

Lesson 3B: Capabilities & limitations | AI Fluency: Framework & Foundations Course

7 MIN • YOUTUBE • [HTTPS://WWW.YOUTUBE.COM/WATCH?V=W5CGA7XIPRI](https://www.youtube.com/watch?v=W5cga7xipRI)

<https://www.youtube.com/watch?v=W5cga7xipRI>

SUMMARY

Generative AI, particularly large language models (LLMs) like Claude, demonstrate remarkable versatility in language tasks, enabling users to craft emails, summarize reports, and explain complex topics seamlessly. However, these models also have limitations, including a reliance on outdated training data, potential inaccuracies (hallucinations), and constraints in processing information.

- *LLMs excel at diverse language tasks, including writing, summarizing, and translating.*
- *They can maintain conversational context, recalling earlier discussions like a human.*
- *Models have a knowledge cutoff, meaning they lack awareness of events or developments post-training.*
- *Inaccuracies can arise from training data, leading to "hallucinations" where the AI confidently states incorrect information.*
- *LLMs have a context window limit, restricting the amount of information they can process at once.*
- *Responses can be non-deterministic, resulting in variability in answers to the same question.*
- *Complex reasoning tasks, especially those requiring multi-step logic, remain challenging, though improvements are underway.*
- *Understanding AI's strengths and limitations is crucial for effective collaboration, leveraging human judgment alongside AI capabilities.*

Let's now examine what generative AI can and cannot do. Focusing on LLM such as Claude, think of this as getting to know a new colleague. Understanding their strengths and limitations help you collaborate more effectively. To start, we'll focus on what these systems do remarkably well. You might be amazed at how versatile modern language models can be. They're skilled with language in ways that seemed impossible just a few years ago. Crafting emails that capture your voice, condensing lengthy reports into clear summaries, translating between languages, and explaining complex topics across countless fields from microbiology to marketing strategy.

What's particularly notable is how these models can shift between different tasks without needing additional training. The very same system that helps you write poetry or brainstorm ideas for your birthday party can turn around and help you understand quantum computing concepts or analyze quarterly business trends all through simple conversation. These models can also maintain the thread of a conversation, remembering what you discussed earlier and building upon it.

If you mention your project deadline in passing, for example, and refer back to it later within the conversation, the AI typically understands what you're talking about, much like a human conversation partner would. Many

modern LLMs can now also reach beyond their own knowledge by connecting to external tools and information sources, allowing them to search the web, process files, or even use other applications to enhance their capabilities. This dramatically expands what they can help with. However, just like any technology, LLMs as exist today also have certain limitations. First, AI models are bounded by their training data. LLMs have a knowledge cutoff date based on when they were trained, the point after which they have no innate knowledge of the world. For example, a model with a cut off date of November 2024 means that it wasn't trained on any data after November 2024. Imagine someone who went into a retreat without internet access at a specific date. They wouldn't know about events that happened after they left.

Models need tools like web search to learn more about recent developments. Additionally, the training process doesn't verify every fact in the training data. This means models can sometimes learn and reproduce inaccuracies that were present in their training data. They can also make mistakes when trying to piece together information they've learned. This leads to what is often called a hallucination. AI confidently stating something that sounds plausible but is actually incorrect. Unlike search engines that simply retrieve existing documents, LLMs generate responses based on statistical patterns, sometimes producing hallucinations. Imagine a friend who tells a story with absolute confidence only to have the details completely wrong. AI can sometimes be like that.

Another important constraint is the context window we mentioned earlier. As a reminder, that's the amount of information an AI can process at one time. Every LLM has a maximum limit to how much information it can consider during a single interaction. If this limit is exceeded, the AI won't be able to remember information that falls outside the window, usually on a first in first out basis. Depending on the size of the model, this can limit its ability to process large documents or remember the entire conversation.

Furthermore, unlike traditional software that produces identical outputs given the same inputs, LLM are somewhat unpredictable by default, also known as non-deterministic. Ask the same question twice and you might get slightly different responses each time. This variability stems from the nature of how these models generate text. They're making probabilistic decisions about what text should come next based on patterns in their training data and certain settings that developers can tweak. This creative variability can be great for brainstorming and generating diverse ideas, but requires awareness when consistency or accuracy are critical. Some LLM interfaces also offer settings to control this randomness when needed. This setting is often referred to as temperature. Additionally, while these models are improving rapidly, they've historically shown limitations with complex reasoning tasks, particularly with mathematical or logical problems requiring multiple steps. The good news is that newer reasoning or extended thinking models specifically designed to think stepby step are showing strong progress in these areas. And finally, while models like Claude can now access external tools, they may still lack access to specific data sources or specialized tools that would be needed for certain tasks. It's like having a brilliant colleague who can't access your company's internal database. Their ability to help will be limited no matter how smart they are. If a model doesn't have access to a piece of data or tool that is needed to answer a question, then it should not come as a surprise that it won't be able to help answer the question. The field of generative AI is rapidly evolving.

Researchers are working to address current limitations through techniques like retrieval augmented generation which connects models to external knowledge and data sources as well as expanding their ability to use tools and improving their reasoning capabilities. That said, some limitations will likely remain for the foreseeable future, even if we don't know exactly what those limitations will be. Understanding what AI can or cannot do is essential for AI fluency and helps you determine when and how to best incorporate these systems effectively into your work and daily life. The most effective applications will leverage the complimentary strengths of humans and AI. We bring critical thinking, judgment, creativity, and ethical oversight that AI may struggle to replicate. While AI offers speed, scale, pattern recognition, and the ability to process vast amounts of information. These complimentary strengths will evolve as the technology evolves. That's why continued learning and experimentation are so valuable.

They help you stay a breast of these changes and discover new possibilities. In these exercises across this course, you'll have a chance to explore these concepts firsthand through conversations with Claude. This direct experience will help you develop an intuitive feel for what generative AI can do, can't do, and how best to work with it.