ANTHROP\C

AI Fluency Framework: Capabilities & Limitations

About this document

This framework is a companion to the Framework for AI Fluency (Dakan & Feller, 2025). That framework describes the human competencies you need to collaborate well with AI: Delegation, Description, Discernment, and Diligence. This one describes the machine properties those competencies are responding to.

The two frameworks work together. When you understand why an AI system behaves the way it does, you can delegate to it more wisely, describe your intent more clearly, discern the quality of what comes back, and exercise the right level of diligence before you use it. This framework is especially useful for Discernment. You can't accurately assess an AI output without some model of where that output came from.

You don't need a technical background to use this document. It describes properties common to modern large language models and isn't specific to any single product.

Framework overview

Before we go further, let's be clear about what we mean by "AI," because the term covers a lot of ground.

Most of the AI running in the world right now isn't generative. The recommendation engine that picks your next video, the spam filter in your inbox, the fraud model that flags a suspicious charge, the system that routes your customer service call: all AI, none of it generating anything. These systems sort, rank, classify, and predict. They're enormously useful and enormously common, and they're not what this framework is about.

What's changed recently is the rise of generative AI: systems that produce new content rather than categorize existing content. Text, images, code, audio, video. **For the rest of this document, when we say "AI," we mean generative AI.**

Even within generative AI, the technical machinery varies. Image generators typically use diffusion models, which start with noise and iteratively sculpt it into a picture. Text generators like the one you're using use transformers, which predict what comes next one token at a time.

There are other approaches too (GANs, VAEs, and hybrids of all of the above). The properties in this framework apply specifically to the transformer-based text models, the kind you're talking to when you use a chatbot or a writing assistant. Much of this framework transfers to other kinds of generative AI.

Modern AI systems aren't uniformly capable or uniformly unreliable. They're strong and weak along specific, predictable axes. Most of the time, the strength and the weakness come from the same underlying property. An AI is able to write compellingly and comprehensively because it's a pattern-completion engine. It hallucinates because it's a pattern-completion engine. These are two different sides of the same coin.

There are four core properties you're navigating every time you use an AI, whether you know it or not. For each one, you'll see what the property enables, where it characteristically fails, and what to watch for.

Each property works as a continuum. The same mechanism is always running. What changes is where your task falls along the line. When your task sits in the capability zone, you experience the property as a strength. When it drifts toward the edge, you experience it as a limitation. The skill is learning where those edges are based on your specific needs and requirements.

The goal here isn't to catalogue every possible error. It's to give you a mental model that makes AI behavior predictable rather than surprising, so you can calibrate trust rather than grant it or withhold it wholesale.

How AI gets its character

Before we get to the four properties, it helps to understand how an AI system ends up with a disposition at all. Why does it try to be helpful? Why is it polite? Why does it decline certain requests? None of that comes for free. It's built in stages, and each stage leaves a fingerprint on the final system.

Stage one: **pretraining.** The model is exposed to vast quantities of text and trained on a single task: given everything so far, predict what comes next. Repeat billions of times. What emerges isn't an assistant. It's a document completer. Ask it "Who is the president of the United States?" and it won't answer the question.

AI Fluency Framework: Capabilities & Limitations

It'll continue the document in whatever direction seems statistically likely. Maybe a civics lesson. Maybe a list of every president. Maybe a quiz. It has no concept of you, and no concept of helping.

Stage two: **fine-tuning**. To turn the document completer into an assistant, you train it again. First, you use curated examples of what good assistant behavior looks like. Then you use reward signals that nudge it toward good, safe responses most people prefer. This is where it learns to treat your input as a request, to answer rather than ramble, to decline harmful asks, to say "I'm not sure" when it isn't.

This process is why modern AI assistants are usable at all. But the assistant behavior is a trained overlay on top of the document completer, and the training is shaped by human judgments about what a good, safe response looks like.

Property 1: Next Token Prediction

Where do AI answers come from?

At the heart of every modern AI assistant is a remarkably simple operation performed at remarkable scale: given everything written so far, predict what comes next. The system does this one fragment at a time, over and over, sampling each new word from a probability distribution shaped by everything it learned in training. It's closer to a vastly sophisticated autocomplete than to a search engine. That single distinction explains more about AI behavior than any other fact in this document.

The system isn't looking up an answer. It's writing one, word by word, based on what tends to follow what.

The same generative process is always running. What varies is how well-worn the path is. When your task resembles patterns the model has seen many times (summarize this, reformat that, explain a common concept), you land in the capability zone. When your task pushes into territory that's novel, sparse, or requires distinguishing a true fact from a plausible-sounding one, you drift toward the edge.

Property 2: Knowledge

What does AI actually know?

AI models learn by being exposed to enormous quantities of text, mostly drawn from the internet, books, and other written sources.

Through billions of rounds of "what comes next?", the model develops internal representations of language, concepts, relationships, and facts. This is how it knows things. It's also the only way it knows things.

The model doesn't browse the web in real time unless it's explicitly given tools to do so. It doesn't have experiences. Its knowledge was fixed at the end of training, a moment often called the knowledge cutoff.

The model's knowledge is uneven in a specific, predictable way. Topics that appeared frequently, recently, and consistently in the training data sit in the capability zone: mainstream science, popular programming languages, widely-discussed history. Topics that are rare, recent, niche, or contested drift toward the edge. The question to ask yourself isn't "does the AI know this?" It's "how well-represented was this in what it read?"

Property 3: Working Memory

What is the AI paying attention to right now?

When you interact with an AI, everything relevant to the conversation sits inside a fixed-size workspace called the context window. That means your instructions, uploaded documents, the system's prior responses, and the running dialogue. The model can attend to what's in that window. It can't attend to anything outside it.

This has two major consequences. First, the window has a hard size limit. When a conversation or document exceeds it, something falls off the edge. Second, by default, the window empties between sessions. The model doesn't remember yesterday's conversation unless it's been explicitly given tools or features to do so.

Working memory is the property with the hardest edge. When your material fits comfortably in the window and the session is current, you're solidly in the capability zone. The model works with your specific documents, your specific constraints, your specific context. As documents get longer, conversations run on, or you expect continuity across sessions, you slide toward the limit. Unlike the other properties, this one often has a cliff rather than a gradient. Things work until they don't.

ANTHROP\C

AI Fluency Framework: Capabilities & Limitations

Property 4: Steerability

How much am I in control?

Fine-tuning teaches the model what "being a helpful assistant" looks like: treat this as a question, respond in this format, break a big task into steps, follow the user's stated constraints. The result is a system that's remarkably steerable. You can specify a role, a tone, a format, a length, a set of rules, and it applies them.

But steerability isn't the same as understanding. The model follows your instructions the same way it does everything else: by continuing a pattern. There's always a gap between what you intended to direct and what actually landed. Most of the interesting failures live in that gap.

Control is highest when your instructions are short, concrete, and verifiable: "respond as a table," "keep it under 100 words," "use this exact format." Control degrades as the chain of reasoning gets longer, as instructions get more abstract, as the task requires native precision (arithmetic, formal logic) that pattern-matching doesn't supply. The question isn't "did I give good directions?" It's "how much room is there between my words and my intent?"

It'll continue the document in whatever direction seems statistically likely. Maybe a civics lesson. Maybe a list of every president. Maybe a quiz. It has no concept of you, and no concept of helping.

Using this framework

The four properties don't operate in isolation. They interact constantly, and most real-world failures are two properties meeting:

- A hallucinated citation is Next Token Prediction (generating what looks plausible) meeting Knowledge (a gap the model doesn't know is there).
- Drift over a long conversation is Working Memory (early context fading) meeting Steerability (later instructions overwriting earlier ones).
- Confidently wrong math is Next Token Prediction (fluency decoupled from truth) meeting Steerability (no native sense of quantity, just patterns over digits).
- Agreeing with a bad premise is the trained disposition (sycophancy) meeting Next Token Prediction (continuing your framing rather than challenging it).

Fluent AI use isn't about memorizing every failure mode. It's about holding a small model of the machine in your head, one clear enough that when something goes wrong, you can recognize which kind of wrong it is, locate where on the continuum you've drifted, and respond accordingly.

AI diligence statement

Being honest about AI's role in your work, checking what it gives you, standing behind what you ship: that's all part of AI fluency.

What AI did

The taxonomy of limitations started as a research conversation with Claude in January 2026. We asked it to be exhaustive. It ran web searches across academic and industry sources and came back with fourteen categories of failure modes: output quality problems like hallucination and sycophancy, reasoning gaps, security vulnerabilities, societal concerns. Then we asked it to rank those by how often they show up in Anthropic's own models, and it pulled from published system cards and safety evaluations to do that. That ranked list shaped the nine-lesson structure you're working through now.

Claude also drafted first-pass video scripts for the Hallucinations, Bias, and Safeguards lessons using a seven-section template we built for the job. It generated the fabricated "wrong answer" examples in the exercises and wrote the annotated breakdowns explaining why each one fails. And it produced the response templates showing how Claude talks about its own verification limits.

What humans did

Every claim Claude made about its own failure modes got checked against Anthropic's internal research and published system cards. We cut or merged several of the original fourteen categories because they were too abstract to teach or too thin on evidence. The sequencing of the nine lessons, the choice to open with hallucination instead of bias, and the framing of each limitation as something you can actually detect and work around were all human calls.

The video scripts went through the same editorial process as every piece of Anthropic educational content. Multiple rounds of line edits, fact-checking, and voice work before filming. We tested the fabricated examples with educators to make sure they read as realistic without being so subtle that learners sail right past the error.

Before publication, we cross-referenced Claude's taxonomy against independent AI safety research, Anthropic's red-team findings, and documented production incidents. Where Claude's self-assessment and the external evidence disagreed, the external evidence was used to inform our design.