Agentic Engineering

Working With AI, Not Just Using It

Kilo Code

The Paradigm Shift

From autocomplete to teammates

The Paradigm Shift

From autocomplete to teammates

2020

Autocomplete

Finish your line

2022

Copilots

Suggest functions

2025+

Agents

Execute tasks

90% of tech workers now use AI at work

— Google DORA Report

The mental model shift:
AI as an **energetic, enthusiastic junior developer**

> *"We are no longer just using machines, we are now working with them."*
>
> — Armin Ronacher, creator of Flask

The Agent-Model Architecture

Three-way collaboration: You + Agent + Model

You

Direction & judgment

The thinking

Agent

Runs locally in your IDE

Kilo Code, Cursor, Claude Code

Model

Runs in the cloud

Claude, GPT, Gemini

⚠️ Agents are an **illusion** — models are stateless

The only way to get better output is to put better tokens in

Context Engineering

What you put in determines what you get out â□□ and it costs more than you think

> *"Context engineering is the delicate art and science of filling the context window with just the right information for the next step."* *â□□ Andrej Karpathy*

Problem 1: Context is expensive

Every token you add to the context window costs money â€” and the cost scales fast.

10K	200K	600K	1M+
tokens	tokens	tokens	tokens
~$0.03	~$0.60	$1.50+	ðŸ’¸ðŸ’¸ðŸ’¸

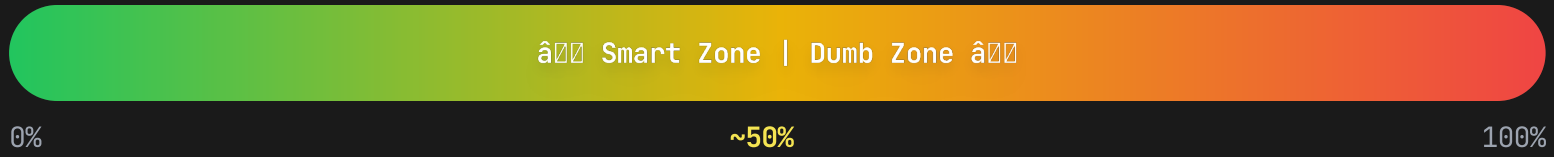
More context â†’ better results

You're paying for noise, not signal

And even if cost weren't an issue â€” more context actually makes the model worse.

Problem 2: More context makes the model dumber

It's not just about money. Quality degrades well before you hit the context limit.



â ï¸ The MCP Trap

Loading every MCP server you have means you're starting every task already in the dumb zone â before you've typed a single word.

â The Benchmark

Stay under 50% context utilization when possible. The model gets worse and your costs get worse as you approach the limit.

â Dex Horthy

But context bloat isn't the only risk. Sometimes the problem isn't quantity â it's quality.

Problem 3: Bad context poisons everything

Once bad information enters the context window, the model treats it as ground truth â□□ and builds on it.

â□ ï¸□ Common sources

Task mixing
Leftover context from a previous, unrelated task

Outdated comments
Stale docs and TODOs that no longer reflect reality

Massive debug logs
Thousands of lines of noise drowning out the signal

Compounding hallucinations
An early AI mistake that every subsequent response builds on

â□□ The fix

ð□□□

Start a new session

You can't think your way out of a poisoned context. Don't try to fix it â□□ reset it.

So context is expensive, it degrades quality, and bad context corrupts outputs. How do you manage it?

The solution: four ways to manage context

Every context engineering decision falls into one of these categories.

â□□ï¸□ Write Context

Persist information *outside* the window so it's available without filling it up

`Scratchpads, memory files, CLAUDE.md`

ð□□¯ Select Context

Pull only what's relevant for *this step* â□□ not everything that might be useful

`RAG, @-mentions, file references`

ð□□□ï¸□ Compress Context

Reduce what's already in the window to just the tokens the task actually needs

`Summarization, trimming logs`

ð□□□ Isolate Context

Split work across separate sessions so no single window accumulates everything

`Parallel agents, task separation`

The habit that ties it together

You don't need to think about all four strategies on every task. You need one default behavior.

ð□□±

One task per session

Finish the task, close the session. Start fresh for the next one.

ð□□□

Watch the meter

If you're past 50%, you're already in the dumb zone. Wrap up or restart.

ð□□□

When in doubt, restart

Something feels off? Don't debug the AI. Start a new session.

Frequent new sessions aren't a sign of confusion â□□ **they're a sign of discipline.**

"Context engineering is the delicate art and science of filling the context window with just the right information for the next step." — *Andrej Karpathy*

The Cost Reality

Context scales quadratically — twice the context requires 4x computational power

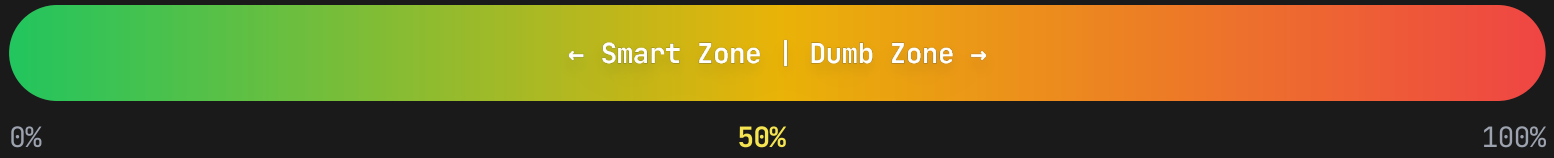
10K	200K	600K	1M+
tokens	tokens	tokens	tokens
~$0.03	~$0.60	$1.50+	💸💸💸

More context ≠ better results

You're paying for noise, not signal

The "Dumb Zone"

Around 40% context utilization, you start seeing diminishing returns



⚠️ The MCP Trap

Too many MCP servers = working entirely in the dumb zone

Each server adds to your system prompt

✅ The Rule

Stay under 50% context utilization when possible

— Dex Horthy

Context Poisoning

Once mistakes enter context, they contaminate everything

☠️ Sources of Poison

Task mixing
Unrelated context from previous tasks

Outdated comments
Stale documentation and TODOs

Massive debug logs
Thousands of irrelevant lines

Hallucinations
AI mistakes that compound

✅ The Only Reliable Fix

Start a new task

Clean context = clean thinking

Don't try to "fix" poisoned context — just start fresh

The Four Strategies

✍️ Write Context

Save information outside the context window

`Scratchpads, memories, plan files`

🎯 Select Context

Pull the right information into the window

`RAG, file references, @-mentions`

🗜️ Compress Context

Retain only tokens required for the task

`Summarization, trimming`

🔀 Isolate Context

Split context across sub-agents or sandboxes

`Parallel agents, task separation`

The Rule of Thumb

One task per session

Start fresh for each new piece of work

Watch the meter

Keep context under 50% — stay in the smart zone

When in doubt, restart

A new session costs nothing. A poisoned context costs everything.

Frequent new sessions aren't a sign of confusion — **they're a sign of discipline.**

Research → Plan → Implement

The workflow that went viral on HackerNews

The Classic Mistake

❌

What most people do

"Hey AI, implement this feature"

Jump straight to code → wrong assumptions → wasted time → frustration

✅

What works

Research → Plan → Implement

Understand first → explicit steps → execute with confidence

"A bad line of research = potentially hundreds of bad lines of code"

— Dex Horthy

Phase 1: Research

🔍 What You Do

Understand the system
How does this actually work?

Find the right files
Where does this code live?

Stay objective
No implementation yet — just learning

📄 Output

Research Document

- Exact files involved
- Line numbers for key functions
- How data flows through the system
- Dependencies and side effects
- Edge cases to consider

Use **Ask Mode** — prevents overeager changes

Phase 2: Plan

📝 What You Do

Outline exact steps
File names, function names, code snippets

Include testing
How to verify after every change

Be explicit
What will change, what won't

📄 Output

Plan File

- Step-by-step instructions
- Specific code changes
- Test commands for each step
- Rollback strategy

> So clear that even a "dumb model" couldn't screw it up

Use **Architect Mode** — planning and design focus

Phase 3: Implement

⚡ What You Do

Execute the plan
Step by step, no improvising

Keep context low
Fresh task for each major step

Review each change
Before moving to the next step

Commit frequently
Small, reversible changes

🎯 The Key Insight

Human review at **research** and **planning** stages is the highest leverage use of your time

By the time you're implementing, the hard thinking is done

Use **Code Mode** — file editing and implementation

"AI cannot replace thinking. It can only amplify the thinking you have done — or the lack of thinking you have done."

— Dex Horthy

Configuring Your Agent

Modes, rules, and customization

Modes and Customization

Same model, different instructions — modes define the AI's role

❓

Ask Mode

Questions & exploration

Prevents overeager changes

💻

Code Mode

File editing & implementation

The workhorse

🏗️

Architect Mode

Planning & design

Big picture thinking

🌍 Global Rules

Apply to all projects

📁 Workspace Rules

Project-specific context

🎭 Mode Rules

Behavior per mode

Top Settings & Configuration

Settings that can make a difference in your Kilo experience

🤖 Agent Behavior

- Modes configuration
- MCP servers
- Rules & workflows
- Skills

⚙️ Auto-Approve

- Check-in frequency
- Read/write permissions
- Execution controls

✨ Autocomplete

- Chat autocomplete
- Background editing (experimental)

💡 Start with defaults, then tune as you learn what works for your workflow

Mental Model for Agent Configuration

Where does configuration belong in your repo?

📋 AGENTS.md

Always-on rules for the repo

`"README for AI agents"`

- Project conventions
- Build commands
- Testing requirements

🎯 SKILLS.md

Task-specific workflows (on-demand)

`Reusable playbooks`

- "api-design"
- "code-review"
- .kilocode/skills/

🎭 Modes

Personas with different behaviors

`Role-based configurations`

- Architect
- Coder
- Debugger

📢 Memory Bank is being deprecated in favor of AGENTS.md + SKILLS.md

Kilo Code Power User Tips

Get more out of every interaction with these features

📎 @ Mentions for Context

@file — Pull specific files into context
@commit / @URL — Reference commits or web pages
@terminal — Include terminal output directly

⚡ Slash Commands

/mode — Switch between Ask, Code, Architect
/newtask — Start fresh with clean context
/condense — Compress context when it gets long

💡 Code Selection

Highlight code → right-click to add to context

✨ Autocomplete

Works in prompts and code

Beyond the IDE: CLI, Cloud Agents, Slack Bot

Use Kilo Code wherever you work — terminal, browser, or team chat

Model Context Protocol (MCP)

The fundamental problem: AI models only receive text and return text

🔌 What MCP Does

- Open-source standard from Anthropic
- Connects AI to external resources
- One server works with all supporting agents
- Read files, run commands, call APIs

📦 Examples

GitHub MCP
67 tools for issues, PRs, workflows

Context7
Up-to-date framework documentation

Amazon
Thousands of internal MCP servers

⚠️ Each MCP server adds to your system prompt

If you don't use it, disable it — this is context engineering applied to tooling

Working with Internal Platform APIs

How to give Kilo context about platform APIs you're building against

1️⃣ Swagger / OpenAPI Spec

Pull it in directly via @file mention

Best for: Standard REST APIs

2️⃣ Convert to Markdown

Document in AGENTS.md for always-on access

Best for: Stable internal APIs

3️⃣ Reference URL/URI

Point to hosted docs that update frequently

Best for: Rapidly changing APIs

4️⃣ Build an MCP Server

For complex, multi-step interactions

Best for: Multi-project platforms

Working Alongside Kilo

How do I avoid getting my manual edits overwritten?

⚠️ The Problem

Editing the same files Kilo is working on leads to conflicts

✅ The Solution

Use CLI with parallel mode (git worktrees)

📁
You work here
Main directory

↔️
Isolated
No conflicts

🤖
Kilo works here
.kilocode/worktrees/

Review Kilo's work as a **separate PR**

Roadmap

Expect improvements to the key features you use getting richer and deeper

🌐 Expanded Surfaces

VS Code

JetBrains
Re-architected extension

CLI
New release

App Builder & Cloud Agents

🎯 Current Focus

🔧 **Feature Depth**
Building out richness of each feature

🔗 **Interconnectedness**
Better integration between features

✨ **Agent Experience**
Consistent quality across all surfaces

Thank You

🎯 **Pick one tool and get reps**

Mastery comes from practice

📋 **Try Research → Plan → Implement**

On your next feature

> *"I'm having more fun programming than I ever had in 52 years."*
>
> — Kent Beck

Questions?