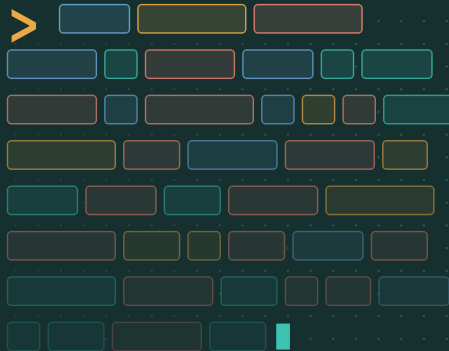


ACC 300 · ARTIFICIAL INTELLIGENCE

How generative AI works



> The model uses two kinds of information

A large language model (LLM) predicts each answer from both.

1 Weights

LEARNED IN TRAINING, BEFORE THE CHAT

- Patterns from past examples of text, stored as numbers.

2 Context

SUPPLIED DURING THIS CHAT

- Your prompt, your files, and search results for this one answer.

> Weights hold what the model learned in training

Training happens before your chat. It sets the weights.

TRAINING

- Past examples of text teach the model patterns.
- Training stores those patterns as numbers called weights.

CUTOFF DATE

- The data stops at a cutoff date.
- The model learned nothing directly after it.

YOUR CHAT

- The weights stay the same during your chat. They change only when the model is trained again.
- You can't use the weights to search for source material.

> Context holds the material for this one answer

The product you use, such as ChatGPT, NotebookLM, or Claude Code, supplies it.

- The context changes with each task.
- An upload can shape this answer. It does not become permanent knowledge.
- A web search adds newer pages to the context. It does not change the weights.

THIS ANSWER'S CONTEXT

your prompt

chat history

uploaded files

pages from a web search

> The model works in tokens

A token is a small piece of text, not always a whole word.

A sentence, split into tokens:

The company reported gross profit .

A long word can be several tokens:

depreciation

- The model reads the tokens in the context and predicts the next token.
- An answer is a long run of those predictions, one token at a time.
- Each model splits text in its own way. These splits are illustrations.

> The model ranks many possible next tokens

The model estimates a probability for every possible next token.

Everyday:

> The sauce has a gross 

taste			likely
flavor			possible
profit			unlikely

> Context changes the probability of the next token

The model estimates a probability for every possible next token.

Everyday:

> The sauce has a gross 

taste		likely
flavor		possible
profit		unlikely

Accounting:

> The company reported gross 

profit		likely
margin		likely
taste		unlikely

→ Context shifts the odds. Likely text still needs verification.

> Probability $\neq$ truth

A likely answer can cite the wrong paragraph.

> Which ASC paragraph requires a valuation allowance for deferred tax assets? ■

MODEL OUTPUT

ASC 740-10-25-6. It applies a more-likely-than-not threshold, a likelihood of more than 50 percent.

CHECK THE PARAGRAPH

ASC 740-10-25-6 uses those words for uncertain tax positions. The valuation allowance requirement is ASC 740-10-30-5(e).

→ The two paragraphs share wording, so the wrong one is a likely answer. Open the cited paragraph and read it.

> The same request can produce different results

LLM systems are non-deterministic in practice.

> Explain why a deferred tax asset may require a valuation allowance. ■

RUN 1

A company reduces a deferred tax asset with a valuation allowance when it is **more likely than not** that some or all of the asset will not be realized. The company weighs **all available evidence, positive and negative**.

RUN 2

A company records a valuation allowance when **future taxable income is uncertain**. The allowance reduces the deferred tax asset to the amount the company **expects** to realize.

> Check both runs against the source

Judge each answer by the paragraph, not by the other run.

ASC 740-10-30-5(e)

"Reduce deferred tax assets by a valuation allowance if, based on the weight of available evidence, it is more likely than not (a likelihood of more than 50 percent) that some portion or all of the deferred tax assets will not be realized."

RUN 1

States the more-likely-than-not threshold and the weighing of all available evidence.

RUN 2

Leaves out both. Uncertain future taxable income is not the test.

→ Compare outputs against evidence, not against each other.

> Many things change between runs

Results can change even when your request stays the same.

Results can change with:

sampling

reasoning effort

hidden instructions

tool results

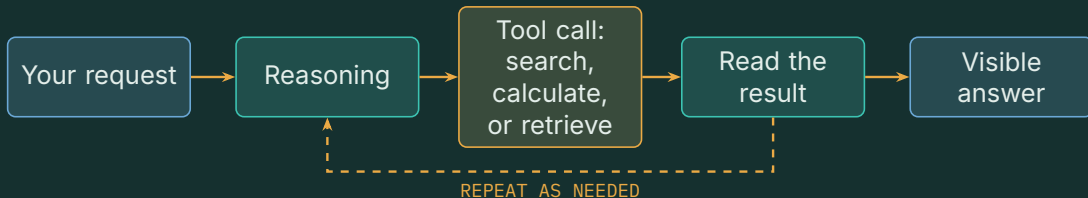
model updates

changed context

- A rerun can be useful for comparison. It does not validate the first answer.
- Two answers that agree can both be wrong.

> Reasoning and tools can precede the answer

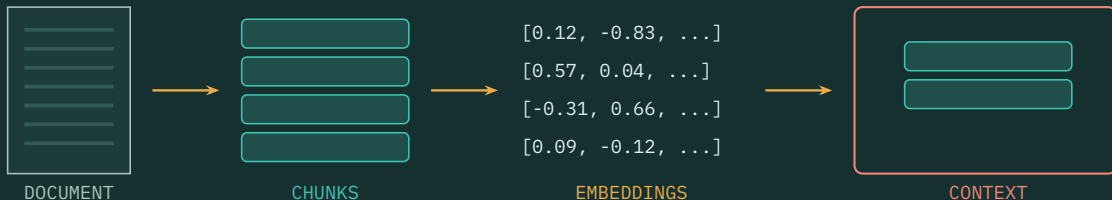
Some products expose a reasoning-effort setting; others choose it automatically.



→ More effort can improve analysis on a hard task. It does not verify the conclusion.

> Large documents enter context in pieces

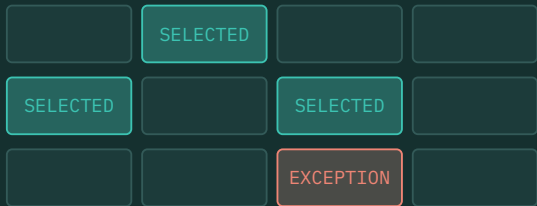
Search and file tools fetch passages for the model to read in this response.



- **Chunking**: a product splits a long document into smaller passages called chunks.
- **Embeddings**: the system represents chunks by patterns of meaning so it can search for related material.
- Retrieved material extends this response. It does not update the model's weights.

> Retrieval can miss an exception

Relevant material may still be incomplete.



- The system selected three chunks. The exception sits in a chunk it did not select.
- The model can also retrieve the wrong passage, omit a qualification, misread a source, or apply it badly.
- Open the source and check what the answer left out.

➤ Language fluency and deterministic calculation are different

Equipment costs \$50,000, has a \$5,000 salvage value, and a 5-year life. Straight-line depreciation.

LLM DRAFT

Cost	50,000
Life	5 years
Annual	10,000

Clear and fluent. It left out the salvage value.

SPREADSHEET, SAME INPUTS

=B1/B2
Annual 10,000

The formula repeats the arithmetic exactly, including the missing input.

YOUR CHECK

Cost	50,000
Salvage	(5,000)
Life	5 years
Annual	9,000

Trace each input to the facts, then recompute.

➔ A tool can make arithmetic deterministic. It cannot make the model's interpretation automatically correct.

> The context window is temporary working space

It holds the tokens active for the current response.

A short chat:

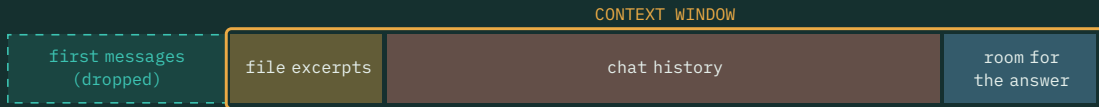


- It includes your prompt, files, retrieved material, chat history, and room for the new answer.

> The context window is temporary working space

It holds the tokens active for the current response.

A long chat:



- It includes your prompt, files, retrieved material, chat history, and room for the new answer.
- Long chats and large uploads can push earlier material out of the window.
- Some products also use saved memory or summaries; those are separate features and may be incomplete or stale. Check coverage against the source.

> The mechanism tells you what to check

Each failure mode calls for a different check.

MECHANISM	WHAT CAN GO WRONG	CHECK
Next-token prediction	A plausible fact or citation that is wrong	Open every paragraph number and quotation the answer cites
Variation between runs	Two answers differ, or agree and are both wrong	Do not count agreement between runs as support
Reasoning and tools	A careful-looking process with a wrong conclusion	Read what the tool returned, not the model's summary of it
Retrieval	A missed exception or qualification	Search the full document for exceptions the answer did not mention
Calculation	Correct arithmetic on the wrong inputs	Trace each input to a stated fact, then recompute
Context window	An earlier instruction or fact dropped	In a long chat, restate the key instructions or start a new chat

> Match the verification to the consequence of the work

The question is not whether AI helped. The question is whether you can support the conclusion.

LOW STAKES

Brainstorming,
plain-language
explanation, formatting

Read it for usefulness.

MODERATE STAKES

Study guide, worked
example, draft analysis

**Check against class
materials and
recompute.**

HIGH STAKES

Accounting conclusion,
submission, client or
public work

**Use authoritative
sources, documented
workpapers, and
human review.**

→ Confidence in the wording does not set the level of assurance.

BEFORE YOU RELY ON AN AI ANSWER

Use AI as a capable collaborator. Keep the work of proof.

1. Ask what the system had access to: its weights only, your files, or a web search.
2. Find the source: the filing page, the ASC paragraph, or the class example.
3. Recompute every amount from the facts.

INSTRUCTIONS

FILE EXCERPTS

CHAT HISTORY

NEW ANSWER

CONTEXT WINDOW