SkyeStaq.ai

info@skyestaq.ai

Anatomy of an LLM Turn

From keystroke to response... what actually happens when you talk to an AI...

"What is a Turn?"
A turn in AI is one complete cycle: your message in, the AI's response out. Every turn re-reads the entire conversation history, which is why longer chats consume more resources and eventually hit limits.

1 Your Message

- You hit send
- Your text gets tokenized (broken into chunks, roughly 1 token per 4 characters in English)
- Attached files/images also get tokenized

2 Context Assembly

- System Prompt — Core instructions (hidden from you)
- Project Instructions — Your custom instructions
- Conversation History — All prior messages (grows each turn)
- Tool Descriptions — Available tools (search, code, etc.)
- Current Message — What you just sent

Tool Execution

If LLM calls a tool:

- Tool executes (web search, file creation, etc.)
- Tool result gets added to conversation history
- Process loops back to step 2 with the tool result now in the input
- LLM generates another response
- Each loop = new input tokens + new output tokens

Optional Tool Usage

3 Processing

- The assembled context gets fed into LLM's neural network
- LLM processes all input tokens simultaneously (not sequentially like reading)
- The model generates a probability distribution for what token should come next

4 Response Generation

- LLM starts producing tokens one at a time
- Each new token is:
  - Generated based on all input + previously generated output
  - Added to the output
  - Counted toward output tokens
- This continues until LLM:
  - Decides it's finished (generates a stop token)
  - Hits a length limit
  - Calls a tool (tool calls count as output tokens)

5 Final Response

- You see LLM's complete response
- Both input and output tokens are logged

Key Insights

- Project instructions re-read every turn
- Conversation history grows; input tokens snowball
- Tool calls = extra input+output cycles
- Context window is finite (~200K tokens); old messages eventually truncated

Comprehension of a "turn" is foundational to understanding the concepts of token engineering and token usage.

Note: Illustrated using Anthropic Claude's terminology. The underlying process is the same across all foundational LLMs.

February 2026