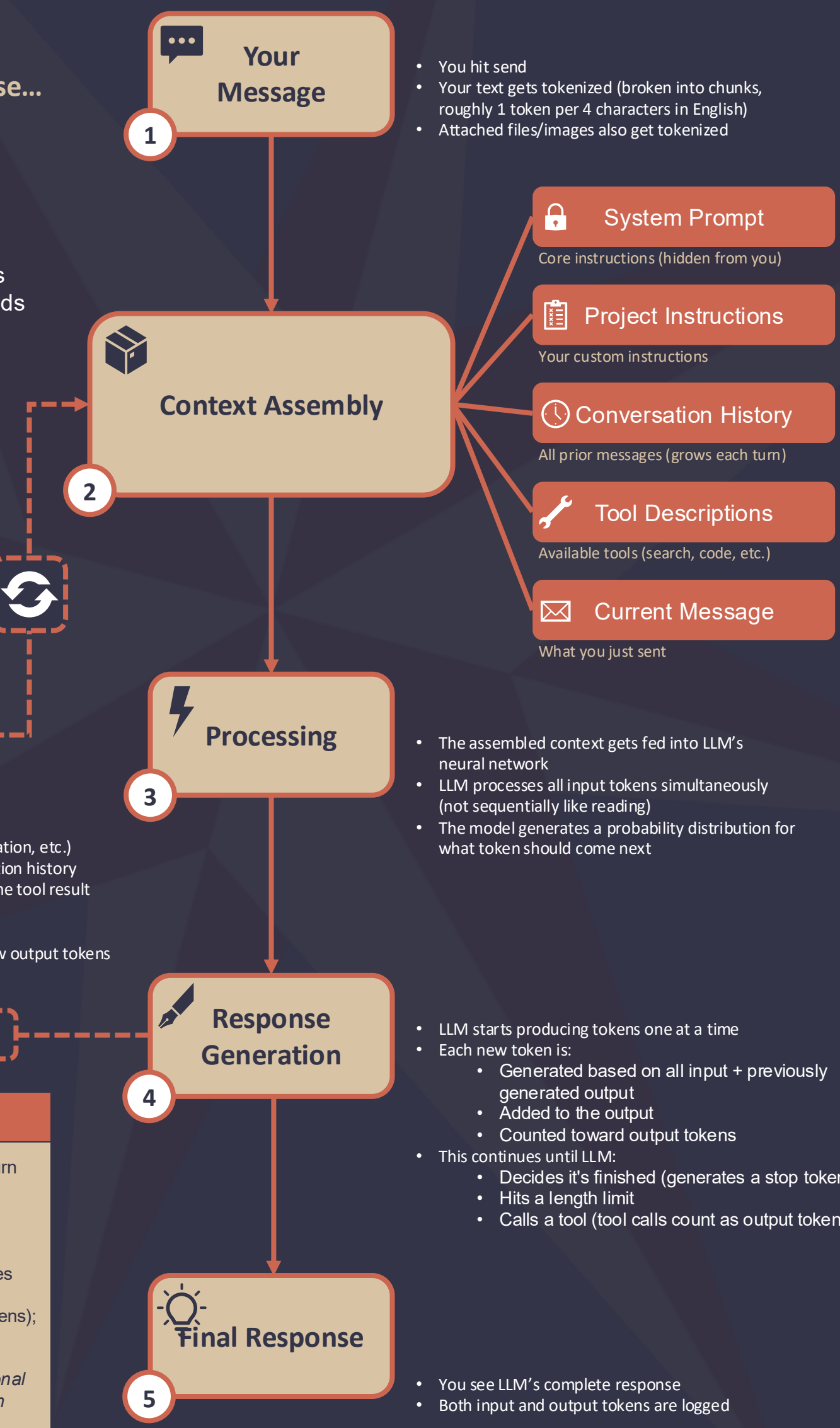
SkyeStaq.ai

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.



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