Context Engineering for Agents

LangChain
129K subscribers

Subscribe

55 Share Clip Save

617 views Jul 2, 2025
Agents need context (e.g., instructions, external knowledge, tool feedback) to perform tasks. Context engineering is the art and science of filling the context window with just the right information at each step of an agent's trajectory. In this video, we break down some common strategies — write, select, compress, and isolate — for context engineering by reviewing various popular agents and papers. We then explain how LangGraph is designed to support them!

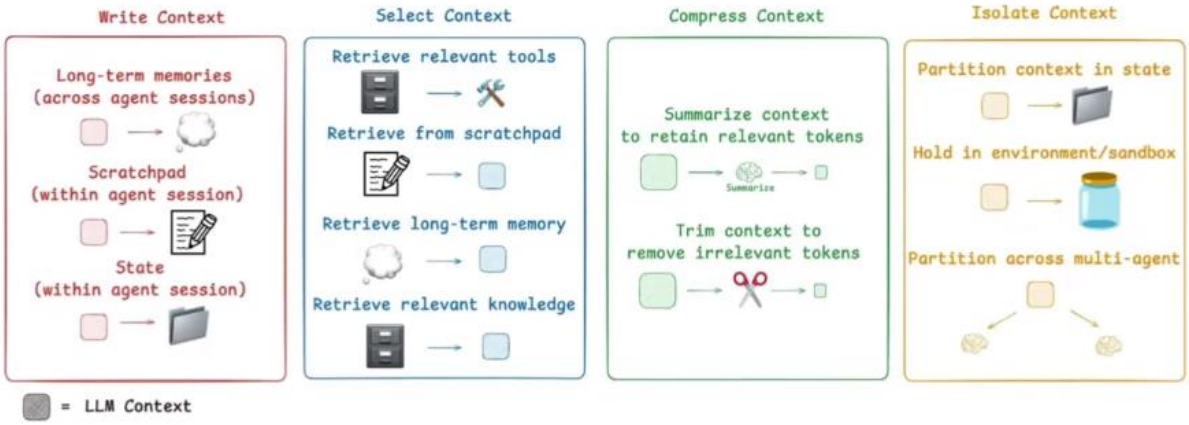
Context Engineering for Agents

Context Engineering

- Agents need context (e.g., instructions, external knowledge, tool feedback) to perform tasks

> ***Context engineering is the art and science of filling the context window with just the right information at each step of an agent's trajectory***

- Common strategies
  - *Writing context - saving it outside the context window to help an agent perform a task.*
  - *Selecting context - pulling it into the context window to help an agent perform a task.*
  - *Compressing context - retaining only the tokens required to perform a task.*
  - *Isolating context - splitting it up to help an agent perform a task.*
- LangGraph is designed to support them



Context Engineering Defined

tobi lutke @tobi · Follow

I really like the term "context engineering" over prompt engineering.

It describes the core skill better: the art of providing all the context for the task to be plausibly solvable by the LLM.

8:01 PM · Jun 18, 2025

8.1K Reply Copy link

Read 337 replies

Andrej Karpathy @karpathy · Follow

+1 for "context engineering" over "prompt engineering".

People associate prompts with short task descriptions you'd give an LLM in your day-to-day use. When in every industrial-strength LLM app, context engineering is the delicate art and science of filling the context window Show more

> **tobi lutke** @tobi
> I really like the term "context engineering" over prompt engineering.
>
> It describes the core skill better: the art of providing all the context for the task to be plausibly solvable by the LLM.

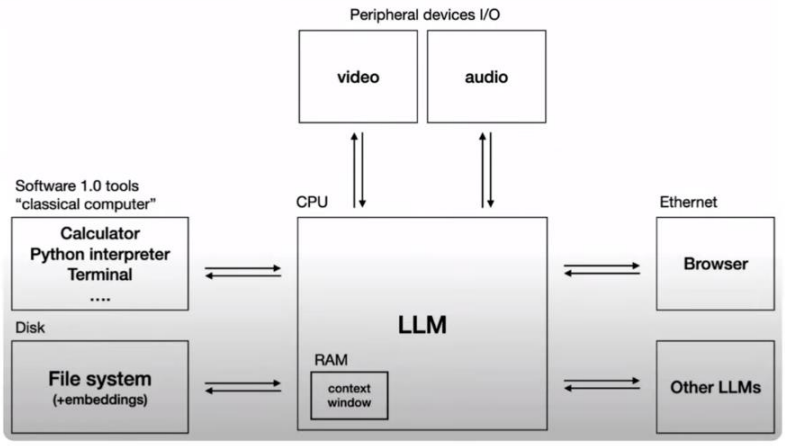
Definition

Markdown Copy Caption

```
[Context engineering is the]"…delicate art and science of filling the context wi
ndow with just the right information for the next step."
```

Analogy

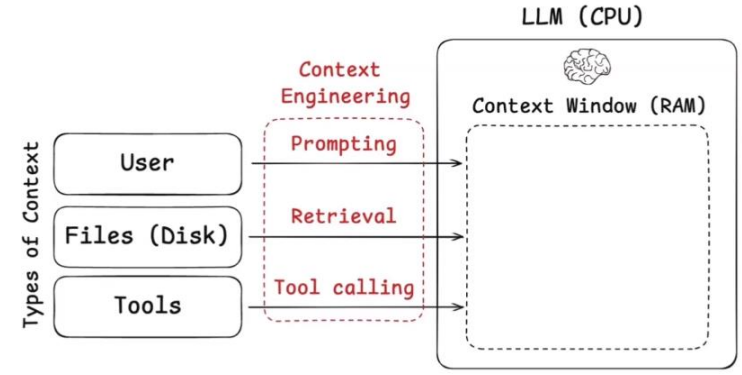
- Karpathy: LLMs are a new kind of OS
  - LLM is CPU
  - Context window is RAM or "working memory" and has limited capacity to handle context
  - Curation of what fits into RAM is analogous to "context engineering" as mentioned above



Types of context

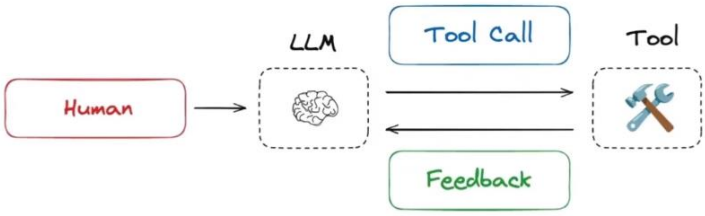
Umbrella discipline that captures a few different types of context:

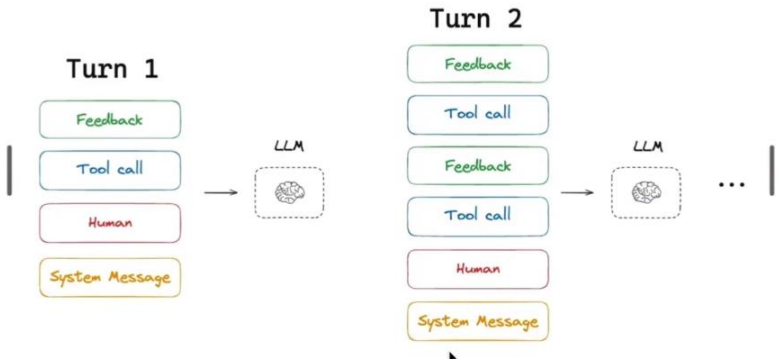
- **Instructions** – prompts, memories, few-shot examples, tool descriptions, etc
- **Knowledge** – facts, memories, etc
- **Tools** – feedback from tool calls



Why this is harder for agents

- Long-running tasks and accumulating feedback from tool calls
- Agents often utilize a large number of tokens!





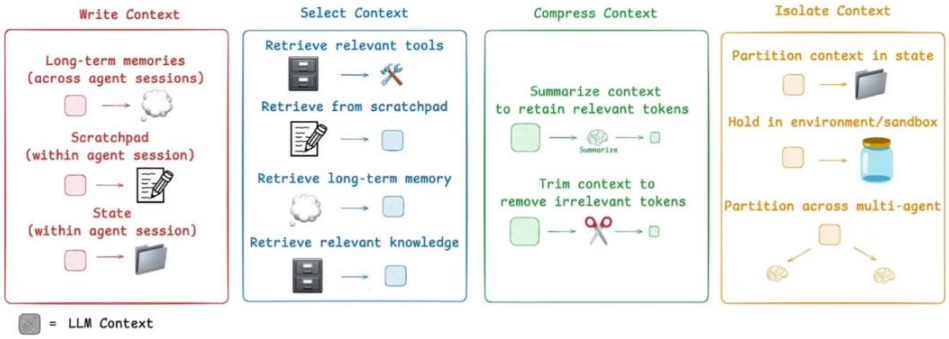
- Drew Breunig nicely outlined longer context failures:
  - Context Poisoning: When a hallucination makes it into the context
  - Context Distraction: When the context overwhelms the training
  - Context Confusion: When superfluous context influences the response
  - Context Clash: When parts of the context disagree
- Context engineering is critical when building agents!

Cognition **Cognition | Don't Build Multi-Agents**

> *Context Engineering is effectively the #1 job of engineers building AI agents.*

Approaches

- *Writing context means saving it outside the context window to help an agent perform a task.*
- *Selecting context means pulling it into the context window to help an agent perform a task.*
- *Compressing context involves retaining only the tokens required to perform a task.*
- *Isolating context involves splitting it up to help an agent perform a task.*



Write

- *Writing context means saving it outside the context window to help an agent perform a task.*
- When humans solve tasks, we take notes and remember things for future, related tasks
- Notes → Scratchpad
- Remember → Memory

Scratchpads

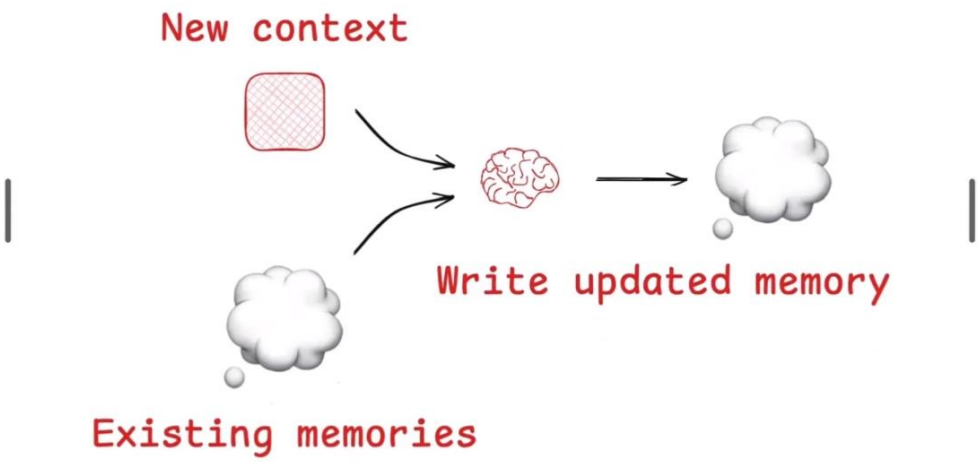
- Persist information while an agent is performing a task
- Anthropic's multi-agent researcher

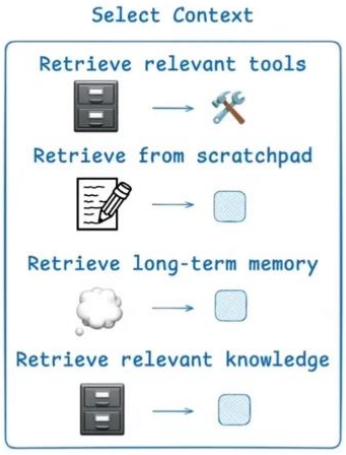
> *The LeadResearcher begins by thinking through the approach and saving its plan to Memory to persist the context, since if the context window exceeds 200,000 tokens it will be truncated and it is important to retain the plan.*

- Use a runtime state object or file

Memories

- Generative Agents synthesized memory from collections of past agent feedback
- ChatGPT, Cursor, and Windsurf all auto-generate memories





Select

- *Selecting context means pulling it into the context window to help an agent perform a task.*

Scratchpads

- Tool call
- Read from state

Memories

- Few-shot examples (episodic memories) for examples of desired behavior
- Instructions (procedural memories) to steer behavior
- Facts (semantic memories)

Memory Type	What is Stored	Human Example	Agent Example
Semantic	Facts	Things I learned in school	Facts about a user
Episodic	Experiences	Things I did	Past agent actions
Procedural	Instructions	Instincts or motor skills	Agent system prompt

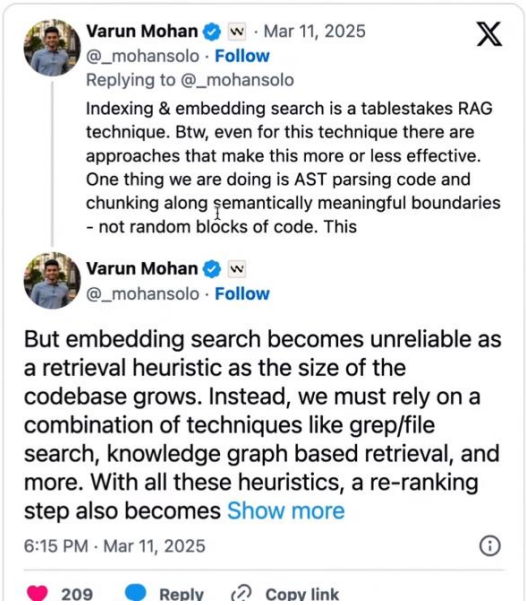
- Instructions → Rules files / CLAUDE.md
- Facts → Collections

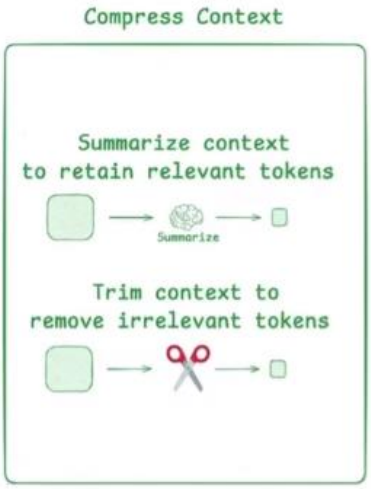
Tools

- RAG on tool descriptions: recent papers have shown this can improve selection 3x

Knowledge

- RAG is a big topic
- Code agent some of the large-scale RAG apps



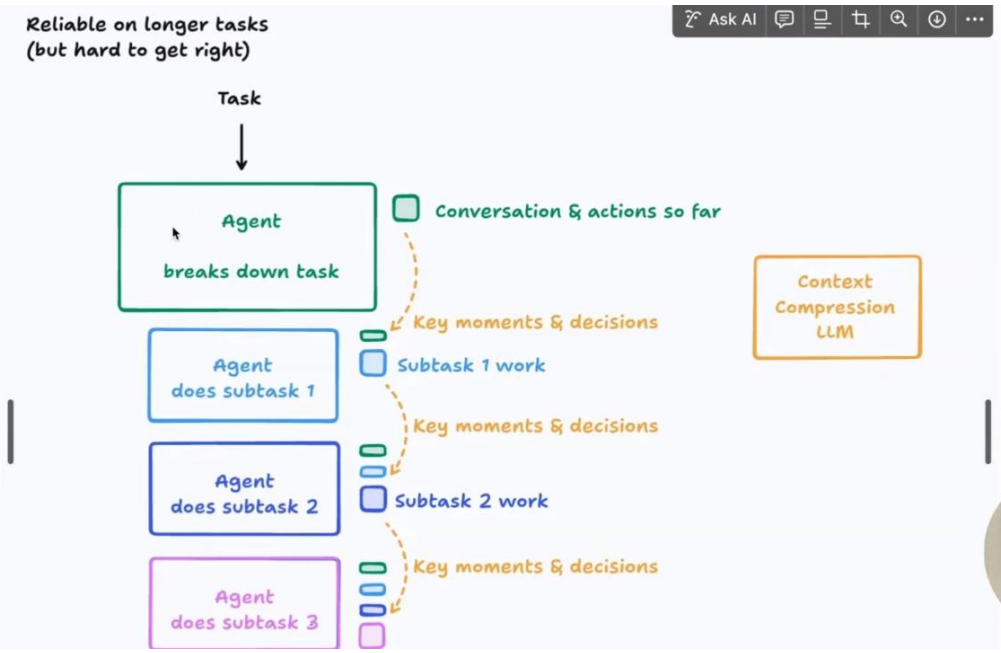


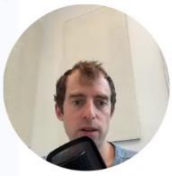
Compress

- *Compressing context involves retaining only the tokens required to perform a task.*

Summarization

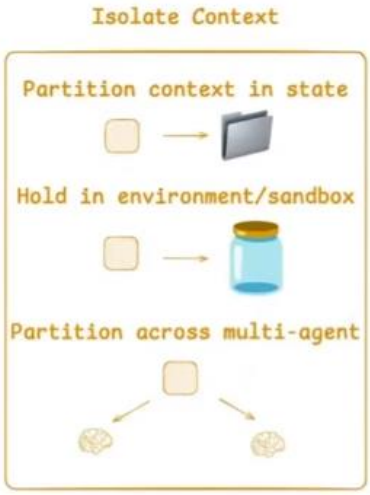
- Claude Code "auto compact" Anthropic Manage costs effectively - Anthropic
- Completed work sections AnthropicAI How we built our multi-agent research system
- Passing context to linear sub-agents Cognition Cognition | Don't Build Multi-Agents





Trimming

- Heuristics: Recent messages
- Learned: arXiv.org Provence: efficient and robust context pruning for...



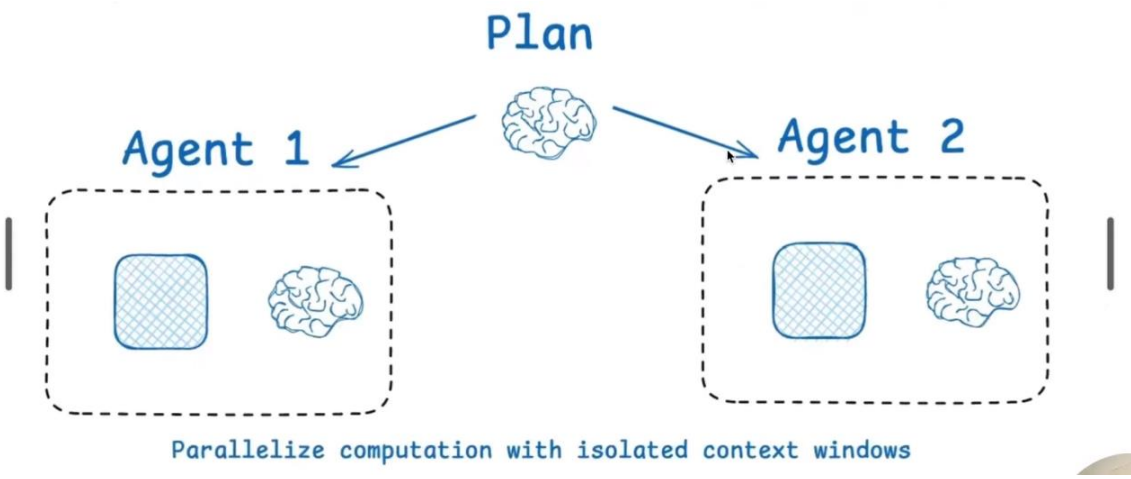
Isolate

- *Isolating context involves splitting it up to help an agent perform a task.*

Multi-agent

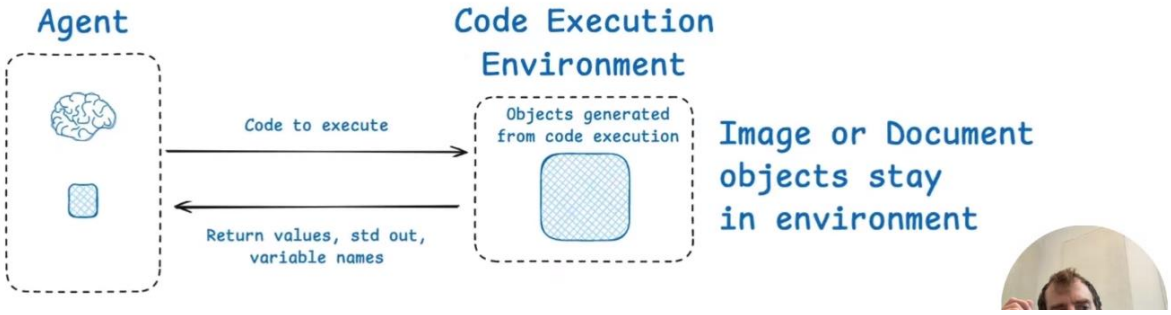
- swarm "separation of concerns" where each agent has their own context
- AnthropicAI How we built our multi-agent research system

> *[Subagents operate] in parallel with their own context windows, exploring different aspects of the question simultaneously.*



Environment

- huggingface Open-source DeepResearch – Freeing our search agents



State

- Models - Pydantic

Context Engineering + LangGraph

Tracing + Eval

- Get started with LangSmith | 🦜🛠️ LangSmith

Write

Scratchpad

- Checkpointing to persist agent state across a session

Memory

- Long-term memory to persist context *across many sessions*

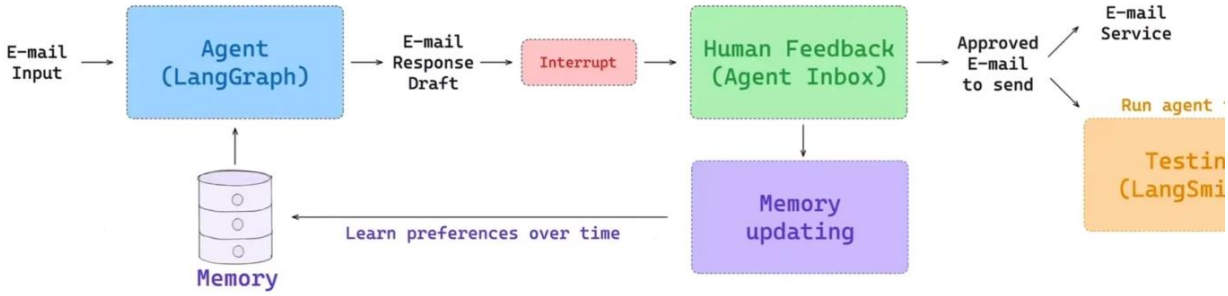
Select

Scratchpad

- Retrieve from state in any node

Memory

- Retrieve from long-term memory in any node
- deeplearningai_ Long-Term Agentic Memory with LangGraph - DeepLearning.AI
- LangChain Academy Building Ambient Agents with LangGraph



Tools

- langgraph-bigtool

Knowledge

- Agentic RAG

Compress

Summarization + Trimming

- Summarizing, trimming message history: Add memory
- Low-level framework, gives flexibility to define logic within nodes
  - Post-processing tool execution: utils.py langchain-ai/open_deep_research

Isolate

Multi-Agent

- langgraph-supervisor-py
- langgraph-swarm-py
- LangChain Conceptual Guide: Multi Agent Architectures
- LangChain Multi-agent swarms with LangGraph
- LangChain Hierarchical multi-agent systems with LangGraph

Environment

- LangGraph + E2B GitHub GitHub - jacoblee93/mini-chat-langchain
- Pyodide LangChain LangChain Sandbox: Run Untrusted Python Safely for AI Agents

State

- State object: Overview Define graph schema

Summary

- *Writing context means saving it outside the context window to help an agent perform a task.*
- *Selecting context means pulling it into the context window to help an agent perform a task.*
- *Compressing context involves retaining only the tokens required to perform a task.*
- *Isolating context involves splitting it up to help an agent perform a task.*