reflection and tools use of agentic workflows are both examples of multi-agent collaborations, as is the planning agent. In each case, the agents have specific tasks and operate independently, and then collaborate to produce an outcome.

1. Andrew Ng, Founder of DeepLearning.AI and Stanford University Adjunct Professor, shared these classifications in his review of agentic AI here: https://www.youtube.com/watch?v=sal78ACtGTc

Compounding Latency

Agentic AI applications have the potential to tackle all sorts of new types of problems for enterprises, but a problem can arise with them when they scale. Whereas a slow response of a second or two in a typical "query and response" AI user experience might be annoying, it may not render the solution useless. But in agentic workflows, a slow response compounds as its occurence multiplies – enough to bring the entire user experience crashing down.

Take the simple and powerful Mixture-of-Agents (MoA) implementation developed by Groqster Soami Kapadia. It runs a query through three different AI agents simultaneously. Each agent has a different prompt and runs on a different model. Agent 1 is asked to think through its response step by step and uses Llama 3 8B, Agent 2 is asked to start its response with a thought and runs on Gemma 7B, while Agent 3 is instructed to "always take a logical approach" and also runs on Llama 3 8B.

Once these agents perform their tasks, their various outcomes are aggregated by the main LLM (in this case, Llama 3 70B) to generate a single, optimal result. This can then be fed through the agents again for another layer of review, starting the cycle over, or on to the user. The number of layers is set by the user.

The multi-agent approach generates superior outcomes, but it entails many more LLM calls. Setting the layer parameter to three means the solution will make a total of 10 LLM calls (three agents x three layers + one final LLM review), which means a slow model response is amplified by 10. It takes ChatGPT 3.5 about four seconds to complete its response to the query "write me 10 sentences that end in apple." Multiply that by 10 and you get a 40 second response wait time, which is untenable for most use cases. On Groq, it takes just three seconds.

Agentic AI & Groq

The MoA demo doesn't run on ChatGPT 3.5, it runs on Llama 3 8B, Gemma 7B, and, as its main model that delivers the final outcome, Llama 3 70B. Most important from a speed perspective, it runs on Groq. These factors – smaller models and Groq ultra-low latency – mean the MoA demo will answer the query about sentences ending in apple in about three seconds.

10X the number of tokens generated yet 25% faster. That's the difference between an AI solution that works well in the lab but can't scale and one that makes a transformative impact on enterprises and companies.

The combination of exceptional performance and agentic use of smaller, more efficient models means Groq offers another advantage besides speed: affordability. Task-oriented agents with more limited scope can usually get great outcomes using smaller, often specialized LLMs. A general travel query might need all of a large model's capabilities to generate a high-quality result, but an agent just focused on, say, cruises in the Caribbean can use a much smaller (and cheaper) LLM.

> The combination of exceptional performance and agentic use of smaller, more efficient models means **Groq offers another advantage besides speed: affordability.**

The Groq pricing table reflects this. Our price for inference for Llama 3 70B is $0.79 / one million tokens, while the price for Llama 3 8B is $0.08. But 8B provides a comparable result for many focused tasks, so why not take advantage of that?

For example, Groqster Benjamin Klieger developed an agentic app called Infinite Bookshelf, which writes a full-length book based on a simple query, "What do you want the book to be about?" The solution employs one agent to develop the structure of the book and another to generate the book, one chapter at a time. The structure-generating agent runs on Llama 3 70B – the bigger model is used to develop the best possible narrative outline. However, the chapter-generating agent runs on Llama 3 8B, which delivers strong content at exceptional speed. Infinite Bookshelf's judicious use of agents and models turns what would have otherwise been a costly and slow AI exercise, writing a book, into something that is affordable and timely.

Questions AI Business Technologists Should Be Asking

Agentic AI solutions have the potential to help business technologists apply LLMs and AI to tackle even bigger, more complex enterprise problems.

HERE ARE THE QUESTIONS TO ASK:

Which of my enterprise challenges could be best addressed by agentic AI?

These are likely situations which require multiple disciplines to address and have less specific, more general objectives.

How can I get my AI team started today in learning to develop and scale agentic AI solutions?

Agentic AI is here now. The best AI teams are already learning how to use it. Staying on the sidelines is not the best option.

What is the best inference strategy that delivers the speed and affordability required to scale agentic AI solutions successfully?

Due to their high volume of token generation, agentic AI solutions demand much higher speed and energy efficiency than traditional AI solutions. They also require the flexibility to support multiple models.

Do I have to sacrifice speed for cost?

No. Groq LPU AI inference technology is the ideal inference solution for agentic AI workflow solutions. It provides the speed, affordability, and flexibility to scale agentic solutions and power enterprises to transform their businesses and tackle new challenges.