



Taming the AI Beast: How Big Companies Build Safe AI Agents

Thought Leadership • October 15, 2026

First published on [Enterprise IT World](#).

Authored by Subroto Kumar Panda

Imagine you gave a supersmart robot a credit card and told it to “buy the best stuff for my party.” It might buy balloons and cake, or it might decide the “best” thing is to hire a live elephant and accidentally drain your bank account. That’s the problem big companies face today with Artificial Intelligence (AI). They want AI to do complex jobs, but they absolutely must keep it under control. Here is how the world’s biggest tech companies keep their AI agents

The Inference Layer: Constrained Decoding

This is where the real magic happens. When an AI generates a response, it is literally calculating the math probabilities of which word should come next. The Inference Layer is the software engine that runs this math. Engineers use a technique called Constrained Decoding. Imagine the AI is about to say the word “apple,” but the system rules say it can only output a number. Constrained Decoding reaches into the AI’s math and forces

the probability of the word “apple” to zero. The AI is mathematically blocked from making a mistake or writing computer code when it’s only supposed to write a list of numbers. It’s like bowling with the bumpers up; the ball simply cannot go in the gutter.

The Application Layer: Object Validation (Pydantic / Zod)

Constrained Decoding ensures the AI speaks the correct “language” (JSON), but it doesn’t check if the meaning makes sense. What if the AI outputs a perfectly formatted JSON message that says, “Age: 999 years old”? That’s where the Application Layer comes in. Engineers use software tools like Pydantic (for the Python coding language) or Zod (for JavaScript). These tools act as a strict bouncer at a club. They look at the AI’s output and validate the object. If the age isn’t between 0 and 120, Pydantic immediately rejects the AI’s answer and throws an error before the bad data can break the company’s database. It guarantees that the AI’s data makes logical sense in the real world.



The Network Layer: AI Gateways

In big corporations, the AI doesn't talk directly to the user. All messages pass through a digital checkpoint called an AI Gateway at the Network Layer. Think of it like airport security. Before the AI's answer is allowed to leave the company's servers, the AI Gateway scans it. If the AI was somehow tricked into revealing a customer's Adhar Number, the Gateway detects the sensitive info, blocks the message, and triggers an alarm. It can also act as a circuit breaker; if an attacker is bombarding AI with malicious prompts, the Gateway will cut off their connection completely to protect the system.

Making Safety Fast: The Speed Enhancements

Checking every single word an AI says for safety is used to slow things down massively. If you must pause an AI a thousand times a second to make sure it's not breaking a rule, the computer freezes up. But tech companies hate slow software, so they invented brilliant ways to make this security happen at lightning speed.

Ahead-of-Time (AOT) Grammar Compilation

Instead of figuring out the rules while the AI is talking, engineers do it beforehand. This is Ahead-of-Time Grammar Compilation. When the server first turns on, it looks at the strict JSON rules it needs to follow and builds a map of every possible legal move the AI can make. It's like studying the entire maze before you enter it. Because the computer already knows exactly what shapes the answers are allowed to take, it doesn't have to waste time thinking about it later.

GPU-Native Bitmasking

The "brain" of the AI lives on a super-fast computer chip called a GPU (Graphics Processing Unit). In the old days, to check if a word was safe, the system had to send the word from the fast GPU over to the slower main computer (CPU) to check the rules. This took forever. Now, engineers use GPU-Native Bitmasking. They put the safety rules directly onto the fast GPU chip itself. As the AI thinks of a word, a digital mask drops down instantly, blocking all the bad words right at the source, without ever having to ask the slower parts of the computer for permission.

Structured Speculative Decoding

This is the coolest trick of all. Since the AI is forced to speak in a strict JSON structure, much of what it says is predictable. For example, if it has to output a user's profile, it *must* write the characters {"username": ". Instead of making the huge AI brain slowly calculate those exact characters one by one, the system uses Structured Speculative Decoding. It essentially realizes, "I know exactly what the



AI has to say next,” and it fast-forwards through those predictable parts. By skipping the heavy math for the boring formatting stuff, the AI actually runs faster when it is highly restricted!

Asynchronous Out-of-Band Validation

We talked about Pydantic checking if the AI’s answers make sense (like checking if the age is a real human age). Doing this normally forces the whole system to wait until the AI finishes its entire paragraph before checking the first sentence. To fix this, engineers use Asynchronous Out-of-Band Validation. This means the system checks the AI’s work **while** the AI is still talking, using a separate part of the computer. If the AI

makes a mistake on word 5, the validator catches it immediately and tells the AI to stop, instead of waiting for it to finish a 1,000-word essay. It saves time and computer power

The Final Picture

By combining pre-compiled FSMs (the ahead-of-time maze map), native GPU masking (the lightning-fast word blocker on the chip), and speculative fast-forwarding (skipping the boring parts), enterprise engineers have achieved something amazing. They can force a wild, creative AI to follow strict, mathematically perfect

This is why big companies can confidently deploy AI agents today. They aren’t just hoping for the AI to act nicely but are building a physical, mathematical cage around it—from the Dual-LLM architecture down to the GPU-native bitmasking—ensuring that if the AI tries to do something dangerous, the system catches it, blocks it, and hits the kill switch long before any damage is done.



RELATED INDUSTRIES

[ARTIFICIAL INTELLIGENCE](#)