ימי AI
11-22.10 בטכניון

TECHNION
Israel Institute of Technology

DEPARTMENT OF CHEMICAL ENGINEERING
TECHNION – ISRAEL INSTITUTE OF TECHNOLOGY

Claude Code: skills and agents

Advanced users workshop

QUESTIONS · SLIDO

Assoc. Prof. Alon Grinberg Dana

Department of Chemical Engineering · Technion · alon@technion.ac.il

AI DAYS · TECHNION | OCTOBER 2026

Acknowledgments

Prof. Benny Kimelfeld

Dean of Undergraduate Studies, Technion

Dr. Efrat Nativ-Ronen

Registrar for Undergraduate Studies, Technion

Shikma Kalfon

Project Coordinator, School of Undergraduate Studies

QUESTIONS · SLIDO

— THE REQUEST

An email lands in your inbox

FROM	Prof. S. Or
TO	Alon Grinberg Dana
SUBJECT	**A tool for our lab visitors**

Hi Alon,

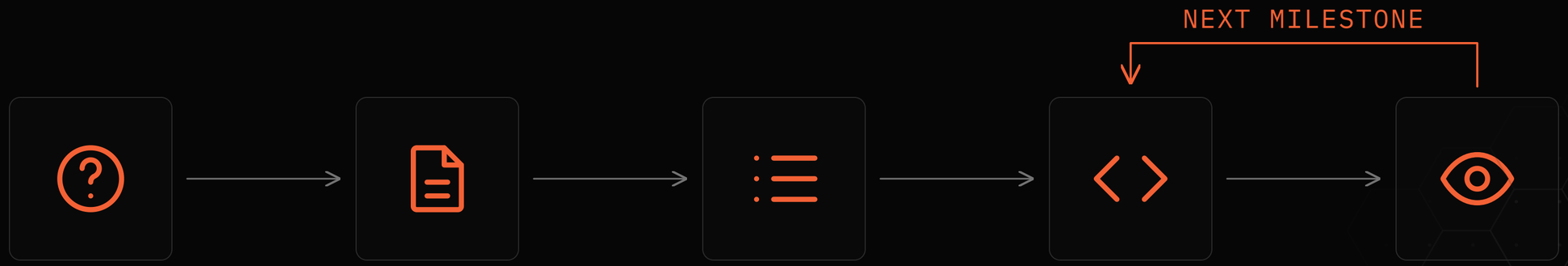
Visitors keep asking what our lab actually does. Can you build a tool that explains it to them? Something smart, with AI.

We need it before the open day.

Thanks,
Prof. S. Or

— THE WORKFLOW

From idea to implementation



Today's setup, in brief

WSL 2 · Ubuntu

Linux inside Windows.
Claude Code runs there.

Headroom

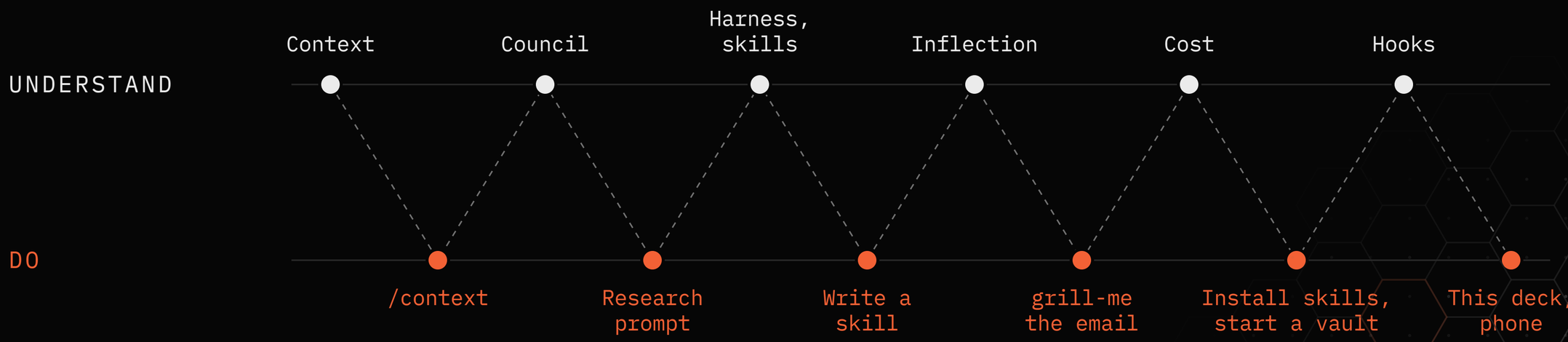
A local proxy that compresses
what the model reads.

My skills repo

github.com/DanaResearchGroup/
agent-skills

— TODAY

Understand it, then do it



Each idea is followed by your turn.

The context window is working memory

Everything the model sees at once, **counted in tokens**.

Hidden system instructions
System prompt, tools, MCP servers

Your setup
CLAUDE.md, skill descriptions

The conversation
Your prompts and its replies

The work
Every file read, every tool output

Careful: it overflows.



THE CONTEXT WINDOW

REVIEW: credit the video this view follows

Every session starts already partly full

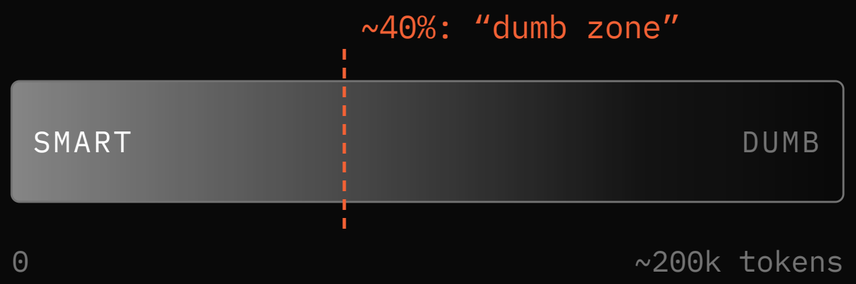


After A. Kapetanović, “Context Engineering…”, 20 Jan 2026, antekapetanovic.com/blog/context-engineering

Quality fades gradually; newer models fade later

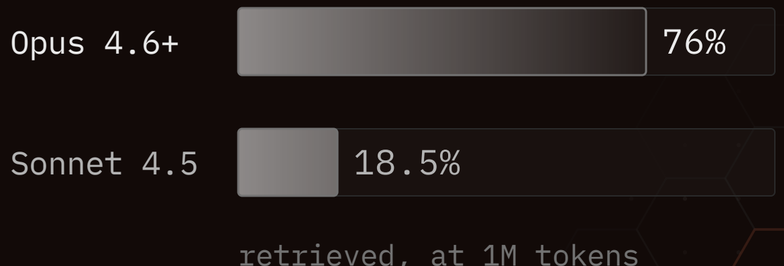
2025-ERA MODELS

~200k window



1M-NATIVE MODELS

8 needles, at 1M



No fixed threshold. Watch for **the symptoms**; compact on purpose.

Hong, Troynikov, Huber, Context Rot, Chroma, July 2025 · D. Horthy, No Vibes Allowed, AI Engineer Code 2025
8 needles at 1M (MRCR v2): Anthropic, Claude Opus 4.6, Feb 2026, anthropic.com/news/claude-opus-4-6

Long tasks run in rounds



Compaction is lossy. The handoff keeps what matters.

A paused session goes cold

25×

A cold turn against a hot one.
Opus 5.5, API prices.

	150K CONTEXT	600K CONTEXT
HOT cache read	$0.03	$0.12
COLD cache expired, re-written	$0.75	$3.00

1 h
main chat, subscription

5 min
API key, usage credits

5 min
subagents, always

Before a break: **/handoff** if needed.
After it: /handoff or /compact, then go on.

API prices: platform.claude.com/docs/en/about-claude/pricing, checked 15 Oct 2026 · TTL: code.claude.com/docs/en/prompt-caching

Your turn: /context and a % status line

FIRST

`/context`

What fills the window, in tokens.

THEN

A % status line

`/statusline show context %` writes it for you.

FINALLY

Watch it climb

Read a few files, then look again.

GUI Desktop: **Context %** in the session rail. VS Code: the **context indicator** in the prompt box.

Ask a council, not one model

Several models answer
The same question, independently

One consolidator
Merges the answers, checks the sources

Hallucinations differ.



After A. Karpathy, llm-council, github.com/karpathy/llm-council, 2025 · Image: generated with Gemini

— COUNCIL · YOUR TURN

Two skills run the council

`/deep-research-prompt`
Writes one prompt with a fixed answer table

Paste it, unchanged
ChatGPT, Gemini, Claude: Deep Research

Save each answer
One file per tool, in `responses/`

`/deep-research-consolidate`
Merges, checks every citation, shows disagreement

TRY IT TOGETHER

```
/deep-research-prompt
Do microplastics cross
the human blood-brain
barrier, and at what
levels?
```

Skills: github.com/DanaResearchGroup/agent-skills (deep-research-prompt, deep-research-consolidate)

The model thinks. The harness acts.

Claude Code is not the model. It is **the program around it**.

Model
The weights. It predicts the next token.

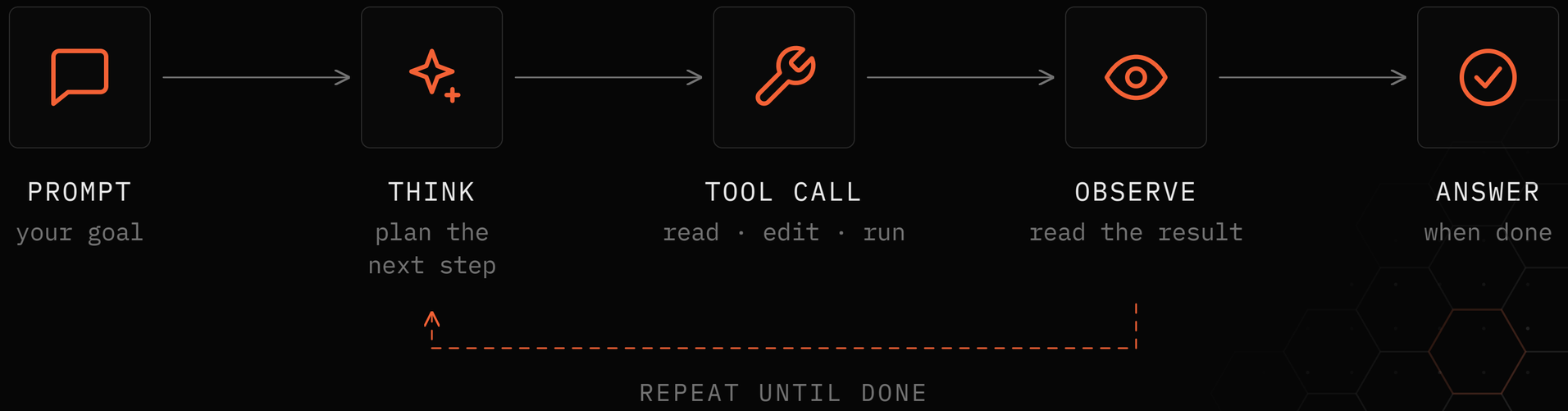
Harness
Tools, files, permissions, the loop.

You
The operator: engages with the harness, and through it with a model.



Chat vs. Agent

A chat answers once. An agent goes around until the job is done.



the classic agent: **perceive** (observe) → **reason** (think) → **act** (tool call)

Russell & Norvig, *Artificial Intelligence: A Modern Approach* · Yao et al., “ReAct: Synergizing Reasoning and Acting in Language Models”, 2023

Same models, different harnesses

Claude Code

Anthropic. The one we use today.

Codex CLI

OpenAI. Open source, Apache-2.0.

Antigravity

Google. An agent-first IDE, public preview Nov 2025.

OpenCode

MIT. 75+ providers, including free models.

OpenClaw

A personal agent. Clawdbot → Moltbot → OpenClaw, 2026.

What transfers

Skills, CLAUDE.md / AGENTS.md, and your habits.

OpenClaw: github.com/openclaw/openclaw · OpenCode: MIT · Codex CLI: Apache-2.0

A skill is instructions loaded on demand

A folder with a **SKILL.md**.
Only its description waits in context.

Frontmatter
A name, and a description of when to use it

Body
The instructions, loaded when it fires

Files beside it
Scripts, references, templates

Many skills, little context.

writing-git-commits/SKILL.md

```
---
name: writing-git-commits
description: Use when
  committing, staging, or
  splitting changes into
  commits… keeps commit
  messages short.
---
# Writing git commits
1. One logical unit per commit
2. …
```

ALWAYS IN CONTEXT: the frontmatter

Every skill costs someone's context

```
# default, neither line: the model fires it, and you can type /name
disable-model-invocation: true   # human only: the description leaves context
user-invocable: false            # model only
```

MODEL-INVOKED

Costs the model's context

The description sits in context every turn.
It fires by itself.

HUMAN-INVOKED

Costs your memory

Zero context. But you must remember to call it.

Balance the two burdens.

Write a skill with a skill

`/writing-great-skills`

Your repo's rules: lint, layout, when it fires

Wraps `/writing-for-agents`

The craft: the description is the trigger; steps end in "done when"; cut every line that changes nothing

Then test it

Give it a paragraph. Does it fire? Does it sound like you?

TRY IT TOGETHER

```
/writing-great-skills
Create an academic-voice
skill. Learn my voice from
the papers in ./papers/.
Use it whenever I draft
or edit a manuscript.
```

Skills: github.com/DanaResearchGroup/agent-skills (writing-great-skills, writing-for-agents)

— INFLECTION

Decisions hide in prose; /ask-me finds them

It sweeps the chat for decisions **never put to you**:

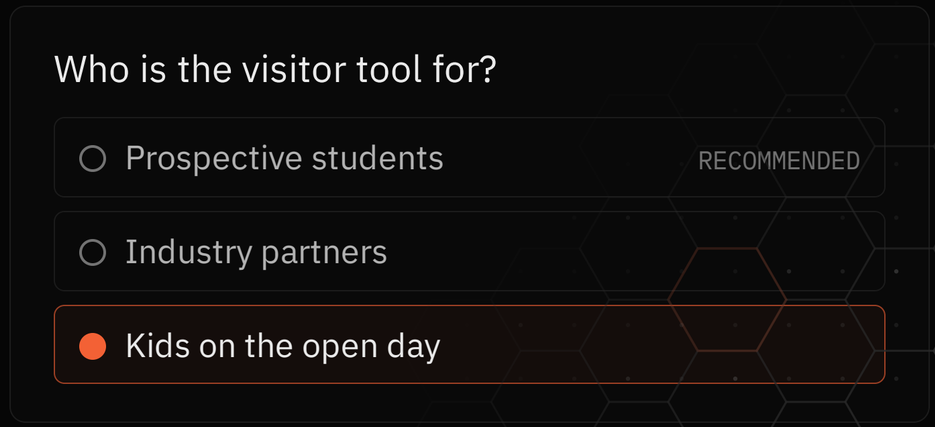
Asked in prose
A question mark inside a paragraph

Softened
"Let me know if you'd rather..."

Deferred
"We can decide that later"

Silently settled
A fork it took without telling you

EACH ONE BECOMES A QUESTION



Skill: github.com/DanaResearchGroup/agent-skills (ask-me, with framing-decisions) · built-in tool: AskUserQuestion

— INFLECTION

It asks good questions.
Your answers decide the product.

01 Don't auto-accept

Its recommendation is a guess, not your goal.

02 Ask it to reframe

“Give me three other ways to see this.”

03 Redirect the wrong question

“You're not in a job interview.”

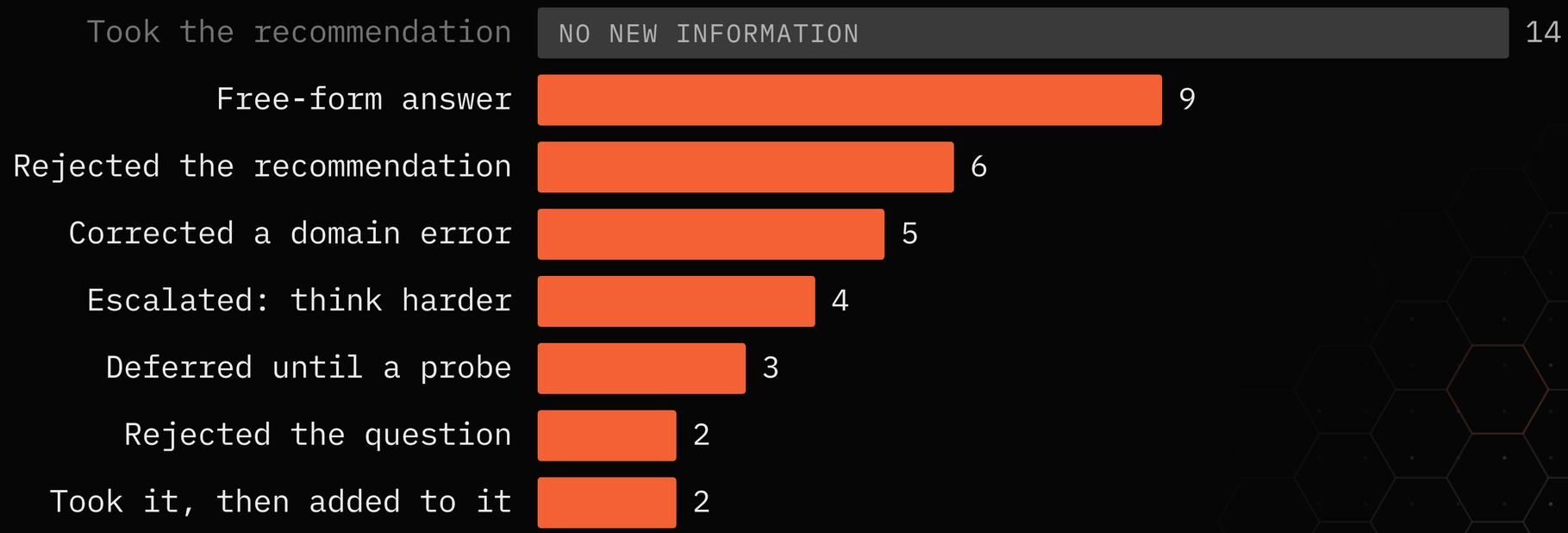
04 Challenge it

Ask what would prove it wrong.

— INFLECTION

Taking the recommendation adds nothing

What I did with the agent's 39 questions. The ideas came from the other answers.



5-day design session, Aug 2026; meta-analysis skill. 39 questions; an answer can count twice (sum 45).

— INFLECTION

Two real examples for inflection points

REDIRECT

AGENT

The limit is 5 pages.
What should be cut out?

ME

We're building the skeleton now. Focus on the science, not on length.

REFRAME

AGENT

The hypothesis is [A].
How do we test it?

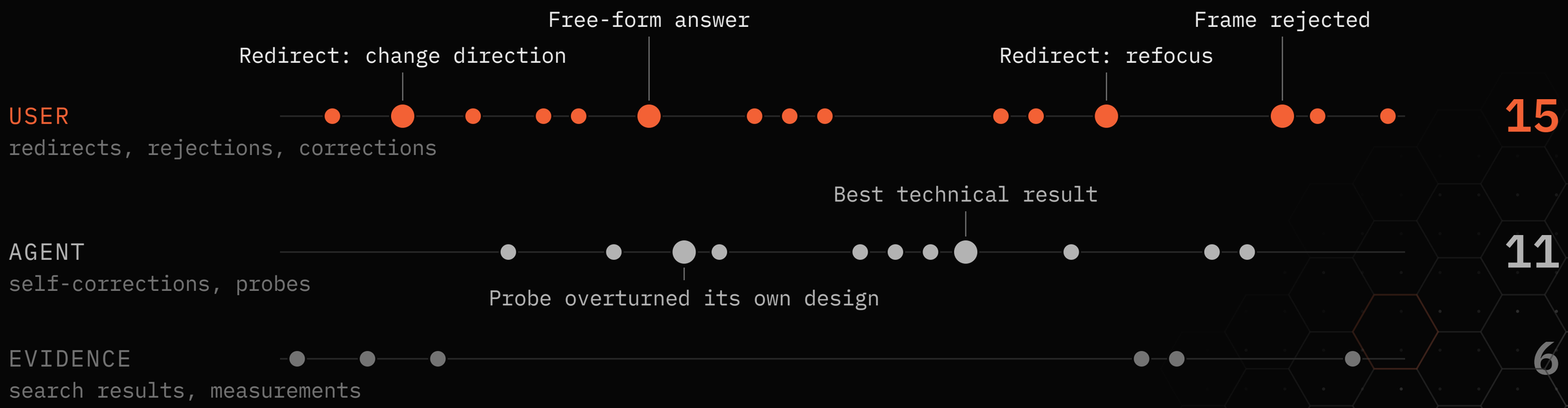
ME

No, reframe. The hypothesis is different: [B].

Real exchanges from Alon's Claude Code sessions, 2026; hypotheses withheld.

— INFLECTION

The turns came from the user



The same session: 32 turning points in 5 days, in order; tallied by a meta-analysis skill.

Pick the smallest model that does the job

MY LADDER	TIER	HARNESS · MODEL	EFFORT	
8	lead	claude · claude-opus-5-5 [1M]	high	Strategy, architecture, final verdicts
7	review	codex · gpt-6.1-sol	high	Verdicts only: review, go/no-go
6	hard	codex · gpt-6-astra	high	Hard integration; retries cost more
5	craft	codex · gpt-6.1-sol	medium	Ambiguous multi-file work
4	build	codex · gpt-6-luna	medium	Bounded coding, with tests DEFAULT
3	triage	codex · gpt-6-luna	low	Triage, finding files
2	free	oc-free · Nemotron 3 Ultra	—	Work a test already checks
1	local	ollama · qwen3.6:35b	—	Private data; never leaves
0	script	no LLM	—	Plain code: exact checks

Start at **build**. Promote one rung on evidence, not hope.

My ladder.yaml (model-ladder skill, main as of 18 Oct 2026) · Claude Code models and effort: code.claude.com/docs/en/model-config

— COST

Compress what the model reads, before you pay for it



12–28%
typical

90%+
on repetitive logs

reversible
nothing is lost

github.com/headroomlabs-ai/headroom · savings measured locally

Skills worth taking home

`/deep-research-*`
One prompt to several research tools; one audited table.

`/spar`
A second model attacks your plan.

`/obsidian-vault`
Search and write notes in your vault.

`/grill-me`
It interviews you until the plan holds.

`/ask-me`
Finds decisions it settled without you.

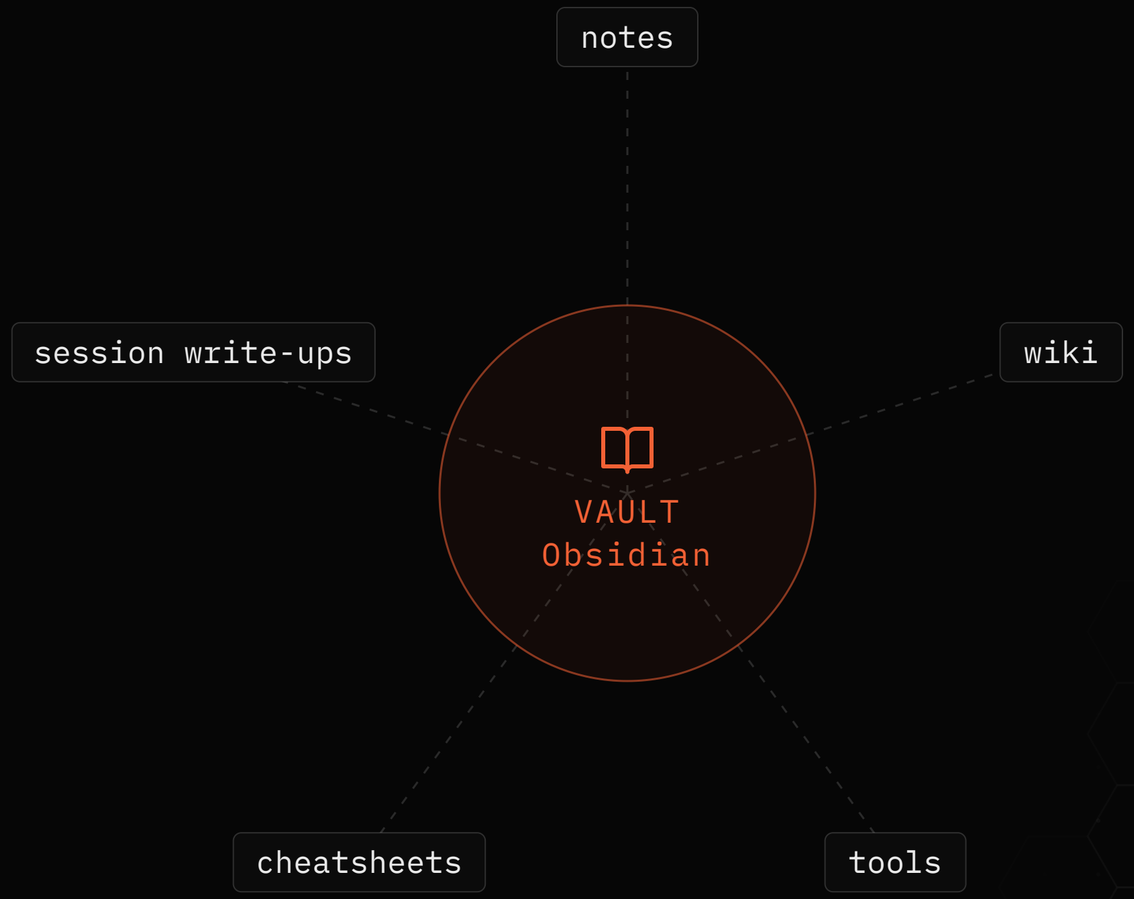
`/pr`
A pull-request body: summary, evidence, merge danger.

github.com/DanaResearchGroup/agent-skills

Install: README, step 1 (Mac, Linux, Windows). Then open a new session.

A vault is memory across every project

Plain markdown in one folder, opened in **Obsidian** (free app, obsidian.md). Every session reads it.



Start a vault this week



OBSIDIAN
install, open a folder as a vault

CLAUDE.md
how the vault is organised

SKILL
obsidian-vault

WRITE UP
“write this session up as a wiki page”

SEARCH FIRST
before you build

EVERY SESSION ADDS A PAGE

CLAUDE.md asks. A hook enforces.

A hook is **a command the harness runs** at an event.

In
JSON about the event

Out
Allow, deny, or add context

Where
Terminal, IDE, Desktop, cloud sessions

CLAUDE.md is advice. A hook is a rule.



code.claude.com/docs/en/hooks

This deck was built with Claude Code



Corrections become rules

PRESENTATIONS/CLAUDE.MD · ALON'S RECURRING CORRECTIONS

4×4 text, short hand-broken lines

Few clicks: the rest appears by itself

Large images: trim text, not art

Bold lecturer cues; ports keep sub-points

Illustrations on most slides

Nothing lost in the port

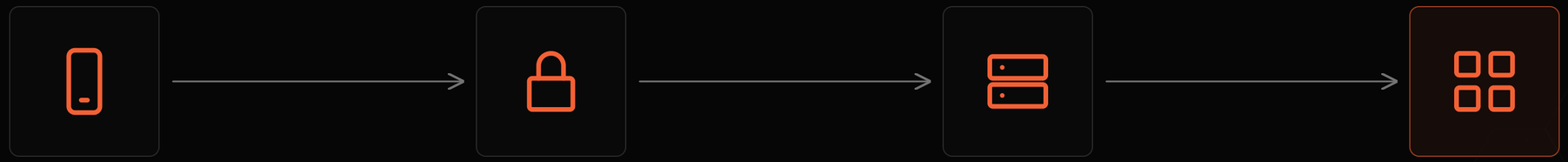
Concrete wording for non-native readers

Runs hold still: fixed positions

Each was corrected more than once.

— IN PRACTICE

Your agents, from your phone



working | blocked: needs you | done

Three things to take home

01

Guard the context

Watch it.
Compact on purpose.

02

Argue with it

The answers you add
make the product.

03

Turn habits into skills

Write it once. It fires by itself.

Assoc. Prof. Alon Grinberg Dana

Department of Chemical Engineering · Technion · alon@technion.ac.il