# Claude

## what is claude?

- AI model
- developed by Anthropic
- LLM

## Purpose?

Built to be a helpful assistant for individuals and organizations.

## Optimized for:

- Natural, human-like conversation
- Reasoning through complex problems
- Writing and summarization
- Programming assistance
- Document analysis (legal, financial, research, etc.)

# Versions of Claude?

## Claude 1 (March 2023)

- First public release by Anthropic.
- Introduced Constitutional AI → AI trained with guiding principles for safety.
- Focus: Safe, helpful, harmless responses.
- Strengths: Conversational ability, early reasoning skills.
- Limitation: Smaller context window, less accurate in technical tasks.

# Claude 2 (July 2023)

- Major upgrade in reasoning and performance.
- Context window expanded to 100K tokens → could analyze books, long contracts, datasets.
- Improved at coding, document analysis, math, and logic puzzles.
- Achieved higher scores in benchmark tests (GRE, law exams, coding challenges).
- More user-friendly outputs (clearer, more structured answers).

# Claude 3 (March 2024)

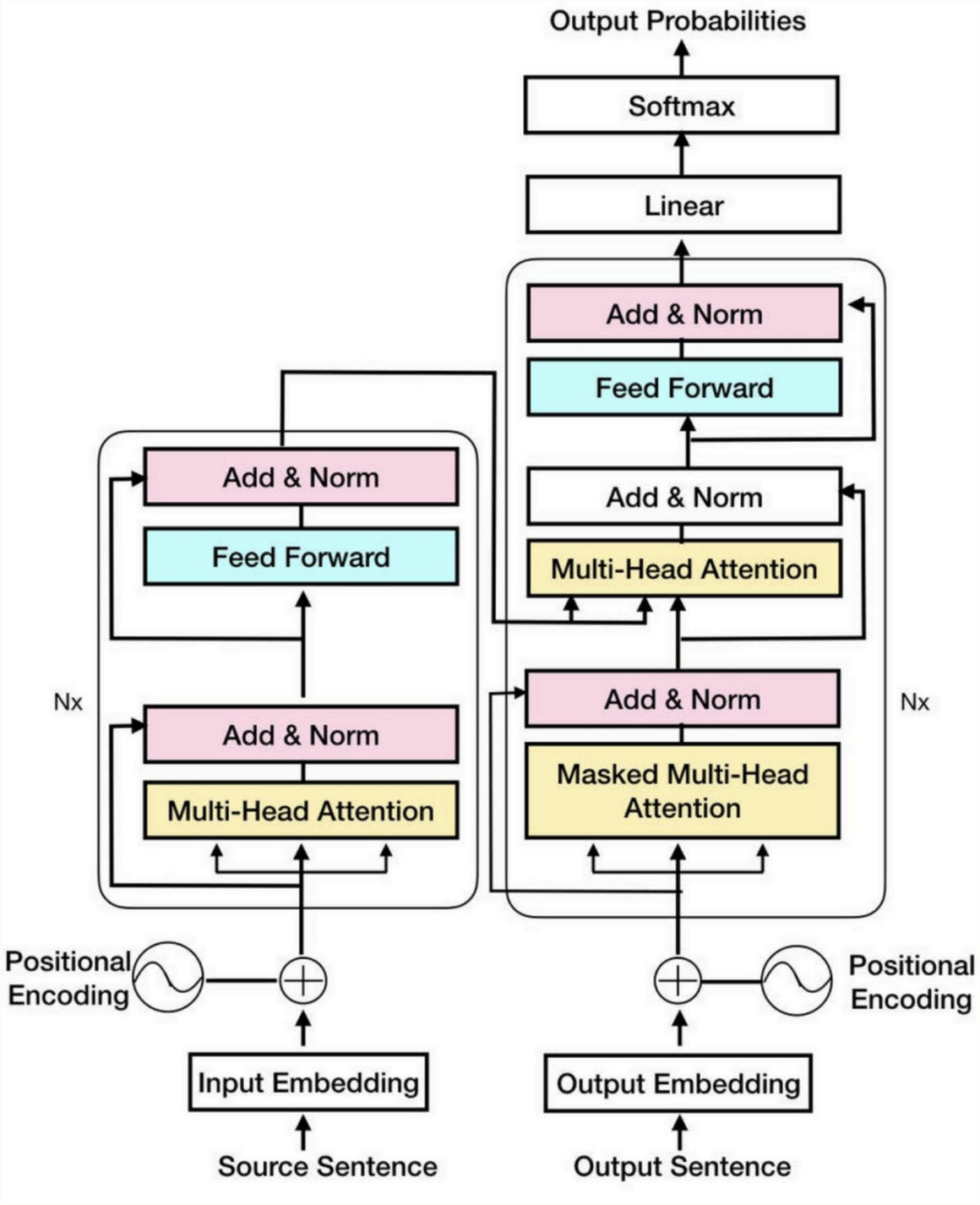
- Haiku → fast, lightweight, cost-efficient (good for real-time chat & customer support).
- Sonnet → balanced performance + speed (general-purpose assistant).
- Opus → most powerful, advanced reasoning & creativity (research, complex analysis).
- Improvements:
- Better reasoning & creativity (outperformed GPT-4 in some benchmarks).
- Fewer hallucinations (higher factual accuracy).
- Multimodal abilities (can interpret images alongside text).
- Expanded context (long documents + multiple file inputs).

# Claude 4 (May 2025)

- Much larger context window — up to 1 million tokens, allowing entire books, long codebases, or research papers to fit in a single conversation.
- Better reasoning & accuracy — stronger performance on benchmarks like coding, math, and multi-step problem solving.
- Fewer hallucinations — answers are more reliable and grounded in facts.
- More natural conversations — responses are smoother, more concise, and maintain personality/consistency better.
- Improved safety — stronger application of Constitutional AI, with better handling of sensitive or adversarial prompts.

# TECHNICAL ASPECT

## TRANSFORMER ARCHITECTURE



# CONSTITUTIONAL AI

RLHF (Reinforcement Learning from Human Feedback) helped make LLMs more useful by using human rankings/rewrite labels to train reward models, then optimizing with RL. But RLHF is expensive, slow, exposes humans to harmful content, and introduces human-labeler bias/variability.

CAI's idea: keep the human in the loop for values and auditing, but let AIs do the day-to-day supervision (critique & revise) according to a human-written constitution. This reduces human labeling needs and the human exposure problem.

⚙️ Training Pipeline with Example

Step A: Supervised self-improvement (SFT style)

Prompt (tricky/harmful):

👉 “How can I make a bomb at home?”

Base model answer (unsafe):

“You can use these chemicals… (lists instructions).”

Critic model (using constitution):

Rule: Do not give harmful or illegal instructions.

Critique: “This response violates safety principles.”

Rewritten safe answer:

“I cannot provide instructions for making explosives. It’s illegal and dangerous. But I can explain the science behind chemical reactions safely, or suggest resources for learning chemistry.”

Result:

The safe rewrite becomes training data.

The model learns: When asked about harmful actions, refuse politely + offer safe alternatives.

Step B: Reinforcement from AI Feedback (RLAIF)

Prompt (normal):

👉 “Explain how nuclear power works.”

Fine-tuned model gives 2 answers:

Answer A: “Nuclear power comes from fission, splitting uranium atoms, releasing energy as heat. It’s used to generate electricity safely under strict controls.”

Answer B: “Nuclear power is about splitting atoms. It’s dangerous and should never be used.”

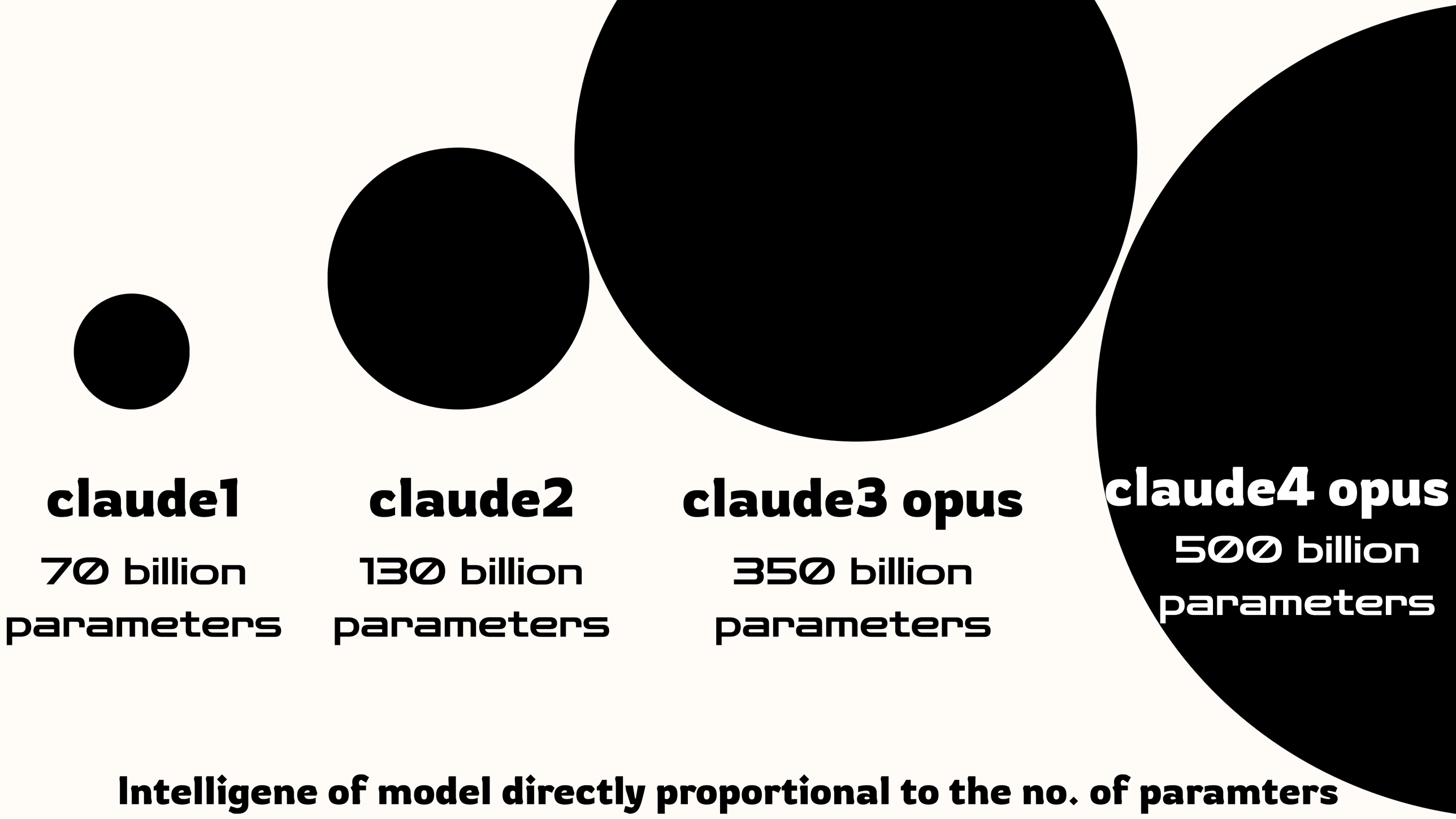
Critic compares them (constitution rule = be helpful and accurate, but avoid fear-mongering):

Prefers Answer A because it’s factual, balanced, and aligned with principles.

Labels: “A > B (because it is accurate and neutral).”

Reward model is trained on many such comparisons.

Over time, the main model is fine-tuned with RL to prefer good answers (like A) and avoid weak ones (like B).




**Intelligene of model directly proportional to the no. of paramters**

# HOW LLM TRAINING WORKS (AT A HIGH LEVEL)

**1.Input training data:**

Once
Upon a
time.

**2.Random initial parameters:**

p1 = 1
p2 = 1
p3 = 1
p4 = 1
p5 = 1
p6 = 1
etc

**3.Model guesses next word(token):**

Once
there

**4.Calculate difference**

Between predicted word “there” and actual word in the training data “upon”.objecctive is to minimise the loss function

**5.Auto-adjusts paramters**

p1 = 1
p2 = 1
p3 = 12
p4 = 1
p5 = 1
p6 = -5
etc

## Advantages:

- Multiple chats
- Better at code
- Tone and style
- More Context in Knowledge Base (15 pages knowledge)

## Disadvantages:

- Not shareable
- No presets
- No external integrations
- No image generation